UNIVERSIDAD DE GRANADA

Facultad de Ciencias

Aportaciones a la estimación en áreas pequeñas. Estimación de proporciones.

Tesis Doctoral

Directores de tesis:
Profr. Dr. D. Antonio Arcos Cebrian
Profa. Dra. D. María del Mar Rueda García

Agustín Santiago Moreno

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Granada 2011

Editor: Editorial de la Universidad de Granada
Autor: Agustín Santiago Moreno
D.L.: GR 492-2012
ISBN: 978-84-694-6002-3

Aportaciones a la estimación en áreas pequeñas. Estimación de proporciones.

Memoria presentada por Agustín Santiago Moreno para obtener el título de Doctor por la Universidad de Granada.

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

UNIVERSIDAD DE GRANADA

Granada 2011

Fdo.: Agustín Santiago Moreno

Directores de Tesis:

V.B.
Prof. Dr. Antonio Arcos Cebrian

V.B.
Profa. Dra. D. María del Mar Rueda García

Índice general

Capítulo 1

Introducción

Uno de los objetivos de la estimación en áreas pequeñas, consiste en proponer estimadores, que sin aumentar el tamaño de muestra, permitan obtener buenas estimaciones, es decir, el error de estimación sea menor que el de los estimadores conocidos. Hasta ahora, en el contexto de la estimación en áreas pequeñas, se han desarrollado distintos métodos de estimación, bajo el diseño y bajo el modelo, para abordar este problema. La estimación basada en modelos, que ligan la información auxiliar a la variable de interés con la cual se supone relacionada, ha demostrado ser más eficiente que la estimación basada en el diseño, si es que se cuenta con información auxiliar completa, es decir, se observan la variable respuesta y un vector de p variables auxiliares que son conocidas para todos los individuos de la población. Desgraciadamente, en la práctica este supuesto no es muy común, siendo más plausible que los datos asociados a las variables auxiliares se obtengan de censos y ficheros administrativos que proporcionan diferentes parámetros de estas variables auxiliares. Así, es común disponer de diversas medidas de posición (medias, medianas, momentos, etc.) pero es difícil tener acceso a los datos originales de cada individuo, fundamentalmente por motivos de privacidad o cuando se trata de información relacionada con políticas públicas. En este contexto es en el que proponemos un conjunto de estimadores de proporciones para áreas pequeñas, si es que se dispone de información auxiliar en forma de totales o proporciones a nivel de la población o a nivel de dominio.

Hemos organizado la revisión de la estimación en áreas pequeñas en dos grandes bloques, el primero trata de los estimadores para variables cuantitativas y el segundo, trata de los estimadores para variables cualitativas. Pretendemos que la forma en cómo presentamos la estimación para variables cualitativas, contribuya a la organización sistematizada de los estimadores de proporciones para áreas pequeñas, donde las variables de interés sean binarias o de frecuencias.

En la primera parte se inicia con la definición del problema de la estimación en áreas pequeñas, los objetivos y alcances del trabajo de investigación. En el segundo capítulo se describen los métodos de estimación en áreas pequeñas para variables cuantitativas, empezando por los estimadores directos, estimadores directos modificados, estimadores sintéticos y estimadores combinados. Se sigue con los estimadores mixtos y estimadores EBLUP, que hacen uso de información auxiliar ajustada a un modelo de regresión generalizado (GREG), con objeto de obtener mejores estimadores o predictores lineales insesgados en dos etapas. Se describen los estimadores de bayes empírico (BE) y bayes jerarquico (BJ) como herramientas que permiten hacer estimaciones exactas en áreas pequeñas, usando modelos de enlace apropiados y algoritmos recursivos para la estimación de los parámetros de interés y para la estimación de las varianzas. Una excelente recopilación de estos métodos puede consultarse en Rao (2003) y desarrollos posteriores, sobre todo de modelos lineales mixtos y modelos de bayes, en los cuales se asume para la distribución apriori y a postriori, distribuciones distintas a la normal. Ver por ejemplo Larsen (2003), González-Manteiga, et al.(2007), Xie, et al. (2007), Trevisani, et al. (2007), Liu (2009) y Chandra et al. (2009). Modelos no paramétricos pueden consultarse en Chambers, et al. (2006) y Salvati, et al. (2010).

En el capítulo 3 se describen las técnicas de estimación en áreas pequeñas para variables cualitativas. Se inicia con la descripción de métodos basados en modelos y se deja para el final la estimación basada en el diseño. Hemos organizado la presentación de esta forma, porque la contribución nuestra se enmarca en el enfoque basado en el diseño. La estimación de proporciones basada en modelos, para los modelos básicos de nivel de área, se estudia para modelos de efectos mixtos y estimadores EBLUP, para estimadores bajo el enfoque de bayes, haciendo énfasis en los modelos mixtos lineales generalizados, que han sido usados especialmente para el ajuste de datos binarios y de frecuencias, haciendo uso de distintos modelos de enlace y distribuciones a priori, incluyendo la distribución binomial, normal, logit, poisson, gamma, exponencial y doble exponencial. Finalmente, se describe la estimación de proporciones basada en el diseño, destacando las contribuciones en la estimación directa, sintética y combinada. Se presentan los modelos LGREG, propuestos por Lehtonen y Veijanen (1998) para la estimación en dominios y los propuestos por Duchesne (2003) para la estimación de proporciones en dominios.

El capítulo 4 está destinado a la presentación de los estimadores de proporciones de áreas pequeñas que proponemos. Como señalamos antes, la estimación se hace bajo el diseño, empezando por la estimación directa de proporciones de dominio, estimadores de razón, de regresión, de diferencia y de calibración, con información auxiliar en forma de totales o proporciones a nivel de población o a nivel de dominio. Se halla también el correspondiente

estimador sintético para cada caso y estimadores de sus varianzas. Para los estimadores de razón, de regresión y de diferencia, se proponen estimadores combinados y estimadores de sus varianzas. Se proponen, además, las extensiones a un diseño general de muestreo y al caso multivariado.

En el capítulo 5 se describe la población artificial, a partir de la cual se hacen las simulaciones, para la obtención de medidas de precisión como error cuadrático medio, sesgo relativo y eficiencia relativa de los estimadores. Se repite este proceso con una muestra disponible de casos confirmados de dengue para el dominio 6, que corresponde a una región de la Costa chica, y se estima su error cuadrático medio usando Jackknife. Identificados los estimadores más eficientes, se hace la estimación de proporciones para las siete regiones del Estado de Guerrero, México, de pacientes infectados por dengue, a partir de una base de datos de casos confirmados del año 2006, registrados por el Laboratorio Estatal de Salud Pública del estado de Guerrero (LESPEG), institución encargada del control epidemiológico de la enfermedad. Se discuten los resultados y se emiten conclusiones al respecto.

1.1. El problema de la estimación en áreas pequeñas

En la actualidad, prácticamente todos los gobiernos de los países del mundo realizan encuestas nacionales o estatales para obtener información sobre distintas características de la población. Esta información es utilizada en la formulación de políticas y programas, en la evaluación, en la asignación de fondos gubernamentales y en la planificación regional. Las agencias estadísticas y el sector privado hacen uso de información de encuestas para apoyar las decisiones de negocios, reorientar una campaña política, posicionar un producto en el mercado, etc. Normalmente estas encuestas se diseñan para producir estimaciones para la población objetivo de muestreo y para subgrupos poblacionales grandes. Las estimaciones de muestreo estándar de poblaciones finitas para subgrupos grandes de la población, se denominan estimaciones basadas en el diseño o estimaciones directas, debido a que la estimación está basadas solamente en los datos de muestreo y las probabilidades de selección para la muestra en el subgrupo de interés. La inferencia estadística bajo el diseño no depende de la validez de un modelo estadístico, a diferencia de lo que ocurre en la mayoría de las áreas de la estadística. Esta situación resulta ventajosa en la estimación, toda vez que no necesitamos suponer un modelo apriori sobre los datos. Sin embargo, la inferencia bajo el diseño se torna problemática cuando el tamaño de muestra en los subgrupos de interés es muy pequeño o incluso cero. En esta situación, se han propuesto y se usan cada vez más, diversos

métodos basados en modelos para ajustar las variables de interés y la información auxiliar asociada, con objeto de producir los denominados estimadores de áreas pequeñas o estimadores indirectos.

El término "área pequeña" o "dominio" usualmente se refiere a un área geográfica pequeña tal como un estado, provincia, municipio, distrito escolar, área metropolitana, o también puede tratarse de un grupo específico de edad-sexo, sexo-distrito electoral-preferencia política, u otro grupo de interés dentro de un área geográfica grande. La realización de encuestas nacionales que también sean representativas de las áreas geográficas de nivel inferior es posible, pero puede no ser viable desde el punto de vista de los costos.

Un ejemplo lo tenemos en la encuesta PISA, de la cual recientemente ha tenido lugar la publicación de los resultados del último informe correspondiente al año 2009, que ha sido concebida como un recurso para ofrecer información abundante y detallada, que permita a los países miembros adoptar las decisiones y políticas públicas necesarias para mejorar los niveles educativos, servir de base para la investigación y análisis, destinados a mejores políticas en el campo de la educación.

El tamaño de la muestra de esta encuesta es moderado (25887 alumnos para España) y los estimadores dados permiten realizar inferencias válidas del país en su totalidad y de algunas comunidades autónomas que han ampliado la muestra (como Andalucía con 1416 alumnos encuestados), pero no permiten inferencias fiables por regiones, provincias y áreas pequeñas de interés. Los procedimientos para la asignación de pesos en PISA reflejan los estándares de las buenas prácticas de las encuestas, si bien no usan la información auxiliar disponible en la fase de estimación, debido, fundamentalmente, a la enorme variedad de tal información auxiliar según país, región, comunidad, comarca, provincia, etc. Sin embargo, usando técnicas indirectas avanzadas, es posible obtener estimadores complejos que usen la información auxiliar disponible de centros, familias, localidades y alumnos, mediante la formulación de modelos adecuados en el área considerada.

Las estadísticas de áreas pequeñas se realizan desde hace mucho tiempo en países como Inglaterra y Canadá, que las han usado en sus censos y registros administrativos desde hace varios siglos. Los demógrafos llevan mucho tiempo usando una gran variedad de métodos indirectos para hacer estimaciones de áreas pequeñas de poblaciones, pero prescindiendo de las técnicas de muestreo clásicas. Actualmente la mayoría de países europeos hacen uso de ellas, las economías en transición, como los países de Europa central y oriental, los países de la ex unión soviética y varios países latinoamericanos.

Durante las pasadas tres décadas, la demanda de estimaciones para áreas pequeñas se ha incrementado en diferentes áreas de aplicación, incluyendo ingreso y pobreza, educación, salud, uso de sustancias y agricultura. La razón

principal para el incremento en la demanda de estadísticas de áreas pequeñas, se encuentra en la reciente tendencia en la política federal para objetivos sociales y programas económicos a un nivel más local.

1.1.1. Objetivos

La mayoría de métodos propuestos hasta ahora para estimar en áreas pequeñas, funcionan muy bien para variables continuas, incluyendo los métodos de estimación basados en modelos, que hacen uso de información auxiliar disponible a partir de censos, registros administrativos o encuestas anteriores. La inserción de esta información auxiliar disponible en la formulación de los estimadores mediante distintas técnicas (como regresión, calibración, regresión no paramétrica o verosimilitud empírica), normalmente produce un aumento considerable en la precisión de los estimadores de la media o el total poblacional.

En todos estos trabajos se asume que la información auxiliar está disponible para cada persona en el área. Este supuesto no es muy común, siendo más plausible que los datos asociados a las variables auxiliares se obtengan de censos y ficheros administrativos que proporcionan diferentes parámetros de estas variables auxiliares.

Los objetivos en los cuales centraremos nuestro trabajo son los siguientes:

i Proponer estimadores para proporciones de áreas pequeñas que permitan la incorporación de la información auxiliar, en forma de totales o proporciones, proveniente de una variable binaria.

ii Hacer una comparación de estos estimadores con otros que incluyen información auxiliar completa a partir de medidas de precisión.

iii Organizar de forma sistemática los estimadores de áreas pequeñas para proporciones, incluyendo los propuestos.

1.1.2. Notación y conceptos básicos

En el presente trabajo denotaremos a la *población* finita de interés como U, con N unidades distintas e identificadas, es decir, $U = \{1, \cdots, j, \cdots, N\}$. La lista que nos permite identificar cada una de las unidades de la población se llama *marco*. Las unidades poblacionales poseen muchas características de interés, algunas conocidas y otras desconocidas. Representamos por la variable y (o A, si se trata de un atributo) la característica de la población que deseamos estudiar, y que llamaremos *variable de estudio*. El valor que toma dicha variable sobre la población es desconocido, pero vendrá dado por el valor que cada unidad j asigna a dicha característica, y_j,

$$y = \{y_1, y_2, \cdots, y_j, \cdots, y_N\}$$

o si se trata de un atributo A,

$$A = \{A_1, A_2, \cdots, A_j, \cdots, A_N\}.$$

Además, podemos registrar las características conocidas de una variable x o de un atributo B, que supondremos p-dimensional, al disponer de p características de cada unidad poblacional. Así, para cada unidad j en la población tenemos un vector de información adicional, $\mathbf{x}_j = (x_{j1}, \cdots, x_{jp})$. Por ello, la variable,

$$\mathbf{x} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$$

o

$$\mathbf{B} = \{\mathbf{B}_1, \cdots, \mathbf{B}_N\}$$

recibe el nombre de *variable de información adicional*. Esta información auxiliar, que se supone relacionada con la variable de interés, debe emplearse para ayudar a conocer los valores asociados a la variable de interés.

Toda vez que la variable de interés que queremos estudiar representa una característica "no conocida" de la población U, para "su conocimiento total" sólo hay un procedimiento, la investigación de todas la unidades de la población. Esta investigación exhaustiva recibe el nombre de *censo*. No obstante, es frecuente que no deseemos "toda" la información, sino alguna parte de ella o bien alguna función de los valores $\{y_1, y_2, \cdots, y_j, \cdots, y_N\}$, que denominamos *parámetros* y que se representan genéricamente por $\theta(y)$. Una inferencia sobre los parámetros poblacionales puede realizarse al obtener la información de una parte de la población, que recibe el nombre de *muestra* y será designada por s.

La determinación de la muestra que permita conocer la información deseada sobre toda la población, es un punto crucial de toda la teoría estadística, pues como es lógico se desea que con el menor número de elementos en la

muestra, lo que se conoce como *tamaño muestral* (denotado por n) y por ello con el menor número de elementos a estudiar, se obtenga un juicio aceptable sobre el verdadero valor del parámetro, el cual se habría obtenido al realizar la investigación sobre toda la población. El valor obtenido a partir de la muestra se llama *estimador* del correspondiente parámetro poblacional.

Una muestra s de U, de tamaño n_s, seleccionada de acuerdo a un diseño de muestreo especificado, que asigna una probabilidad conocida $p(s) > 0$, tal que para todo $s \in S$, donde S es el conjunto de las posibles muestras s, cumple con la condición $\sum_{s \in S} p(s) = 1$. Designemos por π_j, la probabilidad de que de unidad j pertenezca a la muestra s y π_{jk} representa la probabilidad de que la unidades j y k pertenezcan a la muestra s. A esta última cantidad se le llama *probabilidades de inclusión de segundo orden*. Otras cantidades que serán usadas con frecuencia en las expresiones de los estimadores y sus errores son los pesos básicos de diseño $w_j = \pi_j^{-1}$ y $\triangle_{jk} = \pi_{jk} - \pi_j \pi_k$.

Consideremos la población U particionada en D subconjuntos, $U_1, \cdots, U_d, \cdots, U_D$, llamados *dominios*. Designemos por N_d, el tamaño del dominio U_d, desconocido en la mayoría de casos prácticos. Si seleccionamos una muestra aleatoria de la población U, de acuerdo a un diseño de muestreo especificado y probabilidades de inclusión π_j, π_{jk}, parte de la muestra s cae en el dominio U_d. Esta parte de la muestra se denotará por s_d, es decir

$$s_d = s \cap U_d.$$

Denotemos por n_{s_d} el tamaño de s_d. Debe cumplirse que

$$s = \cup_{d=1}^{D} s_d; \; n_s = \sum_{d=1}^{D} n_{s_d},$$

donde n_{s_d} es una variable aleatoria y posiblemente bastante pequeña.

Estamos interesados en la estimación de totales, medias o proporciones, por lo que designaremos por total poblacional a

$$Y = \sum_{j=1}^{N} y_j,$$

total de dominio

$$Y_d = \sum_{j=1}^{N_d} y_j,$$

la media

$$\overline{Y} = N^{-1} \sum_{j=1}^{N} y_j,$$

la media de dominio

$$\overline{Y}_d = N_d^{-1} \sum_{j=1}^{N_d} y_j.$$

Para proporciones tenemos

$$\overline{P} = N^{-1} \sum_{j=1}^{N} A_j,$$

la proporción de dominio

$$\overline{P}_d = N_d^{-1} \sum_{j=1}^{N_d} A_j.$$

Los estimadores de Horvitz-Thompson de la media y al proporción para el dominio se designarán por

$$\overline{y}_{d_{HT}} = N_d^{-1} \sum_{j \in s} w_j y_j$$

y

$$\overline{p}_{d_{HT}} = N_d^{-1} \sum_{j \in s} w_j A_j,$$

respectivamente, donde w_j son los pesos de diseño.

Capítulo 2

Estimación para variables cuantitativas

En este capítulo haremos un resumen de los estimadores de áreas pequeñas desarrollados para estimar características poblacionales, considerando variables cuantitativas. Discutiremos en primer lugar los estimadores basados en el diseño y en seguida los basados en modelos, aunque en algunos casos abordaremos estimadores que poseen ambas características, es decir, son basados en el diseño asistidos por modelos que posibilitan la incorporación de información auxiliar en la fase de estimación.

2.1. Estimación basada en el diseño

En este tipo de estimación se usa un diseño de muestreo para seleccionar una muestra s de U con probabilidad $p(s)$. La probabilidad de selección de la muestra $p(s)$ puede depender de variables de diseño conocidas tales como variables indicadoras de estratos o la medida del tamaño del conglomerado. Comúnmente los diseños de muestreo usados comprenden muestreo aleatorio simple, estratificado y muestreo estratificado multietápico.

La población U consiste de N distintos elementos (o unidades finales) identificadas mediante las etiquetas $j = 1, \cdots, N$. Asumimos que la característica de interés , y, asociada con el elemento j puede medirse exactamente por observación del elemento j, es decir, asumimos que se mide sin error. El parámetro de interés es el total $Y = \sum_U y_j$ o la media $\overline{Y} = Y/N$ poblacionales, donde $\sum_U$ denota la suma sobre los j elementos de la población.

Para hacer inferencias sobre la media poblacional $\overline{Y}$, observamos lo valores y asociados a la muestra seleccionada s. En este enfoque, un estimador $\widehat{\overline{Y}}$ de $\overline{Y}$ se dice que es insesgado bajo el diseño (o p-insesgado) si la esperanza de $\widehat{\overline{Y}}$

bajo el diseño es igual a $\overline{Y}$; es decir

$$E_p(\widehat{\overline{Y}}) = \frac{1}{N}\sum p(s)y_s = \overline{Y}, \tag{2.1}$$

donde la suma es sobre todas las muestras posibles s bajo el diseño especificado y y_s es el valor de y para la muestra s.

La varianza de $\widehat{\overline{Y}}$ bajo el diseño se denota como $V_p(\widehat{\overline{Y}}) = E_p[\widehat{\overline{Y}} - E_p(\widehat{\overline{Y}})]^2$. Un estimador de $V_p(\widehat{\overline{Y}})$ se denota como $v(\widehat{\overline{Y}}) = s^2(\widehat{\overline{Y}})$, y la varianza del estimador es p-insesgada para $V(\widehat{\overline{Y}})$ si $E_p[v(\widehat{\overline{Y}})] \equiv V_p(\overline{Y})$.

Para el caso de estimación en dominios, consideremos una partición de la población $U = \{1, \cdots, j, \cdots, N\}$ en D subconjuntos, $U_1, \cdots, U_d, \cdots, U_D$ llamados dominios. Designemos por N_d el tamaño de U_d, es decir N_d es el número de elementos en U_d de tal manera que $N = \sum_{d=1}^{D} N_d$. Sea $P_d = N_d/N$ el tamaño relativo de U_d, es decir, la proporción de elementos de U que pertenecen a U_d. Asumimos que N es conocida y N_d puede ser conocida o no, como con frecuencia ocurre en la práctica. Si los elementos en U_d pueden ser identificados de antemano y listados, es decir, N_d es conocida, podemos seleccionar una muestra directamente del dominio de acuerdo a un diseño de muestreo conveniente. El dominio, entonces, es designado como estrato, y la selección muestral en el dominio es controlado por la elección del diseño en el estrato. Sin embargo, aún cuando U_d pueda distinguirse como un estrato, en la práctica no siempre es posible seleccionarlo. Si el número de dominios D es grande, el costo de selección controlada en cada dominio puede ser muy alto. En este trabajo daremos mayor énfasis al caso en que resulta imposible (por ejemplo, por falta de marco de muestreo) o impráctico (por ejemplo, por razones de costo) tratar al dominio como un estrato.

2.1.1. Estimadores directos

En el contexto de encuestas por muestreo, un estimador de dominio se dice directo si está basado solamente en los datos de la muestra en el dominio. Un estimador directo puede usar también información auxiliar conocida, tal como el total de una variable, x, relacionada a la variable de interés, y. Los estimadores basados en el diseño hacen uso de los pesos muestrales, y las inferencias asociadas están basadas en la distribución de probabilidad inducida por el diseño de muestreo con los valores poblacionales considerados fijos. A nivel de dominio podemos hacer estimaciones directas, siempre que los tamaños de muestra dentro del dominio sean suficientemente grandes. Los pesos del dominio $w_j(s)$ juegan un importante rol en la construcción de estimadores $\widehat{\overline{Y}}$ de $\overline{Y}$ basados en el diseño. Estos pesos básicos dependen de s y del elemento

$j(j \in s)$. Una importante elección de los pesos es considerar $w_j(s) = 1/\pi_j$, donde $\pi_j = \sum_{\{s:j\in s\}} p(s)$, $j = 1, 2, \cdots, N$ son las probabilidades de inclusión de primer orden y $\{s : j \in s\}$ denota todas las muestras s que contienen al elemento j. Para simplificar la notación escribiremos $w_j(s) = w_j$, excepto cuando la notación completa $w_j(s)$ sea necesaria. El peso w_j puede interpretarse como el número de elementos en la población representados por el elemento muestral j.

Si nuestro interés se centra en la estimación del tamaño del dominio N_d o su tamaño relativo P_d, estos parámetros pueden manejarse como casos especiales de la estimación de un total y una media poblacionales (Särndal et al., 1992). Para ver esto, introducimos una variable indicadora de dominio, z_d, cuyo valor para el j-ésimo elemento está definido como

$$z_{d_j} = \begin{cases} 1, & \text{si } j \in U_d; \\ 0, & \text{en otro caso,} \end{cases} \tag{2.2}$$

con $j = 1, \cdots, N$. Ahora

$$\sum_U z_{d_j} = N_d$$

y

$$\overline{z}_{d_j} = \sum_U z_{d_j}/N = N_d/N = P_d,$$

lo cual identifica a N_d como el total poblacional de la nueva variable z_d y P_d como la media poblacional de z_d. Estos resultados pueden usarse para estimaciones directas. Introducimos alguna notación complementaria relacionada con dominios. Designemos por $Q_d = 1 - P_d$, $n_d = \sum_s z_{d_j}$, el número de elementos en la muestra que pertenecen a U_d, $p_d = n_d/n$, la proporción de elementos en la muestra que pertenecen a U_d y $q_d = 1 - p_d$.

A partir e la definición de z_{d_j} y mediante un desarrollo algebraico simple obtenemos que

$$S^2_{z_d U} = \frac{N}{N-1} P_d Q_d \tag{2.3}$$

$$S^2_{z_d s} = \frac{n}{n-1} p_d q_d \tag{2.4}$$

Con frecuencia estamos interesados en la estimación de un total de dominio $Y_d = \sum_{U_d} y_j$ o la correspondiente media de dominio $\overline{y}_{U_d} = \sum_{U_d} y_j/N_d$, cuyos estimadores serán designados por $\widehat{Y}_d$ y $\widehat{\overline{y}}_{U_d}$, respectivamente. Para hacerlo, es necesario redefinir la variable de interés y de la siguiente manera

$$y_{d_j} = \begin{cases} 1, & \text{si } j \in U_d; \\ 0, & \text{en otro caso,} \end{cases}$$

Entonces Y_d es el total (sobre la población completa U) de la nueva variable y_d, es decir

$$Y_d = \sum_{U_d} y_j = \sum_{U} y_{dj}.$$

En ausencia de información auxiliar poblacional, usamos el estimador expandido del total

$$\widehat{Y}_d = \sum_{s} w_j y_{dj} = \sum_{s_d} w_j y_j, \tag{2.5}$$

donde $\sum_s$ denota la suma sobre $j \in s$, $s_d = U_d \cap s$, es decir, s_d es el subconjunto de la muestra s que cae en el dominio U_d, con n_s el tamaño de s y n_{s_d} el tamaño de s_d.

El caso $w_j(s) = 1/\pi_j$ satisface la condición de insesgadez y lleva al conocido estimador de Horvitz-Thompson (H-T) (Cochran, 1977; p. 259).

La varianza de este estimador se da en la forma de un estimador cuadrático insesgado no negativo de la varianza de $\widehat{Y}_d$, construido a partir de un resultado de Rao (1979),

$$v(\widehat{Y}_d) = -\sum_{j<k,\ j,k\in s} \sum w_{jk}(s) b_j b_k \left(\frac{y_j}{b_j} - \frac{y_k}{b_k}\right)^2, \tag{2.6}$$

donde los pesos $w_{jk}(s)$ satisfacen la condición de insesgadez y las constantes b_j distintas de cero son tales que la varianza de $\widehat{Y}_d$ se hará cero cuando $y_j \propto b_j$ para todo j. Por ejemplo, en el caso especial de $w_j = 1/\pi_j$ y un diseño de muestreo fijo, tenemos $b_j = \pi_j$ y $w_{jk}(s) = (\pi_{jk} - \pi_j\pi_k)/(\pi_{jk})$ donde $\pi_{jk} = \sum_{\{s:(j,k)\in s\}} p(s)$ $j < k = 1, \cdots, N$ son las probabilidades de inclusión conjunta asumidas positivas. Bajo estas condiciones, un estimador de la varianza de la media estimada estará dado por

$$\widehat{V}(\widehat{\overline{Y}}_d) = -\frac{1}{2\widehat{N}_d^2} \sum\sum_{s_d} \left(\frac{\pi_{jk} - \pi_j\pi_k}{\pi_{jk}}\right) \left(\frac{y_j - \widehat{\overline{y}}_{U_d}}{\pi_j} - \frac{y_k - \widehat{\overline{y}}_{U_d}}{\pi_k}\right)^2. \tag{2.7}$$

Para un diseño general de muestreo la varianza estimada del estimador del total será

$$\widehat{Var}(\widehat{Y}_d) = \sum_{j,k\in s} \frac{\pi_{jk} - \pi_j\pi_k}{\pi_{jk}} \frac{y_j}{\pi_j} \frac{y_k}{\pi_k}. \tag{2.8}$$

El estimador de la media de y para el dominio estará dado por

$$\widehat{\overline{Y}}_d = \widehat{\overline{y}}_{U_d} = \frac{\widehat{Y}_d}{\widehat{N}_d} = \frac{\sum_s z_{d_j} w_j y_j}{\sum_s z_{d_j} w_j} = \frac{\sum_{s_d} w_j y_j}{\sum_{s_d} 1 w_j} = \frac{\sum_{s_d} y_j/\pi_j}{\sum_{s_d} 1/\pi_j} \tag{2.9}$$

y su varianza aproximada estimada será

$$\widehat{V}(\widehat{\overline{Y}}_d) = \frac{1}{\widehat{N}_d^2} \sum\sum_{s_d} \left(\frac{\pi_{jk} - \pi_j \pi_k}{\pi_{jk}} \right) \left(\frac{y_j - \widehat{\overline{y}}_{U_d}}{\pi_j} \right) \left(\frac{y_k - \widehat{\overline{y}}_{U_d}}{\pi_k} \right). \tag{2.10}$$

2.1.2. Estimador de regresión generalizado para dominios

En este tipo de estimador se asume que se dispone de información auxiliar en forma de totales poblacionales $\mathbf{X} = (X_1, \cdots, X_p)^T$, independientemente de que N_d sea conocido o no, y que el vector auxiliar $\mathbf{x}_j$ para $j \in s$ también es observado, es decir, se observan los datos $(y_j, \mathbf{x}_j)$ para cada elemento $j \in s$. Un estimador que hace eficiente el uso de esta información auxiliar es el estimador de regresión generalizado (GREG) el cual se escribe como

$$\widehat{Y}_{GR} = \widehat{Y} + (\mathbf{X} - \widehat{\mathbf{X}})^T \widehat{\boldsymbol{\beta}}, \tag{2.11}$$

donde $\widehat{\mathbf{X}} = \sum_s w_j \mathbf{x}_j$ y $\widehat{\boldsymbol{\beta}} = (\widehat{B}_1, \cdots, \widehat{B}_p)^T$ es la solución de las ecuaciones de mínimos cuadrados ponderados de la muestra:

$$\left(\sum_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j \right) \widehat{\boldsymbol{\beta}} = \sum_s w_j \mathbf{x}_j y_i / c_j, \tag{2.12}$$

con constantes especificadas $c_j > 0$.

También es usual escribir $\widehat{Y}_{GR}$ en la forma expandida con pesos de diseño w_j cambiados a pesos "revisados" w_i^*. Tenemos

$$\widehat{Y}_{GR} = \sum_s w_j^* y_j \tag{2.13}$$

en notación operador, donde $w_j^* = w_j^*(s) = w_j(s) g_j(s)$ con

$$g_j(s) = 1 + (\mathbf{X} - \widehat{\mathbf{X}})^T \left(\sum_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j \right)^{-1} \mathbf{x}_j / c_j. \tag{2.14}$$

Los pesos revisados $w_j^*(s)$ son el producto de los pesos de diseño $w_j(s)$ y la estimación de pesos $g_j(s)$. La fórmula (2.13) muestra que los mismos pesos w_j^* son aplicados a todas la variables de interés como en el caso del estimador expandido. Esto asegura la consistencia de resultados cuando agregamos diferentes variables, es decir,

$$\widehat{Y}_{GR}(y_1) + \cdots + \widehat{Y}_{GR}(y_r) = \widehat{Y}_{GR}(y_1 + \cdots + y_r),$$

para diferentes variables $y_1, \cdots, y_r$ ligadas a cada elemento.

Una propiedad importante del estimador GREG es que asegura la consistencia con el total auxiliar conocido X en el sentido

$$\widehat{Y}_{GR}(\mathbf{x}) = \sum_s w_j^* \mathbf{x}_j = \mathbf{X}. \tag{2.15}$$

Esta propiedad no se conserva para el estimador expandido básico. Debido a la propiedad (2.15), el estimador GREG es llamado también estimador de calibración (Deville y Särndal, 1992). De hecho, de entre todos los estimadores de calibración de la forma $\sum_s b_j y_j$ con pesos b_j que satisfacen la restricción de calibración $\sum_s b_j \mathbf{x}_j = \mathbf{X}$, los pesos GREG w_j^* minimizan la distancia Chi-cuadrado, $\sum_s c_j (w_j - b_j)^2 / w_j$, entre los pesos básicos w_j y los pesos de calibración b_j. Así los pesos GREG w_j^* modifican los pesos de diseño tan poco como es posible sujeto a las restricciones de calibración.

El estimador GREG toma la forma simple cuando $c_j = \nu^T \mathbf{x}_j$ para todo $j \in U$ y algún vector columna constante ν. En este caso tenemos

$$\widehat{Y}_{GR} = \mathbf{X}^T \widehat{\boldsymbol{\beta}} = \sum_s \tilde{w}_j y_j, \tag{2.16}$$

porque $\sum_s w_j e_j(s) = \widehat{Y} - \widehat{\mathbf{X}}^T \widehat{\boldsymbol{\beta}} = 0$, donde $e_j(s) = e_j = y_j - \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}$ son los residuos muestrales y $\tilde{w}_j = \tilde{w}_j(s) = w_j(s) \tilde{g}_j(s)$ con

$$\tilde{g}_j(s) = \mathbf{X}^T \left(\sum_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j \right)^{-1} \mathbf{x}_j / c_j. \tag{2.17}$$

El estimador GREG cubre muchos estimadores usados como casos especiales. Por ejemplo, en el caso de una sola variable auxiliar conseguimos el estimador de razón

$$\widehat{Y}_R = \frac{\widehat{Y}}{\widehat{X}} X, \tag{2.18}$$

fijando $c_j = x_j$ en (2.17) y notando que $\tilde{g}_j(s) = X/\widehat{X}$. Si solamente se conoce el tamaño de la población N, fijamos $x_j = 1$ en (2.18) para que $X = N$ y $\widehat{X} = \widehat{N} = \sum_s w_j$. Si fijamos $\mathbf{x}_j = (1, x_j)^T$ y $c_j = 1$, entonces $\nu = (1, 0)^T$ y (2.16) se reduce al estimador de regresión lineal

$$\widehat{Y}_{LR} = \widehat{Y} + \widehat{\beta}_{LR}(X - \widehat{X}), \tag{2.19}$$

donde

$$\widehat{\beta}_{LR} = \widehat{\beta}_2 = \sum_s w_j (x_j - \overline{\widehat{X}})(y_j - \overline{\widehat{Y}}) / \sum_s w_j (x_j - \overline{\widehat{X}})^2,$$

con $\widehat{\overline{Y}} = \widehat{Y}/\widehat{N}$ y $\widehat{\overline{X}} = \widehat{X}/\widehat{N}$. El estimador GREG también cubre el estimador post estratificado como un caso especial. Supongamos una partición de U en G post-estratos U_g con tamaños poblacionales conocidos $N_g (g = 1, \cdots, G)$. Entonces fijamos $\mathbf{x}_j = (x_{1j}, \cdots, x_{Gj})^T$ con $x_{gj} = 1$ si $j \in U_g$ y $x_{gj} = 0$ en otro caso para que $\mathbf{X} = (N_1, \cdots, N_G)^T$. Notando que $1^T \mathbf{x}_j = 1$ para todo j tomamos $c_j = 1$ y $\nu = 1$ y el estimador GREG (2.16) se reduce al estimador post estratificado

$$\widehat{Y}_{PS} = \sum_g \frac{N_{\cdot g}}{\widehat{N}_{\cdot g}} \widehat{Y}_{\cdot g}, \tag{2.20}$$

donde $\widehat{N}_{\cdot g} = \sum_{s_{\cdot g}} w_j$ y $\widehat{Y}_{\cdot g} = \sum_{s_{\cdot g}} w_j y_j$, donde $s_{\cdot g}$ denota la muestra de elementos que pertenecen al estrato g.

La estimación GREG de Y se adapta fácilmente también a la estimación del total de un dominio Y_d. Se sigue de (2.13) que el estimador GREG de Y_d es

$$\widehat{Y}_{dGR} = \widehat{Y}_{GR}(y_d) = \sum_{j \in s_d} w_j^* y_j, \tag{2.21}$$

donde el total poblacional del vector auxiliar $\mathbf{x}$ es conocido. Se sigue de (2.21) que el estimador GREG también satisface la propiedad aditiva: $\widehat{Y}_{1GR} + \cdots + \widehat{Y}_{DGR} = \widehat{Y}_{GR}$. El estimador $\widehat{Y}_{dGR}$ es aproximadamente p-insesgado si el tamaño total de la muestra es grande, pero la p-consistencia requiere un tamaño esperado de la muestra de dominio también grande. El caso especial de estimador de razón (2.18) da

$$\widehat{Y}_{dR} = \frac{\widehat{Y}_d}{\widehat{X}} X,$$

cambiando y_d por $a_{dj} y_j$. Similarmente, un estimador post estratificado se obtiene de (2.20) cambiando y_{gj} por $a_{dj} y_{gj}$:

$$\widehat{Y}_{dPS} = \sum_g \frac{N_{\cdot g}}{\widehat{N}_{\cdot g}} \sum_{s_{dg}} w_j y_j,$$

donde s_{dg} es la muestra que cae en la (dg)-ésima celda de la clasificación cruzada de dominios y post-estratos.

Un estimador de la varianza por linealización de Taylor de $\widehat{Y}_{dGR}$ se obtiene simplemente de $v(y)$ cambiando y_d a $e_{dj} = y_{dj} - \mathbf{x}_{dj}^T \widehat{\boldsymbol{\beta}}(y_d)$, donde $\widehat{\boldsymbol{\beta}}(y_d)$ se obtiene de $\widehat{\boldsymbol{\beta}}(y)$ cambiando y_d por y_{dj}. Nótese que $e_{dj} = -\mathbf{x}_{dj}^T \widehat{\boldsymbol{\beta}}(y_d)$ si $j \in s$ y $j \notin U_d$.

Si contamos con información auxiliar específica de dominio, asumimos que los totales del dominio $\mathbf{X}_d = (X_{d1}, \cdots, X_{dp})^T = Y(\mathbf{x}_d)$ son conocidos, donde

$\mathbf{x}_{dj} = \mathbf{x}_j$ si $j \in U_d$ y $\mathbf{x}_{dj} = 0$ en otro caso. En este caso, un estimador GREG de Y_d está dado por

$$Y^*_{dGR} = \widehat{Y}_d + (\mathbf{X}_d - \widehat{\mathbf{X}}_d)^T \widehat{\boldsymbol{\beta}}_d, \tag{2.22}$$

donde $\widehat{\mathbf{X}}_d = \widehat{Y}(\mathbf{x}_d)$ y

$$\left(\sum_{j \in s} w_j \mathbf{x}_{dj} \mathbf{x}_{dj}^T / c_j \right) \widehat{\boldsymbol{\beta}}_d = \sum_{j \in s} w_j \mathbf{x}_{dj} y_{dj} / c_j.$$

Podemos escribir también (2.22) como

$$Y^*_{dGR} = \sum_{j \in s} w^*_{dj} y_{dj}, \tag{2.23}$$

donde $w^*_{dj} = w_j g^*_{dj}$ con

$$g^*_{dj} = 1 + (\mathbf{X}_d - \widehat{\mathbf{X}}_d)^T \left(\sum_{j \in s} w_j \mathbf{x}_{dj} \mathbf{x}_{dj}^T / c_j \right)^{-1} \mathbf{x}_{dj} / c_j.$$

Nótese que los pesos w^*_{dj} dependen ahora de d a diferencia de los pesos w^*_j. Por lo tanto, los estimadores Y^*_{dGR} no suman el total $\widehat{Y}_{GR}$. También, Y^*_{dGR} es aproximadamente p-sesgado a menos que el tamaño de la muestra del dominio sea grande.

En el caso especial de una sola variable auxiliar x con total de dominio conocido X_d, si sustituimos $c_j = x_j$ en (2.23) obtenemos el estimador de razón

$$Y^*_{dR} = \frac{\widehat{Y}_d}{\widehat{X}_d} X_d. \tag{2.24}$$

Si las cantidades N_{dg} de los post-estratos de dominio específico son conocidos, entonces un estimador de la cantidad postestratificada (PS/C) se obtiene del estimador GREG (2.23) como

$$Y^*_{dPS/C} = \sum_g \frac{N_{dg}}{\widehat{N}_{dg}} \widehat{Y}_{dg}, \tag{2.25}$$

donde $\widehat{N}_{dg} = \sum_{s_{dg}} w_j$ y $\widehat{Y}_{dg} = \sum_{s_{dg}} w_j y_j$. Si los totales de celdas X_{dg} de una variable auxiliar x son conocidos, entonces podemos usar un estimador de razón post estratificado (PS/R)

$$Y^*_{dPS/R} = \sum_g \frac{X_{dg}}{\widehat{X}_{dg}} \widehat{Y}_{dg}, \tag{2.26}$$

donde $\widehat{X}_{dg} = \sum_{s_{dg}} w_j x_j$.

Si el tamaño esperado del dominio es grande, entonces un estimador de la varianza linealizada de Taylor de Y^*_{dGR} se obtiene de $v(y)$ cambiando y_j por $e^*_{dj} = y_{dj} - \mathbf{x}^T_{dj}\widehat{\boldsymbol{\beta}}_d$. Nótese que $e^*_{dj} = 0$ si $j \in s$ y $j \notin U_d$ a diferencia de los residuos negativos grandes e_{dj} en el caso de $\widehat{Y}_{dGR}$. Así el estimador GREG de dominio $\widehat{Y}^*_{dGR}$, será más eficiente que $\widehat{Y}_{dGR}$, siempre que el tamaño esperado de la muestra específica de dominio sea grande.

Särndal et al. (1992), propusieron estimadores GREG de dominio, si se tiene disponible información auxiliar en forma del vector $\mathbf{x}_j$ de dimensión p. Como en los casos anteriores ajustamos un modelos de regresión que describa la presunta relación entre la variable de interés y las variables auxiliares. Los valores ajustados $\widehat{y}_j$ se usan para construir un estimador de regresión conveniente. Para el total de dominio (2.21) construimos un estimador como una suma de valores predichos por regresión más un término de ajuste que involucra los residuos de la regresión. Existen dos alternativas

$$\widehat{Y}_{dr} = \sum_{U_d} \widehat{y}_j + \frac{N_d}{\widehat{N}_d} \sum_{s_d} \frac{e_{js}}{\pi_j} \tag{2.27}$$

donde $\widehat{N}_d = \sum_{s_d} 1/\pi_j, N_d$ es conocido, $e_{js} = y_j - \widehat{y}_j = y_j - \mathbf{x}^T_j\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\beta}} = \left(\sum_s \frac{\mathbf{x}_j\mathbf{x}^T_j}{\sigma^2\pi_j}\right)^{-1} \sum_s \frac{\mathbf{x}_j y_j}{\sigma^2\pi_j}$ si se asume que las pendientes son iguales en todos los dominios, aunque es más realista asumir que las pendientes son distintas en cada dominio. Ante esto asumimos que $\widehat{\boldsymbol{\beta}}_d = \left(\sum_{s_d} \frac{\mathbf{x}_j\mathbf{x}^T_j}{\sigma^2_d\pi_j}\right)^{-1} \sum_{s_d} \frac{\mathbf{x}_j y_j}{\sigma^2_d\pi_j}$, para $j \in U_d$, $d = 1, \cdots, D$.

La segunda alternativa es considerar N_d desconocido, entonces, el modelo de regresión será de la forma

$$\widehat{Y}'_{dr} = \sum_{U_d} \widehat{y}_j + \sum_{s_d} \frac{e_{js}}{\pi_j}. \tag{2.28}$$

Si en los modelos (2.27) y (2.28), $\sum_{s_d} \frac{e_{js}}{\pi_j} = 0$, estaríamos ante un estimador sintético y es más conveniente usar (2.28) para la estimación, toda vez que s_d es aleatoria.

La estimación de la media de dominio se sigue inmediatamente sabiendo que $Y_{dr} = N_d\widehat{\overline{Y}}_{dr}$ y que N_d es desconocida.

2.1.3. Estimadores directos de dominio modificados

Consideramos estimadores directos que usan valores y de otro dominio, tales que, permanecen p-insesgados o aproximadamente p-insesgados cuando

el tamaño de muestra global aumenta. En particular, reemplazamos $\widehat{\boldsymbol{\beta}}_d$ en (2.22) por el coeficiente de regresión global $\widehat{\boldsymbol{\beta}}$, dado por (2.12), para conseguir

$$\tilde{Y}_{dGR} = \widehat{Y}_d + (\mathbf{X}_d - \widehat{\mathbf{X}}_d)^T \widehat{\boldsymbol{\beta}} = \sum_{j \in s} \tilde{w}_{dj} y_j \tag{2.29}$$

con

$$\tilde{w}_{dj} = w_j a_{dj} + (\mathbf{X}_d - \widehat{\mathbf{X}}_d) \left(\sum_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j \right)^{-1} (w_j \mathbf{x}_j / c_j),$$

donde a_{dj} es la variable indicadora del dominio. El estimador $\tilde{Y}_{dGR}$ es aproximadamente p-insesgado cuando el tamaño de muestra aumenta, aún cuando el tamaño del dominio sea pequeño. El estimador modificado (2.29) se puede ver también como un estimador de calibración $\sum_s \tilde{w}_{dj} y_j$ con pesos $b_{dj} = \tilde{w}_{dj}$ minimizando una distancia Chi-cuadrada $\sum_s c_j (w_j a_{dj} - b_{dj})^2 / w_j$ sujeta a las restricciones $\sum_s b_{dj} \mathbf{x}_j = \mathbf{X}_d$ (Singh y Mian, 1995).

Un estimador tipo razón de (2.29), en el caso de una única variable auxiliar x con total de dominio conocido X_d, está dado por

$$\tilde{Y}_{dR} = \widehat{Y}_d + \frac{\widehat{Y}}{\widehat{X}} (X_d - \widehat{X}_d) \tag{2.30}$$

Aunque el estimador directo modificado usa información de otro dominio para estimar los coeficientes de regresión, esto no incrementa el tamaño de muestra efectiva, a diferencia de los estimadores indirectos.

El estimador GREG modificado (2.30) puede expresarse como

$$\tilde{Y}_{dGR} = \mathbf{X}_d^T \widehat{\boldsymbol{\beta}} + \sum_{j \in s_d} w_j e_j. \tag{2.31}$$

El primer término $\mathbf{X}_d^T \widehat{\boldsymbol{\beta}}$ es el estimador de regresión sintético y el segundo término $\sum_{j \in s_d} w_j e_j$ corrige aproximadamente el p-sesgo del estimador sintético. Podemos mejorar $\tilde{Y}_{dGR}$ reemplazando el estimador expandido (2.31) por un estimador de razón (Särndal e Hidiroglou, 1989):

$$\widehat{E}_{dR} = N_d \left(\sum_{s_d} w_j e_j \right) \Big/ \left(\sum_{s_d} w_j \right); \tag{2.32}$$

Nótese que $\widehat{N}_d = \sum_{s_d} w_j$. El estimador resultante

$$\tilde{Y}_{dGR}(m) = \mathbf{X}_d^T \widehat{\boldsymbol{\beta}} + \widehat{E}_{dR}, \tag{2.33}$$

sin embargo, está afectado por el sesgo de razón cuando el tamaño de muestra del dominio es pequeño, a diferencia de $\tilde{Y}_{dGR}$.

Un estimador de la varianza por linealización de Taylor de $\tilde{Y}_{dGR}$ se obtiene de $v(y)$ cambiando y_j por $a_{dj}e_j$:

$$v_L(\tilde{Y}_{dGR}) = v(a_d e) \tag{2.34}$$

Este estimador de la varianza es válido incluso cuando el tamaño de la muestra del área pequeña es pequeña, con tal de que el tamaño de la muestra global sea grande.

2.1.4. Estimadores sintéticos

Los estimadores directos conducen a errores estándar grandes inaceptables debido a muestras excesivamente pequeñas de las áreas pequeñas de interés; de hecho, puede ocurrir que en la muestra no se incluya ninguna unidad de algún dominio pequeño de interés. Esto hace necesario hallar estimadores indirectos que incrementen el tamaño de muestra efectivo y así disminuir el error estándar. Este tipo de estimadores reciben el nombre de estimadores indirectos de dominio y están basados en modelos implícitos que proporcionan un enlace para áreas pequeñas relacionadas. Los estimadores que consideraremos bajo esa denominación incluye a los estimadores sintéticos, estimadores combinados y los estimadores de James-Stein.

Un estimador se llama sintético, si un estimador directo apropiado para un área grande, que cubre varias áreas pequeñas, se usa para obtener un estimador indirecto para un área pequeña bajo el supuesto que las áreas pequeñas tienen las mismas características que el área grande (Gonzalez, 1973).

Sin información auxiliar

Supongamos que no disponemos de información auxiliar poblacional y que estamos interesados en estimar la media del área pequeña $\overline{Y}_d$. En este caso un estimador sintético de $\overline{Y}_d$ está dado por

$$\widehat{\overline{Y}}_{dS} = \widehat{\overline{Y}} = \frac{\widehat{Y}}{\widehat{N}} \tag{2.35}$$

donde $\widehat{\overline{Y}}$ es el estimador directo de la media de la población global $\overline{Y}$, $\widehat{Y} = \sum_s w_j y_j$ y $\widehat{N} = \sum_s w_j$. El sesgo bajo el diseño de $\widehat{\overline{Y}}_{dS}$ es aproximadamente igual a $\overline{Y} - \overline{Y}_d$ el cual es relativamente pequeño para $\overline{Y}_d$ si $\overline{Y}_d \approx \overline{Y}$. En el último modelo implícito, en el cual se satisface que la media del área pequeña es aproximadamente igual a la media global, el estimador sintético será muy eficiente porque su error cuadrático medio (ECM) será pequeño. Por otro lado, puede ocurrir que sea sesgado si existen fuertes efectos individuales por áreas,

los cuales, a su vez, pueden conducir a un ECM grande. La condición $\overline{Y}_d \approx \overline{Y}$ puede relajarse a $\overline{Y}_d \approx \overline{Y}(r)$ donde $\overline{Y}(r)$ es la media de un área grande (región) que cubre el área pequeña. En este caso, usamos $\widehat{\overline{Y}}_{dS} = \widehat{\overline{Y}}(r)$ donde $\widehat{\overline{Y}}(r)$ es el estimador directo regional. El sesgo bajo el diseño de $\widehat{\overline{Y}}(r)$ es aproximadamente igual $\overline{Y}(r) - \overline{Y}_d$ el cual es relativamente despreciable para $\overline{Y}_d$.

Con información auxiliar

Si tenemos disponible información auxiliar específica del dominio en forma de totales $\mathbf{X}_d$, entonces el estimador de regresión sintético $\mathbf{X}_d^T\widehat{\boldsymbol{\beta}}$, puede usarse como un estimador del total de dominio Y_d:

$$\widehat{Y}_{dGRS} = \mathbf{X}_d^T\widehat{\boldsymbol{\beta}}, \tag{2.36}$$

donde $\widehat{\boldsymbol{\beta}}$ está dado por (2.12). El sesgo bajo el diseño de $\widehat{Y}_{dGRS}$ es aproximadamente igual a $\mathbf{X}_d^T\widehat{\boldsymbol{\beta}} - Y_d$, donde $\widehat{\boldsymbol{\beta}} = (\sum_U \mathbf{x}_j\mathbf{x}_j^T/c_j)^{-1}(\sum_U \mathbf{x}_j y_j/c_j)$ es el coeficiente de regresión de la población. Este sesgo es relativamente pequeño para Y_d si el coeficiente de regresión del dominio $\widehat{\boldsymbol{\beta}}_d = (\sum_{U_d} \mathbf{x}_j\mathbf{x}_j^T/c_j)^{-1}(\sum_{U_d} \mathbf{x}_j y_j/c_j)$ está próximo a $\widehat{\boldsymbol{\beta}}$ y $Y_d = \mathbf{X}_d^T\widehat{\boldsymbol{\beta}}_d$. La condición anterior $Y_d = \mathbf{X}_d^T\widehat{\boldsymbol{\beta}}_d$ se satisface si $c_j = \nu^T\mathbf{x}_j$ para algún vector columna constante ν. Así el estimador de regresión sintético será muy eficiente cuando el área pequeña d no muestre efectos individuales fuertes con respecto al coeficiente de regresión. Los estimadores sintético s $\widehat{Y}_{dGRS}$ tendrán suma igual al estimador directo GREG $\widehat{Y}_{GR} = \mathbf{X}^T\widehat{\boldsymbol{\beta}}$ cuando $c_j = \nu^T\mathbf{x}_j$.

Un caso especial de (2.36) es el estimador de razón sintético en el caso de una única variable auxiliar x. Se obtiene haciendo $c_j = x_j$ en (2.36) y está dado por

$$\widehat{Y}_{dRS} = X_d\frac{\widehat{Y}}{\widehat{X}}.$$

El sesgo bajo el diseño de $\widehat{Y}_{dRS}$ respecto a Y_d es aproximadamente igual a $X_d(R - R_d)/Y_d$ que será pequeño si la razón de área específica $R_d = Y_d/X_d$ está próxima a la razón global $R = Y/X$. El estimador de razón sintético $\widehat{Y}_{dRS}$ cumple la propiedad aditiva, es decir, la suma de los totales estimados de dominio es igual al estimador directo de razón $\widehat{Y}_R = (\widehat{Y}/\widehat{X})/X$.

Si los tamaños N_{dg} de los post-estratos de dominio son conocidos para los post-estratos $g = 1, \cdots, G$, entonces un estimador sintético se obtiene como un caso especial de $\widehat{Y}_{dGRS}$ haciendo $\mathbf{x}_j = (x_{1j}, \cdots, x_{Gj})$ con $x_{gj} = 1$ si $j \in U_g$ y $x_{gj} = 0$ en otro caso, es decir

$$\widehat{Y}_{dS/C} = \sum_g N_{dg}\frac{\widehat{Y}_{\cdot g}}{\widehat{N}_{\cdot g}}, \tag{2.37}$$

donde $\widehat{Y}_{\cdot g}$ y $\widehat{N}_{\cdot g}$ son estimadores del total del post-estrato $Y_{\cdot g}$ y el tamaño $N_{\cdot g}$ (National Center for Health Statistics, 1968). Si $y_j = 1$ ó 0, entonces se obtiene de (2.37) un estimador sintético de la proporción P_d como

$$\widehat{P}_{dS/C} = \left(\sum_g N_{dg}\widehat{P}_{\cdot g}\right) \Big/ \left(\sum_g N_{dg}\right), \tag{2.38}$$

donde $\widehat{P}_{\cdot g}$ es el estimador directo de la proporción $P_{\cdot g}$ del g-ésimo post-estrato.

Más generalmente, un estimador de razón sintético se obtiene de los totales de celdas X_{dg} de las variables auxiliares conocidas. Haciendo $x_{gj} = x_j$ si $j \in U_g$ y $x_{gj} = 0$ en otro caso en (2.36), conseguimos

$$\widehat{Y}_{dS/R} = \sum_g X_{dg}\frac{\widehat{Y}_{\cdot g}}{\widehat{X}_{\cdot g}}, \tag{2.39}$$

donde $\widehat{X}_{\cdot g} = \sum_{s_{\cdot g}} w_j x_j$ (Ghangurde and Singh, 1977). El sesgo bajo el diseño de $\widehat{Y}_{dS/R}$ es aproximadamente igual a $\sum_g X_{dg}(Y_{\cdot g}/X_{\cdot g} - Y_{dg}/X_{dg}) = \sum_g X_{dg}(R_{\cdot g} - R_{dg})$.

Se han obtenido estimadores sintéticos alternativos, en el contexto de la post-estratificación, cambiando $\widehat{N}_{\cdot g}$ por $N_{\cdot g}$ en (2.37) y $\widehat{X}_{\cdot g}$ por $X_{\cdot g}$ en (2.39)

$$\tilde{Y}_{dS/C} = \sum_g N_{dg}\frac{\widehat{Y}_{\cdot g}}{N_{\cdot g}} \tag{2.40}$$

y

$$\tilde{Y}_{dS/R} = \sum_g X_{dg}\frac{\widehat{Y}_{\cdot g}}{X_{\cdot g}} \tag{2.41}$$

(Purcell y Linacre, 1976; Singh y Tessier, 1976). En muestras grandes, los estimadores sintéticos alternativos son menos eficientes que (2.37) y (2.39) cuando $\widehat{Y}_{\cdot g}$ y $\widehat{X}_{\cdot g}$ (o $\widehat{N}_{\cdot g}$) están correlados positivamente, como en el caso de estimador de razón $\widehat{Y}/\widehat{X}$ comparado con el estimador expandido de $\widehat{Y}$.

Es interesante notar que el método sintético puede usarse también cuando el muestreo no es complicado. Supongamos que Y es conocido, pero no Y_d, de alguna fuente administrativa y que X_d y X son también conocidos. Entonces un estimador sintético de Y_d puede tomarse como $(X_d/X)Y$, cuyo sesgo relativo para Y_d será pequeño cuando $R_d \approx R$. Este estimador no es un estimador en el sentido usual de una cantidad aleatoria.

Estimador sintético de regresión-ajustado

El estimador sintético de regresión ajustado (Levy, 1971), intenta responder a la variación local combinando covariables de área específica con un estimador sintético. Las covariables $\mathbf{z}_d$ para modelar el sesgo relativo $B_d = (\overline{Y}_d - \widehat{\overline{Y}}_{dS})/\widehat{\overline{Y}}_{dS}$ asociado con el estimador sintético $\widehat{\overline{Y}}_{dS}$ de la media $\overline{Y}_d$:

$$B_d = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}^T \mathbf{z}_d + \varepsilon_d,$$

donde los $\boldsymbol{\gamma}$ son los parámetros de regresión y ε_d es un error aleatorio. Dado que B_d es no observable, el modelo de regresión se ajusta por mínimos cuadrados para estimar el valor del sesgo $\widehat{B}_a = (\widehat{\overline{Y}}_a - \widehat{\overline{Y}}_{aS})/\widehat{\overline{Y}}_{aS}$ para áreas grandes $(a = 1, \cdots, A)$ usando un estimador directo apropiado $\widehat{\overline{Y}}_a$ y un estimador sintético $\widehat{\overline{Y}}_{aS}$. Denotando los estimadores de mínimos cuadrados resultantes como $\widehat{\boldsymbol{\gamma}}_0$ y $\widehat{\boldsymbol{\gamma}}$, estimamos B_d como $\widehat{\boldsymbol{\gamma}}_0 + \widehat{\boldsymbol{\gamma}}^T \mathbf{z}_d$ que, a su vez, lleva al estimador de regresión ajustado de $\overline{Y}_d$:

$$\widehat{\overline{Y}}_{dS}(a) = \widehat{\overline{Y}}_{dS}(1 + \widehat{\boldsymbol{\gamma}}_0 + \widehat{\boldsymbol{\gamma}}^T \mathbf{z}_d). \tag{2.42}$$

Nótese que $\widehat{B}_a$ es un estimador apropiado de $B_a = (\overline{Y}_a - \widehat{\overline{Y}}_{aS})/\widehat{\overline{Y}}_{aS}$.

Estimación del ECM de un estimador sintético

La varianza bajo el diseño de un estimador sintético $\widehat{Y}_{dS}$ será pequeña con respecto a la varianza bajo el diseño de un estimador directo $\widehat{Y}_d$ porque depende solamente de la presición del estimador directo a un nivel de área grande. La varianza bajo el diseño se estima sin dificultad usando los métodos estándar basados en el diseño, pero es más difícil estimar el ECM de $\widehat{Y}_{dS}$. Por ejemplo, la varianza bajo el diseño del estimador sintético de razón (2.39) o el estimador sintético de frecuencias (2.37) pueden estimarse usando el método de linealización de Taylor, la varianza bajo el diseño del estimador general de regresión sintética $\widehat{Y}_{dGRS} = \mathbf{X}^T \widehat{\boldsymbol{\beta}}$ puede estimarse usando el resultado de Fuller (1975), sobre la matriz de covarianzas de una muestra grande de $\widehat{\boldsymbol{\beta}}$ o usando un método de remuestreo tal como jackknife.

Un estimador aproximadamente insesgado bajo el diseño del ECM de $\widehat{Y}_{dS}$ puede obtenerse usando un estimador directo insesgado bajo el diseño $\widehat{Y}_d$. Tenemos

$$\begin{aligned} \mathrm{ECM}_p(\widehat{Y}_{dS}) &= E_p(\widehat{Y}_{dS} - Y_d)^2 \\ &= E_p(\widehat{Y}_{dS} - \widehat{Y}_d + \widehat{Y}_d - Y_d)^2 \\ &= E_p(\widehat{Y}_{dS} - \widehat{Y}_d)^2 - V_p(\widehat{Y}_d) + 2\mathrm{Cov}(\widehat{Y}_{dS}, \widehat{Y}_d) \end{aligned}$$

$$= E_p(\widehat{Y}_{dS} - \widehat{Y}_d)^2 - V_p(\widehat{Y}_{dS} - \widehat{Y}_d) + V_p(\widehat{Y}_{dS}). \tag{2.43}$$

Ahora se sigue de (2.43) que un estimador aproximadamente insesgado bajo el diseño de $\mathrm{ECM}_p(\widehat{Y}_{dS})$ es

$$\mathrm{ecm}(\widehat{Y}_{ds}) = (\widehat{Y}_{ds} - \widehat{Y}_d)^2 - v(\widehat{Y}_{ds} - \widehat{Y}_d) + v(\widehat{Y}_{ds}) \tag{2.44}$$

donde $v(\cdot)$ es un estimador insesgado bajo el diseño de $V_p(\cdot)$; por ejemplo, un estimador jackknife. El estimador (2.44), sin embargo, puede ser muy inestable y puede tomar valores negativos. Por consiguiente, es costumbre promediar el ECM del estimador sobre áreas pequeñas ($d = 1, \cdots, D$) que pertenecen a un área grande para conseguir un estimador estable (Gonzalez y Waksberg, 1973). Sea $\widehat{\overline{Y}}_{dS} = \widehat{Y}_{dS}/N_d$ tal que $\mathrm{ecm}(\widehat{\overline{Y}}_{dS}) = \mathrm{ecm}(\widehat{Y}_{dS})/N_d^2$. Tomamos el promedio de $\mathrm{ecm}(\widehat{\overline{Y}}_{dS})$ sobre d como un estimador de $\mathrm{ECM}(\widehat{\overline{Y}}_{dS})$ para que consigamos $\mathrm{ecm}_a(\widehat{Y}_{dS}) = N_d^2 \mathrm{ecm}_a(\widehat{\overline{Y}}_{dS})$ como un estimador de $\mathrm{ECM}(\widehat{Y}_{dS})$, donde

$$\mathrm{ecm}_a(\widehat{\overline{Y}}_{dS}) = \frac{1}{D}\sum_d \frac{1}{N_d^2}(\widehat{Y}_{dS} - \widehat{Y}_d)^2 - \frac{1}{D}\sum_d \frac{1}{N_d^2} v(\widehat{Y}_{dS} - \widehat{Y}_d) + \frac{1}{D}\sum_d \frac{1}{N_d^2} v(\widehat{Y}_{dS}). \tag{2.45}$$

Pero tal medida global de incertidumbre puede ser engañosa dado que se refiere al promedio del ECM en lugar del ECM de área específica.

Una buena aproximación de (2.44) está dada por

$$\mathrm{ecm}(\widehat{\overline{Y}}_{dS}) \approx (\widehat{Y}_{dS} - \widehat{Y}_d)^2 - v(\widehat{Y}_d), \tag{2.46}$$

notando que la varianza del estimador sintético $\widehat{\overline{Y}}_{dS}$ es pequeño con respecto a la varianza del estimador directo $\widehat{\overline{Y}}_d$. Usando la aproximación (2.46)

$$\mathrm{ecm}_a(\widehat{\overline{Y}}_{dS}) \approx \frac{1}{D}\sum_d \frac{1}{N_d^2}(\widehat{Y}_{dS} - \widehat{Y}_d)^2 - \frac{1}{D}\sum_d \frac{1}{N_d^2} v(\widehat{Y}_d). \tag{2.47}$$

Marker(1995), propuso un método simple para obtener un estimador del ECM de $\widehat{Y}_{dS}$ para área específica. Usó el supuesto que el cuadrado del sesgo bajo el diseño $B_p^2(\widehat{\overline{Y}}_{dS})$ es aproximadamente igual al promedio del cuadrado del sesgo:

$$B_p^2(\widehat{\overline{Y}}_{dS}) \approx \frac{1}{D}\sum_d B_p^2(\widehat{\overline{Y}}_{dS}) = B_a^2(\widehat{\overline{Y}}_{dS}) \tag{2.48}$$

el promedio del cuadrado del sesgo puede ser estimado como

$$b_a^2(\widehat{\overline{Y}}_{dS}) = \mathrm{ECM}_a(\widehat{\overline{Y}}_{dS}) - \frac{1}{D}\sum_d v(\widehat{\overline{Y}}_{dS}), \tag{2.49}$$

notando que el promedio del ECM = varianza promedio + promedio $(sesgo)^2$. El estimador de la varianza $v(\widehat{\overline{Y}}_{dS}) = v(\widehat{Y}_{dS})/N_d^2$ se obtiene realmente usando métodos tradicionales como indicamos antes. Ahora se sigue bajo el supuesto (2.48) que el $\text{ECM}_p(Y_{dS})$ puede obtenerse como

$$\text{ecm}_M(\widehat{Y}_{dS}) = v(\widehat{Y}_{dS}) + N_d^2 b_a^2(\widehat{\overline{Y}}_{dS}), \tag{2.50}$$

la cual es específica de área si $v(\widehat{Y}_{dS})$ depende del área. Sin embargo, el supuesto (2.48) no se satisface para áreas que exhiben efectos individuales fuertes. No obstante, (2.50) es una mejora sobre la medida global (2.47), con tal de que el término de la varianza domine el término del sesgo en (2.50). Nótese que ambos $\text{ECM}_M(\widehat{Y}_{dS})$ y $\text{ECM}_a(\widehat{Y}_{dS})$ requieren el tamaño del dominio, N_d.

2.1.5. Estimadores combinados

Una manera natural de balancear el sesgo potencial de un estimador sintético, digamos $\widehat{Y}_{d2}$, frente a la inestabilidad de un estimador directo, digamos $\widehat{Y}_{d1}$, consiste en ponderar los pesos de $\widehat{Y}_{d1}$ y $\widehat{Y}_{d2}$. Tal estimador combinado del total Y_d del área pequeña puede escribirse como

$$\widehat{Y}_{dC} = \phi_d \widehat{Y}_{d1} + (1 - \phi_d)\widehat{Y}_{d2}, \tag{2.51}$$

para una elección apropiada del peso $\phi_d (0 \leq \phi_d \leq 1)$. Muchos de los estimadores propuestos en la literatura, bajo el diseño y bajo el modelo, tienen la forma combinada (2.51).

Estimador combinado Óptimo

El ECM bajo el diseño de un estimador combinado está dado por

$$\begin{aligned}\text{ECM}_p(\widehat{Y}_{dC}) = \phi_d^2 \text{ECM}_p(\widehat{Y}_{d1}) + (1-\phi_d)^2 \text{ECM}_p \widehat{Y}_{d2} \\ +2\phi_d(1-\phi_d)E_p(\widehat{Y}_{d1} - Y_d)(\widehat{Y}_{d2} - Y_d).\end{aligned} \tag{2.52}$$

Minimizando (2.44) con respecto a ϕ_d, conseguimos el peso óptimo ϕ_d como

$$\begin{aligned}\phi_d^* &= \frac{\text{ECM}_p[\widehat{Y}_{d2} - E_p(\widehat{Y}_{d1} - Y_d)(\widehat{Y}_{d2} - Y_d)]}{\text{ECM}_p(\widehat{Y}_{d1}) + \text{ECM}_p(\widehat{Y}_{d2}) - 2E_p(\widehat{Y}_{d1} - Y_d)(\widehat{Y}_{d2} - Y_d)} \\ &\approx \frac{\text{ECM}_p(\widehat{Y}_{d2})}{\text{ECM}_p(\widehat{Y}_{d1}) + \text{ECM}_p(\widehat{Y}_{d2})},\end{aligned} \tag{2.53}$$

asumiendo que el término de la covarianzas $E_p(\widehat{Y}_{d1} - Y_d)(\widehat{Y}_{d2} - Y_d)$ es relativamente pequeño respecto a $\mathrm{ECM}_p(\widehat{Y}_{d2})$. La aproximación óptima ϕ_d^*, dada por (2.53), está en el intervalo $[0, 1]$.

Los pesos óptimos aproximados ϕ_d^* dependen solamente de la razón de los errores cuadráticos medios:

$$\phi_d^* = 1/(1 + F_d), \tag{2.54}$$

donde $F_d = \mathrm{ECM}_p(\widehat{Y}_{d1})/\mathrm{ECM}_p(\widehat{Y}_{d2})$. Además, el ECM de $\widehat{Y}_{dC}$ con pesos óptimos ϕ_d^* se reduce a

$$\mathrm{ECM}_p^*(\widehat{Y}_{dC}) = \phi_d^* \mathrm{ECM}_p(\widehat{Y}_{d1}) = (1 - \phi_d^*)\mathrm{ECM}_p(\widehat{Y}_{d2}). \tag{2.55}$$

Se sigue ahora de (2.55) que la reducción en ECM alcanzada por el estimador óptimo, relativo al más pequeño de los ECM de las componentes del estimador está dado por ϕ_d^* si $0 \leq \phi_d^* \leq 1/2$, e igual a $1 - \phi_d^*$ si $1/2 \leq \phi_d^* \leq 1$. Así la reducción máxima del 50 % se obtiene cuando $\phi_d^* = 1/2$ o $F_d = 1$.

La razón de $\mathrm{ECM}_p(\widehat{Y}_{dC})$ con un peso fijo ϕ_d y $\mathrm{ECM}_p(\widehat{Y}_{d2})$ puede expresarse como

$$\frac{\mathrm{ECM}_p(\widehat{Y}_{dC})}{\mathrm{ECM}_p(\widehat{Y}_{d2})} = (F_d + 1)\phi_d^2 - 2\phi_d + 1. \tag{2.56}$$

Schaible (1978), estudió el comportamiento de la razón del ECM (2.56) como una función de ϕ_d para valores seleccionados de $F_d(= 1, 2, 6)$. Sus resultados sugieren que desviaciones regulares de los pesos óptimos ϕ_d^* no producen incrementos significativos en el ECM del estimador combinado, es decir, la curva (2.56) es bastante monótona en el entorno del peso óptimo. Más aún, la reducción en ECM y el rango de ϕ_d para el cual el estimador combinado tiene un ECM más pequeño que cualquier componente del estimador, dependiendo ambos del tamaño de F_d. Cuando F_d es cercano a uno, conseguimos la mayor ventaja en términos de ambas situaciones.

Es fácil mostrar que $\widehat{Y}_{dC}$ es mejor que cualquier estimador combinado en términos del ECM cuando $\max(0, 2\phi_d^* - 1) \leq \phi_d \leq \min(2\phi_d^*, 1)$. El último intervalo se reduce al rango completo $0 \leq \phi_d \leq 1$ cuando $F_d = 1$, y se reduce más cuando F_d se desvía de uno. El peso óptimo ϕ_d^* será cercano a cero o uno cuando una de las componentes del estimador tenga más grande ECM que las otras, es decir, cuando F_d sea grande o pequeña. En este caso, el estimador con ECM grande proporciona poca información y por consiguiente es mejor usar, preferentemente, la componente del estimador con ECM pequeño, al estimador combinado.

En la práctica, usamos un supuesto a priori del valor óptimo de ϕ_d^* o una estimación de ϕ_d^* de los datos muestrales. Asumiendo que el estimador directo $\widehat{Y}_{d1}$

es p-insesgado o aproximadamente p-insesgado cuando el tamaño de la muestra global crece, podemos estimar el peso óptimo (2.53) usando (2.46). Sustituimos el estimador ECM($\widehat{Y}_{d2}$) dado por (2.46) para el numerador ECM$_p(\widehat{Y}_{d2})$ y $(\widehat{Y}_{d2} - \widehat{Y}_{d1})^2$ para el denominador ECM$_p(\widehat{Y}_{d1})$ + ECM$_p(\widehat{Y}_{d2})$:

$$\widehat{\phi}_d^* = \frac{ECM(\widehat{Y}_{d2})}{(\widehat{Y}_{d2} - \widehat{Y}_{d1})^2}. \tag{2.57}$$

Pero este estimador de ϕ_d^* puede ser muy inestable. Una manera de superar esta dificultad es promediar los pesos estimados $\widehat{\phi}_d^*$ sobre varias variables o áreas "similares" o ambos. El estimador combinado resultante debe funcionar bien en vista de la insensibilidad a las desviaciones de los pesos óptimos.

Estimadores dependientes del tamaño de muestra

Los estimadores dependientes del tamaño de la muestra (SSD) son estimadores combinados con pesos simples ϕ_d que dependen solamente de las frecuencias de dominios $\widehat{N}_d$ y N_d o los totales de dominios $\widehat{X}_d$ y X_d de una variable auxiliar x. Estos estimadores fueron originalmente diseñados para manejar dominios para los cuales el tamaño de muestra esperada es bastante grande para hacer que el estimador directo satisfaga los requerimientos de bondad cuando el tamaño de muestra realizado exceda el tamaño de muestra esperado (Drew, Singh y Couldhry, 1982).

Drew, Singh y Couldhry (1982), propusieron un estimador SSD que usa los pesos

$$\phi_d(S1) = \begin{cases} 1, & \text{si} \quad \widehat{N}_d/N_d \geq \delta; \\ \widehat{N}_d/(\delta N_d), & \text{si} \quad \widehat{N}_d/N_d < \delta. \end{cases} \tag{2.58}$$

donde $\widehat{N}_d = \sum_{s_d} w_j$ es el estimador directo expandido de N_d y δ se escoge subjetivamente para controlar la distribución del estimador sintético. La forma de $\widehat{N}_d$ se incrementa con el tamaño de la muestra del dominio. Otra selección de ϕ_d se obtiene sustituyendo $\widehat{X}_d/X_d$ por $\widehat{N}_d/N_d$ en (2.58). Bajo esta selección, Drew et al. (1982), usaron el estimador de razón post-estratificado (2.26) como el estimador directo $\widehat{Y}_{d1}$ y el estimador de razón sintético (2.39) como el estimador $\widehat{Y}_{d2}$. El estimador directo (2.26), sin embargo, está afectado por el sesgo de la razón a menos que el tamaño de la muestra del dominio no sea pequeña. Para evitar el sesgo de la razón, podemos usar el estimador directo modificado $\widehat{Y}_d + \sum_g (X_{dg} - \widehat{X}_{dg})(\widehat{Y}_g/\widehat{X}_g)$ cuyo p-sesgo tiende a cero cuando el tamaño de la muestra global se incrementa, incluso si el tamaño de la muestra del dominio es pequeña. Generalmente, podemos usar el estimador GREG

modificado $\widehat{Y}_d + (\mathbf{X}_d - \widehat{\mathbf{X}}_d)^T \widehat{\boldsymbol{\beta}}$ como el estimador directo y el estimador sintético de regresión $\mathbf{X}_d^T \widehat{\boldsymbol{\beta}}$ así como el estimador sintético junto con $\phi_d(S1)$. Una alternativa de uso general de δ es $\delta = 1$.

Särndal e Hidiroglou (1989), propusieron el "estimador de regresión amortiguado" el cual se obtiene del estimador GREG modificado en la fórmula (2.33) amortiguando el efecto del componente directo $\sum_{s_d} w_j e_j$ cuando $\widehat{N}_d < N_d$.

Está dado por

$$\widehat{Y}_{dDR} = \mathbf{X}_d^T \widehat{\boldsymbol{\beta}} + (\widehat{N}_d / N_d)^{H-1} \sum_{s_d} w_j e_j \tag{2.59}$$

con $H = 0$ si $\widehat{N}_d \geq N_d$ y $H = h$ si $\widehat{N}_d < N_d$, donde h es una constante convenientemente elegida. Este estimador puede escribirse como un estimador combinado con el estimador GREG modificado mejorado $\mathbf{X}_d^T \widehat{\boldsymbol{\beta}} + (N_d / \widehat{N}_d) \sum_{s_d} w_j e_j$ como el estimador directo y $\mathbf{X}_d^T \widehat{\boldsymbol{\beta}}$ como el estimador sintético junto con los pesos

$$\phi_d(S2) = \begin{cases} 1, & \text{si} \quad \widehat{N}_d / N_d \geq 1; \\ (\widehat{N}_d / N_d)^h, & \text{si} \quad \widehat{N}_d / N_d < 1. \end{cases} \tag{2.60}$$

Una elección de h de uso general es $h = 2$.

Para estudiar la naturaleza de los pesos $\phi_d(S1)$, consideremos el caso especial de muestreo aleatorio simple. En este caso, $\widehat{N}_d = N(n_d/n)$. Tomando $\delta = 1$ en (2.58), se sigue ahora que $\phi_d(S1) = 1$ si n_d es por lo menos tan grande como $E(n_d) = n(N_d/N)$. Por tanto, el estimador SSD no puede usar información auxiliar de otros dominios incluso cuando $E(n_d)$ no sea bastante grande para hacer fiable el estimador directo. Por otro lado, cuando $\widehat{N}_d < N_d$, el peso $\phi_d(S1) = \widehat{N}_d / N_d$ decrece cuando n_d decrece. Como resultado, se da mayor peso al componente sintético cuando n_d decrece. Así en el caso $\widehat{N}_d < N_d$ el peso $\phi_d(S1)$ se comporta bien, a diferencia del caso en que $\widehat{N}_d \geq N_d$. Similar comentario aplica para el estimador SSD basado en el peso $\phi_d(S2)$. Otra desventaja de estimadores SSD es que los pesos no toman en cuenta el tamaño relativo de la variación entre áreas a la variación en las áreas para la característica de interés. Es decir, todas las características reciben el mismo peso sin tener en cuenta sus diferencias respecto a la heterogeneidad entre áreas.

Los estimadores generales SSD proporcionan consistencia cuando se agregan otras características porque se usan los mismos pesos. Sin embargo, su suma no iguala a la de un estimador directo para un nivel de área grande. Una adaptación de razón simple da

$$\widehat{Y}_{dC}(a) = \frac{\widehat{Y}_{dC}}{\sum_d \widehat{Y}_{dC}} \widehat{Y}_{GR}. \tag{2.61}$$

El estimador ajustado $\widehat{Y}_{dC}(a)$ satisface la propiedad aditiva para el estimador director $\widehat{Y}_{GR}$ para un nivel de área grande.

Los estimadores SSD con pesos $\phi_d(S1)$ pueden ser vistos como estimadores de calibración $\sum_s w^*_{dj} y_j$ minimizando la distancia chi-cuadrada

$$\sum_s \sum_{j \in s} c_j [w_j a_{dj} \phi_d(S1) - b^*_{dj}]^2 / w_j \tag{2.62}$$

sujeta a las restricciones $\sum_{j \in s} b^*_{dj} \mathbf{x}_j = \mathbf{X}_d$, $d = 1, \cdots, D$, donde a_{dj} es la variable indicadora de dominio y c_j es una constante especificada, $j \in s$, es decir, el b^*_{dj} "óptimo" es igual a w^*_{dj}. Usando esta medida de distancia, calibramos los pesos amortiguados $w_j a_{dj} \phi_d(S1)$ en lugar de los pesos originales $w_j a_{dj}$. Singh y Mian (1995), usaron el enfoque de calibración para tomar en cuenta diferentes conjuntos de restricciones simultáneamente. Por ejemplo, la restricción aditiva $\sum_d \sum_{j \in s} b^*_{dj} y_j = \widehat{Y}_{GR}$ puede ser introducida desde el principio con las restricciones previas $\sum_{j \in s} b^*_{dj} \mathbf{x}_j = \mathbf{X}_d$, $d = 1, \cdots, D$. Nótese que los pesos de calibración w^*_{dj} para todas las áreas pequeñas bajo consideración son obtenidos simultáneamente usando este enfoque.

La estimación del ECM de los estimadores SSD se encuentra con dificultades similares a aquellas para el estimador sintético y el estimador combinado óptimo. Una aproximación ad hoc para la estimación de la varianza es usar el estimador de la varianza del estimador directo modificado $\widehat{Y}_d + (\mathbf{X}_d - \widehat{\mathbf{X}}_d)^T \widehat{\boldsymbol{\beta}}$, es decir, $v(a_d e)$, como un sobreestimador de la varianza verdadera del estimador SSD usando cualquiera de $\phi_d(S1)$ o $\phi_d(S2)$ como los pesos ligados al estimador directo (Särndal e Hidiroglou, 1989). Otro enfoque es tratar los pesos $\phi_d(S1)$ como fijos y estimar la varianza como $(\phi_d(S1))^2 v(a_d e)$ notando que la varianza de la componente sintética $\mathbf{X}_d^T \widehat{\boldsymbol{\beta}}$ es relativamente pequeña respecto a la varianza de la componente directa. Este estimador de la varianza sobreestima la varianza verdadera, a diferencia de $v(a_d e)$. Pueden usarse también, sin dificultad, métodos de remuestreo, tales como jackknife, para obtener un estimador de la varianza.

2.1.6. Método de James-Stein

Peso común

Otra aproximación para la estimación combinada consiste en usar un peso común, $\phi_d = \phi$, y entonces minimizar el ECM total, $\sum_d \mathrm{ECM}_p(\widehat{Y}_{dC})$, con respecto a ϕ (Purcell y Kish, 1979). Esto garantiza una buena estimación para el grupo de áreas pequeñas en conjunto pero no necesariamente para cada una de las áreas pequeñas en el grupo.

Tenemos

$$\sum_d \mathrm{ECM}_p(\widehat{Y}_{dC}) \approx \phi^2 \sum_d \mathrm{ECM}_p(\widehat{Y}_{d1}) + (1-\phi)^2 \sum_d \mathrm{ECM}_p(\widehat{Y}_{d2}). \tag{2.63}$$

Minimizando (2.63) con respecto a ϕ da el valor óptimo

$$\phi^* = \frac{\sum_d \mathrm{ECM}_p(\widehat{Y}_{d2})}{\sum_d [\mathrm{ECM}_p(\widehat{Y}_{d1}) + \mathrm{ECM}_p(\widehat{Y}_{d2})]}. \tag{2.64}$$

Suponiendo que tomamos $\widehat{Y}_{d1}$ como el estimador directo de $\widehat{Y}_d$ y $\widehat{Y}_{d2}$ como el estimador sintético $\widehat{Y}_{dS}$. Se sigue ahora de (2.46) que ϕ^* puede estimarse como

$$\widehat{\phi}^* = \frac{\sum_d [(\widehat{Y}_{dS} - \widehat{Y}_d)^2 - v(\widehat{Y}_d)]}{\sum_d (\widehat{Y}_{dS} - \widehat{Y}_d)^2} = 1 - \frac{\sum_d v(\widehat{Y}_d)}{\sum_d (\widehat{Y}_{ds} - \widehat{Y}_d)^2}. \tag{2.65}$$

El estimador $\widehat{\phi}^*$ es realmente fiable, a diferencia de $\widehat{\phi}_d^{\,*}$ dado por (2.57), porque agrupamos sobre varias áreas pequeñas. Sin embargo, el uso de pesos comunes puede no ser razonable si las varianzas individuales $V(\widehat{Y}_d)$, varían considerablemente.

El estimador combinado basado en $\widehat{\phi}_d^{\,*}$ es similar al estimador de James-Stein (J-S), quien ha atraído mucho la atención en la corriente dominante de la literatura estadística. Ver Efron y Morris (1972a, 1973, 1975) y Brandwein y Strawderman (1990), para un tratamiento más extenso de este método.

Varianzas iguales $\psi_d = \psi$

En el caso especial de igual varianzas de muestreo, $\psi_d = \psi$, el estimador J-S de θ_d está dado por

$$\widehat{\theta}_{d,JS} = \theta_d^0 + \left[1 - \frac{(D-2)\psi}{S}\right](\widehat{\theta}_d - \theta_d^0), \quad D \geq 3, \tag{2.66}$$

asumiendo θ^0 como fijo, donde $S = \sum_d (\widehat{\theta}_d - \theta_d^0)^2$. Si θ_d^0 es el predictor de mínimos cuadrados entonces reemplazamos $D-2$ por $D-p-2$ en (2.66), donde p es el número de parámetros estimados en la ecuación de regresión. Nótese que (2.66) puede expresarse también como un estimador combinado con peso $\widehat{\phi}_{JS} = 1 - [(D-2)\psi]/S$ ligado a $\widehat{\theta}_d$ y $1 - \widehat{\phi}_{JS}$ al θ_d^0 supuesto a priori. El estimador J-S es llamado también un estimador de reducción porque reduce el estimador directo $\widehat{\theta}_d$ hacia el θ_d^0 supuesto inicialmente.

Varianzas desiguales ψ_d

Ahora consideremos al caso de varianzas muestrales conocidas pero desiguales ψ_d. Una manera sencilla de generalizar el método de James-Stein consiste en definir $\widehat{\delta}_d = \widehat{\theta}_d/\sqrt{\psi_d}$ tal que $\widehat{\delta}_d \overset{\text{ind}}{\sim} N(\delta_d, 1)$ y $\delta_d^0 = \theta_d^0/\sqrt{\psi_d}$ es el valor inicial supuesto de δ_d. Podemos aplicar (2.66) a los datos transformados y entonces retransformar hacia atrás a las coordenadas originales. Esto conduce a

$$\widehat{\theta}_{d,JS} = \theta_d^0 + \left(1 - \frac{D-2}{\tilde{S}}\right)(\widehat{\theta}_d - \theta_d^0), \quad D \geq 3, \tag{2.67}$$

donde $\tilde{S} = \sum_d (\widehat{\theta}_d - \theta_d^0)/\psi_d$. El estimador (2.67) domina al estimador directo $\widehat{\theta}_d$ en términos del ECM ponderado con pesos $1/\psi_d$, pero no en términos del ECM total:

$$\sum_d \frac{1}{\psi_d} \text{ECM}_p(\widehat{\theta}_{d,JS}) < \sum_d \frac{1}{\psi_d} \text{ECM}_p(\widehat{\theta}_d). \tag{2.68}$$

Es más, da el peso común $\tilde{\phi}_{JS} = 1 - (D-2)/\tilde{S}$ a $\widehat{\theta}_d$ y $1 - \tilde{\phi}_{JS}$ al valor inicial supuesto θ_d^0, es decir, cada $\widehat{\theta}_d$ se aproxima hacia el valor inicial supuesto θ_d^0 por el mismo factor $\tilde{\phi}_{JS}$. Esto no resulta atrayente para el usuario como en el caso del estimador combinado con pesos $\widehat{\phi}^*$ dado por (2.65). Nos gustaría tener una mayor reducción del ψ_d que es más grande.

Extensiones de los estimadores J-S

Varias extensiones de los métodos J-S se han estudiado en la literatura. En particular, la dominación de los resultados influyentes bajo simetría esférica o más generalmente para distribuciones elípticas de $\widehat{\theta}$ las cuales incluyen las distribuciones normal, t-student y doble exponencial (ver Bradwein y Strawderman, 1990; Srivastava y Bilodeau, 1989), y para las distribuciones de la familia exponencial (Ghosh y Auer, 1983).

2.2. Estimación basada en modelos

2.2.1. Introducción

Las estimaciones basadas en muestreo en poblaciones finitas clásico, son típicamente inestables debido a los tamaños de muestra pequeño que involucra. Esta deficiencia ha conducido al desarrollo de estimaciones basadas en modelos, los cuales usan información de otras áreas locales relacionadas, para obtener estimaciones que son más exactas. Uno de los primeros de estos procedimientos fue una aproximación sintética basada en el modelo, propuesta por González (1973), el cual fue subsecuentemente usado por González y Hoza (1978). Sin embargo, se reconoce ahora que tal metodología produce estimaciones con una tendencia a ser modelo-sesgadas y las medidas de incertidumbre asociada son engañosas. En las últimas décadas se han desarrollado otros procedimientos basados en modelos para estimar parámetros de áreas pequeñas que usan información de otras áreas relacionadas, entre estas técnicas se encuentran las aproximaciones basadas en bayes empírico y jerárquico. Ghosh y Rao (1994) revisaron la técnicas disponibles para la estimación de áreas pequeñas y mostraron que las estimaciones de áreas pequeñas basadas en bayes empírico y jerárquico, no tienen ninguno de los problemas indeseables asociados con las estimaciones obtenidas usando aproximaciones clásicas insesgadas o sintéticas.

Uno de los primeros usos de métodos de bayes empírico basados en modelos lineales para estimación de áreas pequeñas fue el de Fay y Herriot (1979). También se han propuesto estimadores de bayes jerárquico basados en modelos lineales. Datta y Ghosh (1991) presentaron una teoría bayesiana unificada para los modelos lineales mixtos con énfasis particular en la estimación de áreas pequeñas.

Los métodos tradicionales de estimación indirecta, están basados en modelos implícitos que proporcionan un enlace para relacionar áreas pequeñas mediante datos suplementarios. Consideremos ahora los modelos explícitos de áreas pequeñas que tienen en cuenta la variación específica entre áreas. En particular, introduciremos modelos mixtos que incluyen efectos aleatorios de área, que consideran, además, la variación entre áreas explicada por las variables auxiliares incluidas en el modelo.

Algunas ventajas del uso de modelos en la fase de estimación son las siguientes: pueden usarse modelos de diagnóstico para hallar modelos apropiados que se ajusten bien a los datos; pueden asociarse medidas de presición de área específica con cada estimación de área pequeña, a diferencia de las medidas globales (totales o promedios sobre áreas pequeñas) con frecuencia usadas con estimaciones sintéticas; pueden usarse modelos lineales mixtos así como mo-

delos no lineales , tales como modelos de regresión logística y modelos lineales generalizados con efectos aleatorios de área; pueden manejarse también estructuras de datos complejos, tales como dependencia espacial y estructuras en series de tiempo y pueden utilizarse recientes desarrollos metodológicos para modelos de efectos aleatorios para lograr inferencias exactas de áreas pequeñas.

Aunque presentamos una variedad de modelos para estimación en áreas pequeñas en este capítulo, es importante notar que el especialista de la materia o los usuarios finales deben tener influencia en la elección de modelos, particularmente sobre la introducción de variables auxiliares. También, el éxito de cualquier método basado en el modelo depende de la disponibilidad de buenos datos auxiliares. Debe prestarse más atención, por consiguiente, a la compilación de variables auxiliares que sean buenas predictoras de las variables estudiadas.

Los modelos estudiados pueden clasificarse en dos grandes grupos:

(i) Modelos de nivel de área, que relacionan las medias de área pequeña con las variables auxiliares de área específica. Estos modelos son esenciales si no disponemos de datos a nivel de unidad.

(ii) Modelos de nivel unidad, que relacionan los valores de la variable de estudio a nivel de unidad con las variables auxiliares de unidad específica.

Destacaremos algunos modelos que permiten obtener mejores predictores empíricos lineales insesgados, conocidos como EBLUP y estimadores de bayes empíricos (BE) y de bayes jerárquicos (BJ). Para implementar BJ, se necesitan supuestos adicionales sobre los parámetros del modelo, en la forma de distribuciones a priori. Se incluyen en el presente trabajo los modelos mixtos, es decir, aquellos que incluyen ambas variables auxiliares de nivel unidad y nivel de área.

2.2.2. Modelo básico de nivel de área (Tipo A)

Asumamos que $\theta_d = g(\overline{Y}_d)$, para alguna $g(\cdot)$ especificada, está relacionada con los datos auxiliares de área específica $\mathbf{z}_d = (z_{1d}, \cdots, z_{pd})^T$ a través de un modelo lineal

$$\theta_d = \mathbf{z}_d^T \boldsymbol{\beta} + b_d v_d, \quad d = 1, \cdots, D, \tag{2.69}$$

donde los b_d son constantes positivas conocidas y $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ es el vector de $p \times 1$ coeficientes de regresión. Además, los v_d son efectos aleatorios de área asumidos como independientes e idénticamente distribuídos (iid) con

$$E_m(v_d) = 0, \quad V_m(v_d) = \sigma_v^2 \, (\geq 0), \tag{2.70}$$

donde E_m denota la esperanza bajo el modelo y V_m la varianza bajo el modelo. Denotamos este supuesto como $v_d \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$. Con frecuencia se usa también el supuesto de normalidad de los efectos aleatorios v_d, pero es posible hacer inferencias "robustas" relajando este supuesto. El parámetro σ_v^2 es una medida de homogeneidad de las áreas después de considerar las covariables $\mathbf{z}_d$. En algunas aplicaciones, no todas las áreas pequeñas son seleccionadas en la muestra. Supongamos que tenemos D áreas en la población y solamente son seleccionadas d áreas en la muestra. Asumimos un modelo de la forma (2.69) para la población, es decir, $\theta_d = \mathbf{z}_d^T \boldsymbol{\beta} + b_d v_d, \quad d = 1, \cdots, D$. Además asumamos que la muestra de áreas obedece al modelo de la población, es decir, el sesgo está ausente en la selección de la muestra de áreas para que (2.69) se cumpla para las áreas muestreadas.

Para hacer inferencias acerca de las medias de áreas pequeñas $\overline{Y}_d$ bajo el modelo (2.69), asumamos que disponemos de estimadores directos $\widehat{\overline{Y}}_d$. Como en el método de James-Stein, asumimos que

$$\widehat{\theta}_d = g(\widehat{\overline{Y}}_d) = \theta_d + e_d, \quad d = 1, \cdots, D, \tag{2.71}$$

donde los errores de muestreo e_d son independientes con

$$E_p(e_d|\theta_d) = 0, \quad V_p(e_d|\theta_d) = \psi_d. \tag{2.72}$$

Es costumbre asumir también que las varianzas muestrales, ψ_d, son conocidas. Los anteriores supuestos pueden ser muy restrictivos en algunas aplicaciones. Por ejemplo, el estimador directo $\widehat{\theta}_d$ puede ser sesgado bajo el diseño para θ_d si $g(\cdot)$ es una función no lineal y el tamaño de la muestra de área n_d es pequeño. El supuesto de varianza conocida ψ_d puede relajarse estimando ψ_d de los datos de la muestra de nivel unidad y entonces suavizamos las varianzas estimadas $\widehat{\psi}_d$ para conseguir una estimación más estable de ψ_d. También se asume con frecuencia la normalidad del estimador $\widehat{\theta}_d$, pero no puede ser tan restrictiva como la normalidad de los efectos aleatorios debido al efecto del teorema central del límite en $\widehat{\theta}_d$.

Combinando (2.69) y (2.71) obtenemos el modelo

$$\widehat{\theta}_d = \mathbf{z}_d^T \boldsymbol{\beta} + b_d v_d + e_d, \quad d = 1, \cdots, D, \tag{2.73}$$

Nótese que (2.73) incluye errores inducidos por el diseño e_d así como errores del modelo v_d. Asumimos que e_d y v_d son independientes. El modelo (2.73) es un caso especial de un modelo lineal mixto.

El supuesto $E_p(e_d|\theta_d) = 0$ en el modelo de muestreo (2.71) puede no ser válido si el tamaño de muestra n_d en el área d es pequeña y θ_d es una función

no lineal del total Y_d, incluso si el estimador directo es insesgado bajo el diseño. Un modelo de muestreo más realista está dado por

$$\widehat{Y}_d = Y_d + e_d^*, \quad d = 1, \cdots, D, \tag{2.74}$$

con $E_p(e_d^*|Y_d) = 0$, es decir, $\widehat{Y}_d$ es insesgado bajo el diseño para el total Y_d. En este caso los modelos de muestreo y de enlace no coinciden. Como resultado, no podemos combinar (2.74) con el modelo de enlace (2.69) para producir un modelo lineal mixto de la forma (2.73). Por lo tanto, los resultados estándar en teoría de modelos lineales mixtos no se aplican.

Consideremos ahora varias extensiones del modelo básico de nivel de área (2.73)

Modelo Multivariante de Fay-Herriot

Supongamos que tenemos un vector de $r \times 1$ estimadores de una encuesta $\widehat{\boldsymbol{\theta}}_d = (\widehat{\theta}_{d1}, \cdots, \widehat{\theta}_{dr})^T$ y

$$\widehat{\boldsymbol{\theta}}_d = \boldsymbol{\theta}_d + \mathbf{e}_d, \quad d = 1, \cdots, D, \tag{2.75}$$

donde $\boldsymbol{\theta}_d = (\theta_{d1}, \cdots, \theta_{dr})^T$ con $\theta_{dj} = g_j(\overline{Y}_{dj})$, $j = 1, \cdots, r$ y los errores de muestreo $\mathbf{e}_d = (e_{d1}, \cdots, e_{dr})^T$ son independientes normales r-variados, $N_r(\mathbf{0}, \boldsymbol{\Psi}_d)$, con media $\mathbf{0}$ y matriz de covarianzas $\boldsymbol{\Psi}_d$ sobre $\boldsymbol{\theta}_d$). Aquí $\mathbf{0}$ es el vector nulo de $r \times 1$ y $\overline{Y}_{dj}$ es la d-ésima media de área pequeña para la j-ésima característica. Asumimos además que $\boldsymbol{\theta}_d$ está relacionado con el dato auxiliar de área $\{\mathbf{z}_{dj}\}$ a través del modelo lineal

$$\boldsymbol{\theta}_d = \mathbf{Z}_d\boldsymbol{\beta} + \mathbf{v}_d, \quad d = 1, \cdots, D, \tag{2.76}$$

donde los efectos aleatorios de área $\mathbf{v}_d$ son independientes $N_r(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{Z}_d$ es una matriz de $r \times rp$ con j-ésima fila dada por $(\mathbf{0}^T, \cdots, \mathbf{0}^T, \mathbf{z}_{dj}^T, \mathbf{0}^T, \cdots, \mathbf{0}^T)$ y $\boldsymbol{\beta}$ es el vector de rp coeficientes de regresión. Aquí $\mathbf{0}$ es el vector nulo de $p \times 1$ y $\mathbf{z}_{dj}^T$ ocurre en la j-ésima posición del vector fila (j-ésima fila).

Combinando (2.75) con (2.76), obtenemos un modelo lineal mixto multivariante

$$\widehat{\boldsymbol{\theta}}_d = \mathbf{Z}_d\boldsymbol{\beta} + \mathbf{v}_d + \mathbf{e}_d. \tag{2.77}$$

El modelo (2.77) es una extensión natural del modelo Fay-Herriot (2.73) con $b_d = 1$. Fay (1987) y Datta, Fay y Ghosh (1991), propusieron la extensión multivariante (2.77) y demostraron que puede llevar a estimadores más eficientes de las medias de áreas pequeñas $\overline{Y}_{dj}$ porque se aprovechan las correlaciones entre las componentes de $\widehat{\boldsymbol{\theta}}_d$ a diferencia del modelo univariante (2.73).

Modelo con errores de muestreo correlados

Una extensión natural del modelo Fay-Herriot (2.73) es el caso de errores de muestreo correlados e_d. Sea $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \cdots, \widehat{\theta}_D)^T$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D)^T$ y $\mathbf{e} = (e_1, \cdots, e_D)^T$, y asumimos que

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{e}, \tag{2.78}$$

con $\mathbf{e}|\boldsymbol{\theta} \sim N_D(\mathbf{0}, \boldsymbol{\Psi})$, donde la matriz de covarianzas de los errores de muestreo $\boldsymbol{\Psi} = (\psi_{dj})$ es conocida.

Combinando (2.78) con el modelo (2.69) para los θ_d obtenemos una generalización del modelo Fay-Herriot (2.73). Si $b_d = 1$ para todo d en (2.69) entonces el modelo combinado puede escribirse como

$$\widehat{\boldsymbol{\theta}} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{v} + \mathbf{e}, \tag{2.79}$$

donde $\mathbf{v} = (v_1, \cdots, v_D)^T$ y $\mathbf{Z}$ es una matriz de $D \times p$ con d-ésima fila igual a $\mathbf{z}_d^T$. En la práctica, $\boldsymbol{\Psi}$ se reemplaza por un estimador muestral $\widehat{\boldsymbol{\Psi}}$ o un estimador suavizado, pero la variabilidad asociada con el estimador con frecuencia se ignora.

Series de tiempo y modelos transversales

Muchas encuestas por muestreo se repiten en el tiempo con reemplazo parcial de los elementos de la muestra.

Rao y Yu (1992, 1994), propusieron una extensión del modelo básico de Fay-Herriot (2.73) para manejar series de tiempo y datos transversales. Su modelo consiste de un modelo de error de muestreo

$$\widehat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad t = 1, \cdots, T;\ d = 1, \cdots, D; \tag{2.80}$$

y un modelo de enlace

$$\theta_{dt} = \mathbf{z}_d^T\boldsymbol{\beta} + v_d + u_{dt}. \tag{2.81}$$

Aquí $\widehat{\theta}_{dt}$ es el estimador directo para el área pequeña d en el tiempo t, $\theta_{dt} = g(\overline{Y}_{dt})$ es una función de la media del área pequeña $\overline{Y}_{dt}$, los e_{dt} son los errores de muestreo normalmente distribuidos, dados los θ_{dt}, con medias cero y una matriz de covarianzas diagonal por bloques $\boldsymbol{\Psi}$ con bloques $\boldsymbol{\Psi}_d$, y $\mathbf{z}_{dt}$ es un vector de covariables de área, algunas de las cuales pueden cambiarse por t, por ejemplo, datos administrativos. Además, $v_d \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ y los u_{dt} se asume que siguen un proceso común autoregresivo de primer orden para cada d, esto es

$$u_{dt} = \rho u_{d,t-1} + \varepsilon_{dt}, \quad |\rho| < 1; \tag{2.82}$$

con $\varepsilon_{dt} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Se asume también que los errores $\{e_{dt}\}$,$\{v_d\}$ y $\{\varepsilon_{dt}\}$ son independientes. Los modelos de la forma (2.81) y (2.82) se han usado profusamente en la literatura econométrica (Anderson y Hsiao, 1981), ignorando los errores de muestreo e_{dt}.

El modelo (2.81) sobre los θ_{dt} depende de ambos efectos de área específica v_d y los efectos específicos de área por tiempo u_{dt} los cuales son correlados por tiempo para cada d. Podemos expresar también (2.81) como un modelo distribuído-retardado

$$\theta_{dt} = \rho\theta_{d,t-1} + (\mathbf{z}_{dt} - \rho\mathbf{z}_{d,t-1})^T\boldsymbol{\beta} + (1-\rho)v_d + \varepsilon_{dt}. \tag{2.83}$$

La forma alternativa (2.83) relaciona θ_{dt} con la media del periodo previo $\theta_{d,t-1}$ los valores de las variables auxiliares para los tiempos puntuales t y $t-1$, los efectos aleatorios de área pequeña v_d y los efectos de área por tiempo ε_{dt}. Pueden formularse más modelos complejos sobre los u_{dt} que (2.82) asumiendo un proceso autoregresivo (ARMA), pero la ganancia resultante en eficiencia con respecto a (2.82) es improbable que sea significativa.

Ghosh y Nangia (1993) y Ghosh, Nangia y Kim (1996), también propusieron un modelo de serie de tiempo transversal para estimación de áreas pequeñas. Su modelo es de la forma

$$\widehat{\theta}_{dt}|\theta_{dt} \overset{\text{ind}}{\sim} N(\theta_{dt}, \psi_{dt}), \tag{2.84}$$

$$\theta_{dt}|\boldsymbol{\alpha}_t \overset{\text{ind}}{\sim} N(\mathbf{z}_{dt}^T\boldsymbol{\beta} + \mathbf{w}_{dt}^T\boldsymbol{\alpha}_t, \sigma_t^2), \tag{2.85}$$

y

$$\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1} \overset{\text{ind}}{\sim} N(\mathbf{H}_t\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Delta}). \tag{2.86}$$

Aquí $\mathbf{z}_{dt}$ y $\mathbf{w}_{dt}$ son covariables de área específica, las ψ_{dt} son varianzas muestrales que se asumen conocidas, $\boldsymbol{\alpha}_t$ es un vector de $r \times 1$ de efectos aleatorios de tiempo específico, y $\mathbf{H}_t$ es una matriz conocida de $r \times r$. El modelo dinámico (o espacio de estados) (2.85) en el caso univariante ($r = 1$) con $H_t = 1$ se reduce al modelo de salto aleatorio. El modelo anterior sufre de dos graves limitaciones, a saber:

(i) Los estimadores directos $\widehat{\theta}_{dt}$ se asumen independientes sobre el tiempo para cada d. Este supuesto no es realista en el contexto de las encuestas repetidas con muestras solapadas, tales como la CPS y la LFS.

(ii) No se incluyen efectos aleatorios de área en el modelo, lo cual conduce a un excesiva reducción de estimadores de área pequeña similar a los estimadores sintéticos.

Datta, Lahiri y Maiti (2002) y You (1999), emplearon el muestreo de Rao-Yu y los modelos de enlace (2.80) y (2.81) pero reemplazando el modelo AR(1) (2.80) en los u_{dj} por un modelo de salto aleatorio dado por (2.82) con $\rho = 1$: $u_{dt} = u_{dt-1} + \varepsilon_{dt}$. Datta, Lahiri, Maiti y Lu (1999), consideraron un modelo similar pero agregaron términos extras al modelo de enlace para reflejar la variación temporal en su aplicación.

Pfefferman y Burck (1990), propusieron un modelo general involucrando efectos aleatorios específicos de área por tiempo. Su modelo es de la forma

$$\widehat{\theta}_{dt} = \theta_{dt} + e_{dt}, \tag{2.87}$$

$$\theta_{dt} = \mathbf{z}_{dt}^T \boldsymbol{\beta}_{dt}, \tag{2.88}$$

donde los coeficientes $\boldsymbol{\beta}_{dt} = (\beta_{dt0}, \cdots, \beta_{dtp})^T$ permiten variar particularmente y sobre el tiempo, y los errores de muestreo e_{dt} para cada área d se asume que son sucesivamente incorrelados con media 0 y varianza ψ_{dt}. La variación de $\boldsymbol{\beta}_{dt}$ sobre el tiempo está especificada por el siguiente modelo de espacio de estado:

$$\begin{bmatrix} \beta_{dtj} \\ \beta_{dj} \end{bmatrix} = \mathbf{T}_j \begin{bmatrix} \beta_{d,t-1,j} \\ \beta_{dj} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} v_{dtj}, \quad j = 0, 1, \cdots, p. \tag{2.89}$$

Aquí los β_{dj} son coeficientes fijos, $\mathbf{T}_j$ es una matriz conocida de 2×2 con (0,1) como la segunda fila, y el modelo de errores $\{v_{dtj}\}$ para cada d son incorrelados sobre el tiempo con media 0 y covarianzas $E_m\{v_{dtj}, v_{dtl}\} = \sigma_{vjl}$; $j, l = 0, 1, \cdots, p$.

La fórmula (2.89) cubre varios modelos útiles. Primero, la opción $\mathbf{T}_j = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ da el modelo de regresión de coeficientes aleatorios $\beta_{dtj} = \beta_{dj} + v_{dtj}$ (Swamy, 1971). El familiar modelo de salto aleatorio $\beta_{dtj} = \beta_{d,t-1,j} + v_{dtj}$ se obtiene escogiendo $\mathbf{T}_j = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. En este caso los coeficientes β_{dj} son redundantes y debe omitirse para $\mathbf{T}_j = 1$. La opción $\mathbf{T}_j = \begin{bmatrix} \rho & 1-\rho \\ 0 & 1 \end{bmatrix}$ da un modelo AR(1): $\beta_{dtj} - \beta_{dj} = \rho(\beta_{d,t-1,j} - \beta_{dj}) + v_{dtj}$. El modelo de espacio de estados (2.89) es muy general, pero el supuesto de errores de muestreo serialmente incorrelados e_{dt} en (2.87) es restrictiva en el contexto de encuestas repetidas con muestras solapadas.

Modelos espaciales

El modelo básico de Fay-Herriot (2.73) asume efectos aleatorios de área pequeña v_d iid, pero en algunas aplicaciones puede ser más realista y tomar en consideración modelos que permiten correlaciones entre los valores de v_d.

Los modelos espaciales sobre los v_d son usados cuando pueden definirse áreas "vecinas" de cada área. Tales modelos inducen correlaciones entre los v_d. Por ejemplo, correlaciones que dependen de la proximidad geográfica en el contexto de estimación de enfermedades locales y tasas de mortalidad. Cressie (1991), usó un modelo espacial para estimación de áreas pequeñas en el marco del subconteo en el censo de U.S.

Si A_d denota el conjunto de áreas "vecinas" de el área d, entonces un modelo espacial autorregresivo condicional (CAR) asume que la distribución condicional de $b_d v_d$ dado $\{v_l : l \neq d\}$ está dada por

$$b_d v_d | \{v_l : l \neq d\} \sim N\left(\rho \sum_{l \in A_d} q_{dl} b_l v_l, b_d^2 \sigma_v^2\right). \tag{2.90}$$

Aquí $\{q_{dl}\}$ son constantes conocidas que satisfacen $q_{dl} b_l^2 = q_{ld} b_d^2 \, (d < l)$, y $\delta = (\rho, \sigma_v^2)^T$ es un vector de parámetros desconocidos. El modelo (2.90) implica que

$$\mathbf{B}^{1/2}\mathbf{v} \sim N_m(\mathbf{0}, \mathbf{\Gamma}(\boldsymbol{\delta}) = \sigma_v^2(\mathbf{I} - \rho\mathbf{Q})^{-1}\mathbf{B}), \tag{2.91}$$

donde $\mathbf{B} = \text{diag}(b_1^2, \cdots, b_D^2)$ y $\mathbf{Q} = (q_{dl})$ es una matriz de $D \times D$ con $q_{dl} = 0$ cuando $l \neq A_d$ (incluyendo $q_{dd} = 0$) y $\mathbf{v} = (v_1, \cdots, v_D)^T$ (ver Besag, 1974). Usando (2.91) y (2.73) obtenemos un modelo espacial de área pequeña. Nótese que $\boldsymbol{\delta}$ aparece como no lineal en $\mathbf{\Gamma}(\boldsymbol{\delta})$.

En la literatura geoestadística se han usado estructuras de covarianzas de la forma (i) $\mathbf{\Gamma}(\boldsymbol{\delta}) = \sigma_v^2(\delta_1\mathbf{I} + \delta_2\mathbf{D})$ y (ii) $\mathbf{\Gamma}\boldsymbol{\delta}) = \sigma_v^2[\delta_1\mathbf{I} + \delta_2\mathbf{D}(\delta_3)]$, donde $\mathbf{D} = (e^{-\mathbf{d}_{dl}})$ y $\mathbf{D}(\delta_3) = (\delta_3^{\mathbf{d}_{dl}})$ son matrices de $D \times D$ con $\mathbf{d}_{dl}$ denotando una "distancia" (no necesariamente euclídea) entre las áreas pequeñas d y l. Nótese que en el caso (i) los parámetros δ_1 y δ_2 aparecen como lineales en $\mathbf{\Gamma}(\boldsymbol{\delta})$, mientras que en el caso (ii) δ_2 y δ_3 aparecen como no lineales en $\mathbf{\Gamma}(\boldsymbol{\delta})$. Una desventaja del modelo espacial (2.91) es que dependen de como se defina el vecindario A_d, lo que introduce algo de subjetividad al modelo.

2.2.3. Modelo básico de nivel unidad (Tipo B)

Asumamos disponible el dato auxiliar a nivel de unidad $\mathbf{x}_{dj} = (x_{dj1}, \cdots, x_{djp})^T$ para cada elemento j de la población en cada área pequeña d. Con frecuencia, es suficiente asumir que solamente son conocidas las medias poblacionales $\overline{X}$ o las medias de dominio $\overline{X}_d$. Además, la variable de interés y_{dj}, se asume relacionada con x_{dj} a través de un modelo de regresión lineal de error anidado univariante

$$y_{dj} = \mathbf{x}_{dj}^T\boldsymbol{\beta} + v_d + e_{dj}; \quad j = 1, \cdots, N_d, \quad d = 1, \cdots, D. \tag{2.92}$$

Aquí los efectos de área v_d se asumen como variables aleatorias iid que satisfacen (2.70), $e_{dj} = k_{dj}\tilde{e}_{dj}$ con constantes conocidas k_{dj} y los $\tilde{e}_{dj}$ son variables aleatorias independientes de los v_d y

$$E_m(\tilde{e}_{dj}) = 0, \quad V_m(\tilde{e}_{dj}) = \sigma_e^2. \tag{2.93}$$

Además, con frecuencia se supone normalidad de de los v_d. Los parámetros de interés son la media de área pequeña $\overline{Y}_d$ o el total Y_d. Los modelos de regresión estándar se obtienen haciendo $\sigma_v^2 = 0$ o equivalentemente $v_d = 0$ en (2.92). Tales modelos conducen a estimadores de tipo sintético.

Asumimos que una muestra, s_d de tamaño n_d se toma de las N_d unidades en el área d ($d = 1, \cdots, D$) y que los valores de la muestra también obedecen al modelo asumido (2.92). El último supuesto se satisface bajo muestreo aleatorio simple de cada área o más generalmente para diseños de muestreo que usen la información auxiliar $\mathbf{x}_{dj}$ en la selección de las muestras s_d. Para ver esto, escribimos (2.92) en forma matricial como

$$\mathbf{y}_d^P = \mathbf{X}_d^P\boldsymbol{\beta} + v_d\mathbf{1}_d^P + \mathbf{e}_d^P, \quad d = 1, \cdots, D, \tag{2.94}$$

donde $\mathbf{X}_d^P$ es una matriz de $N_d \times p$, $\mathbf{y}_d^P, \mathbf{1}_d^P$ y $\mathbf{e}_d^P$ son vectores de $N_d \times 1$ y $\mathbf{1}_d^P = (1, \cdots, 1)^T$. Particionemos (2.94) en unidades muestreadas y no muestreadas:

$$\mathbf{y}_d^P = \begin{bmatrix} \mathbf{y}_d \\ \mathbf{y}_d^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_d \\ \mathbf{X}_d^* \end{bmatrix}\boldsymbol{\beta} + v_d \begin{bmatrix} \mathbf{1}_d \\ \mathbf{1}_d^* \end{bmatrix} + \begin{bmatrix} \mathbf{e}_d \\ \mathbf{e}_d^* \end{bmatrix}, \tag{2.95}$$

Donde el superíndice $*$ denota las unidades no muestreadas. Si el modelo se conserva para la muestra, es decir, si la selección se hace sin sesgo, entonces las inferencias sobre $\psi = (\boldsymbol{\beta}^T, \sigma_v^2, \sigma_e^2)^T$ están basadas en

$$f(\mathbf{y}_d|\mathbf{X}_d^P, \psi) = \int f(\mathbf{y}_d, \mathbf{y}_d^*|\mathbf{X}_d^P, \psi)d\mathbf{y}_d^*, \quad d = 1, \cdots, D, \tag{2.96}$$

donde $f(\mathbf{y}_d, \mathbf{y}_d^*|\mathbf{X}_d^P, \psi)$ es la distribución conjunta de y_d y y_d^*. Por otro lado, fijando $\mathbf{a}_d = (a_{d1}, \cdots, a_{dN_d})^T$ con $a_{dj} = 1$ si $j \in s_d$ y $a_{dj} = 0$ en otro caso, la distribución de los datos de la muestra $(\mathbf{y}_d, \mathbf{a}_d)$ está dada por

$$f(\mathbf{y}_d, \mathbf{a}_d|\mathbf{X}_d^P, \psi) = \int f(\mathbf{y}_d, \mathbf{y}_d^*|\mathbf{X}_d^P, \psi)f(\mathbf{a}_d|\mathbf{y}_d, \mathbf{y}_d^*, \mathbf{X}_d^P)d\mathbf{y}_d^*$$

$$= \left[\int f(\mathbf{y}_d, \mathbf{y}_d^*|\mathbf{X}_d^P, \psi)d\mathbf{y}_d^*\right] f(\mathbf{a}_d|\mathbf{X}_d^P),$$

siempre que

$$f(\mathbf{a}_d|\mathbf{y}_d, \mathbf{y}_d^*, \mathbf{X}_d^P) = f(\mathbf{a}_d|\mathbf{X}_d^P),$$

es decir, las probabilidades de selección de la muestra no dependen de $\mathbf{y}_d^P$ pero podrían depender de $\mathbf{X}_d^P$. En este caso, no existe sesgo de selección y podemos asumir que los valores muestrales también obedecen al modelo asumido, es decir, usamos $f(\mathbf{y}_d|\mathbf{X}_d^P, \psi)$ para inferencias sobre ψ (Smith, 1983).

Si las probabilidades de selección de la muestra dependen de una variable auxiliar, digamos $\mathbf{z}_d^P$, la cual no está incluida en X_d^P, entonces la distribución de los datos de la muestra $(\mathbf{y}_d, \mathbf{a}_d)$ es

$$f(\mathbf{y}_d, \mathbf{a}_d|\mathbf{X}_d^P, \mathbf{z}_d^P, \psi) = \left[\int f(\mathbf{y}_d, \mathbf{y}_d^*|\mathbf{X}_d^P, \mathbf{z}_d^P, \psi) d\mathbf{y}_d^*\right] f(\mathbf{a}_d|\mathbf{z}_d^P, \mathbf{X}_d^P).$$

En este caso, las inferencias sobre ψ están basadas en $f(\mathbf{y}_d|\mathbf{X}_d^P, \mathbf{z}_d^P, \psi)$ la cual es diferente de (2.96) a menos que $\mathbf{z}_d^P$ no esté relacionada a $\mathbf{y}_d^P$ dado $\mathbf{X}_d^P$. En este caso tenemos una selección muestral sesgada y por consiguiente no podemos asumir que el modelo (2.94) se conserve para los valores de la muestra. El modelo (2.94) no es apropiado tampoco bajo el muestreo por conglomerados en dos etapas dentro de áreas pequeñas, porque los efectos aleatorios de conglomerados no se incorporan.

Escribimos la media del área pequeña $\overline{Y}_d$ como

$$\overline{Y}_d = f_d \overline{y}_d + (1 - f_d)\overline{Y}_d^*, \tag{2.97}$$

con $f_d = n_d/N_d$ y $\overline{y}_d$ y $\overline{Y}_d^*$ denotando las medias de los elementos muestreados y no muestreados, respectivamente. Se sigue de (2.97) que la estimación de la media del área pequeña $\overline{Y}_d$ es equivalente a estimar la realización de la variable aleatoria $\overline{Y}_d^*$ dado el dato de la muestra $\{\mathbf{y}_d\}$ y el dato auxiliar $\{\mathbf{X}_d^P\}$.

Si el tamaño de la población N_d es grande, entonces podemos tomar las medias de áreas pequeñas como

$$\overline{Y}_d = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + v_d, \tag{2.98}$$

notando que $\overline{Y}_d = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + v_d + \overline{E}_d$ y $\overline{E}_d \approx 0$, donde $\overline{E}_d$ es la media de los N_d errores e_{dj} y $\overline{\mathbf{X}}_d$ es la media conocida de $\mathbf{X}_d^P$. Se sigue de (2.98) que la estimación de $\overline{Y}_d$ es equivalente a la estimación de una combinación lineal de $\boldsymbol{\beta}$ y la realización de la variable aleatoria v_d.

Ahora consideremos algunas extensiones del modelo básico tipo B (2.92)

Modelo de regresión de error anidado multivariante

Asumimos que los datos auxiliares de unidad están disponibles para todos los elementos de la población j en cada área pequeña d. Además asumimos que un vector de $r \times 1$ variables de interés, y_{dj}, están relacionadas a un vector

x_{dj} a través de un modelo de regresión de error anidado multivariante (Fuller y Harter, 1987):

$$y_{dj} = \mathbf{B}x_{dj} + \mathbf{v}_d + \mathbf{e}_{dj}; \quad j = 1, \cdots, N_d; d = 1, \cdots, D. \tag{2.99}$$

Aquí $\mathbf{B}$ es una matriz de $r \times p$ coeficientes de regresión, $\mathbf{v}_d$ son efectos de área asumidos como vectores aleatorios iid con media $\mathbf{0}$ y matriz de covarianzas $\mathbf{\Sigma}_v$, y los vectores de errores aleatorios $\mathbf{e}_{dj}$ son iid con media $\mathbf{0}$ y matriz de covarianzas $\mathbf{\Sigma}_e$ e independientes de $\mathbf{v}_d$. Además, se asume normalidad de los $\mathbf{v}_d$ y $\mathbf{e}_d$. El modelo (2.99) es una extensión natural del modelo de regresión del error anidado univariante (2.92) con $k_{dj} = 1$.

Los parámetros de interés son el vector de medias de áreas pequeñas $\overline{\mathbf{Y}}$ que pueden aproximarse por $\boldsymbol{\mu}_d = \mathbf{B}\overline{\mathbf{X}}_d + \mathbf{v}_d$ si el tamaño de la población N_d es grande. En el último caso, se sigue que la estimación de $\boldsymbol{\mu}_d$ es equivalente a la estimación de una combinación lineal de $\mathbf{B}$ y la realización del vector aleatorio $\mathbf{v}_d$. Como en el modelo de Fay-Herriot multivariante, el modelo de nivel unidad (2.99) puede conducir a estimadores más eficientes aprovechando la correlación entre las componentes de y_{dj}, a diferencia del modelo univariante (2.92).

Modelo lineal con varianza de error aleatorio

Consideremos las componentes de error del modelo (2.92) donde no disponemos de los datos auxiliares $\{\mathbf{x}_{dj}\}$: $y_{dj} = \beta + v_d + e_{dj}$ con $v_d \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ y $e_{dj} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$. El supuesto de igualdad de varianzas del error puede relajarse si hacemos $e_{dj}|\sigma_{ed}^2 \overset{\text{ind}}{\sim} N(0, \sigma_{ed}^2)$ y asumimos la varianza del error σ_{ed}^2 como una variable aleatoria no negativa iid con media σ_e^2 y varianza δ_e (digamos) e independiente de v_d. Aragón (1984), usó tal modelo de varianza de error aleatorio con varianzas del error σ_{ed}^2 de una Gausiana inversa. Kleffe y Rao (1992), usaron el anterior modelo simple para la estimación de área pequeña, y Arora y Lahiri (1979), lo extendieron al caso de regresión (2.92) con $\mathbf{x}_{dj} = \mathbf{x}_d, k_{dj} = 1, e_{dj} = \tilde{e}_{dj}$ y $e_{dj}|\sigma_{ed}^2 \overset{\text{iid}}{\sim} N(0, \sigma_{ed}^2)$.

Modelo de regresión de error anidado doble

Supongamos que el área d pequeña contiene M_d unidades primarias (o conglomerados) y que la j-ésima unidad primaria (conglomerado) en el área d contiene N_{dj} subunidades (elementos). Sean $(y_{djl}, \mathbf{x}_{djl})$ los valores y y $\mathbf{x}$ del l-ésimo elemento en la j-ésima unidad primaria de el área d ($l = 1, \cdots, N_{dj}, j = 1, \cdots, M_d, d = 1, \cdots, D$). Bajo esta estructura de la población es común practicar el empleo de dos etapas de muestreo de conglomerado en cada área pequeña: seleccionamos una muestra s_d de m_d unidades primarias de el área d y si

el j-ésimo conglomerado es muestreado, entonces seleccionamos una submuestra, s_{dj}, de n_{dj} elementos del j-ésimo conglomerado y se observan los valores asociados de y y $\mathbf{x}$. La estructura de la población anterior se refleja mediante el modelo de regresión de error anidado doble (Stukel y Rao, 1999):

$$y_{djl} = \mathbf{x}_{djl}^T\boldsymbol{\beta} + v_d + u_{dj} + e_{djl}; \quad l = 1, \cdots, N_{dj}, j = 1, \cdots, M_d, d = 1, \cdots, D. \tag{2.100}$$

Aquí los efectos de área $\{v_d\}$, los efectos de conglomerado $\{u_{dj}\}$ y los errores residuales $\{e_{djl}\}$ con $e_{djl} = k_{djl}\tilde{e}_{djl}$ y las constantes conocidas k_{djl} se asumen como mutuamente independientes. Además, $v_d \overset{\text{iid}}{\sim}(0, \sigma_v^2)$, $u_{dj} \overset{\text{iid}}{\sim}(0, \sigma_u^2)$ y $\tilde{e}_{djl} \overset{\text{iid}}{\sim}(0, \sigma_e^2)$; se asume con frecuencia normalidad de las componentes aleatorias v_d, u_{dj} y $\tilde{e}_{djl}$. Asumimos que los valores muestrales también obedecen al modelo asumido (2.100) el cual se satisface bajo muestreo aleatorio simple de conglomerados y subunidades muestreadas dentro de conglomerado, o más generalmente, para diseños de muestreo que usan la información auxiliar $\mathbf{x}_{djl}$ en la selección de la muestra. Datta y Ghosh (1991), usaron el modelo (2.100) para el caso especial de covariables espec 'ificas de conglomerado, es decir, $\mathbf{x}_{djl} = \mathbf{x}_{dj}$. Ghosh y Lahiri (1988), estudiaron el caso de no información auxiliar, es decir, $\mathbf{x}_{djl}^T\boldsymbol{\beta} = \beta$.

Los parámetros de interés son las medias de áreas pequeñas

$$\overline{Y}_d = \frac{1}{N_d}\left[\sum_{j\in s_d}\sum_{l\in s_{dj}} y_{djl} + \sum_{j\in s_d}\sum_{l\in s_{dj}^c} y_{djl}^* + \sum_{j\in s_d^c}\sum_{l=1}^{N_{dj}} y_{djl}^*\right] \tag{2.101}$$

donde y_{djl}^* son los valores de y no muestreados, s_d^c y s_{dj}^c denotan los conglomerados y subunidades no muestreadas. Si el número de unidades primarias, N_d es grande, entonces $\overline{Y}_d$ puede ser aproximada como

$$\overline{Y}_d \approx \overline{\mathbf{X}}_d^T\boldsymbol{\beta} + v_d, \tag{2.102}$$

notando que $\overline{Y}_d = \overline{\mathbf{X}}_d^T\boldsymbol{\beta} + v_d + \overline{U}_d + \overline{E}_d$ y $\overline{U}_d \approx 0, \overline{E}_d \approx 0$, donde $\overline{U}_d$ y $\overline{E}_d$ son las medias de área de u_{dj} y e_{djl} y $\overline{\mathbf{X}}_d$ es la media conocida de los $\mathbf{x}_{djl}$. Se sigue de (2.102) que la estimación de $\overline{Y}_d$ es equivalente a la estimación de una combinación lineal de $\boldsymbol{\beta}$ y la realización de la variable aleatoria v_d.

Modelo de dos niveles

El modelo básico de nivel de unidad (2.92) con ordenada en el origen β_1 puede expresarse como un modelo con término ordenada aleatorio $\beta_{1d} = \beta_1 + v_d$ y pendientes comunes $\beta_2, \cdots, \beta_p$: $y_{dj} = \beta_{1d} + \beta_2 x_{dj2} + \cdots + \beta_p x_{djp} + e_{dj}$. Esto sugiere un modelo más general que permita diferencias entre pendientes

así como las ordenadas a través del área pequeña. Introducimos coeficientes aleatorios $\boldsymbol{\beta} = (\beta_{d1}, \cdots, \beta_{dp})^T$ y modelamos entonces $\boldsymbol{\beta}_d$ en términos de las covariables de nivel de área $\tilde{\mathbf{Z}}_d$ para llegar al modelo de área pequeña de dos niveles (Moura y Holt, 1999):

$$y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + e_{dj}, \quad j = 1, \cdots, N_d; d = 1, \cdots, D, \tag{2.103}$$

y

$$\boldsymbol{\beta}_d = \tilde{\mathbf{Z}}_d \boldsymbol{\alpha} + \mathbf{v}_d, \tag{2.104}$$

donde $\tilde{\mathbf{Z}}_d$ es una matriz de $p \times p$, $\boldsymbol{\alpha}$ es un vector de $p \times 1$ parámetros de regresión, $\mathbf{v}_d \overset{\text{iid}}{\sim}(0, \Sigma_v)$ y $e_{dj} = k_{dj}\tilde{e}_{dj}$ con $\tilde{e}_{dj} \overset{\text{iid}}{\sim}(0, \sigma_e^2)$. Podemos escribir (2.103) en forma matricial

$$\mathbf{y}_d^P = \mathbf{X}_d^P \boldsymbol{\beta}_d + \mathbf{e}_d^P. \tag{2.105}$$

El modelo de dos niveles (2.104)-(2.105) integra eficazmente el uso de covariables de nivel unidad y nivel de área dentro de un solo modelo:

$$\mathbf{y}_d^P = \mathbf{X}_d^P \tilde{\mathbf{Z}}_d \boldsymbol{\alpha} + \mathbf{X}_d^P \mathbf{v}_d + \mathbf{e}_d^P. \tag{2.106}$$

Además, el uso de pendientes aleatorias, $\boldsymbol{\beta}_d$, permite mayor flexibilidad en la modelación. Se asume que los valores muestrales $\{(y_{dj}, \mathrm{x}_{dj}); j = 1, \cdots, n_d; d = 1, \cdots, D\}$ obedecen al modelo (2.106), es decir, es una selección muestral sin sesgo. Si N_d es grande, podemos expresar la media $\overline{Y}_d$ bajo (2.106) como

$$\overline{Y}_d \approx \overline{\mathbf{X}}_d^T \tilde{\mathbf{Z}}_d \boldsymbol{\alpha} + \overline{\mathbf{X}}_d^T \mathbf{v}_d. \tag{2.107}$$

Se sigue de (2.107) que la estimación de $\overline{Y}_d$ es equivalente a la estimación de la combinación lineal de $\boldsymbol{\beta}$ y la realización del vector aleatorio $\mathbf{v}_d$ con matriz de covarianzas desconocida $\boldsymbol{\Sigma}_v$.

El modelo (2.106) es un caso especial de un modelo lineal general mixto usado profusamente para datos longitudinales (Laird y Ware, 1982). Estos modelos admiten matrices arbitrarias $\mathbf{X}_{1d}^P$ y $\mathbf{X}_{2d}^P$ para ser asociadas con $\boldsymbol{\alpha}$ y $\mathbf{v}_d$:

$$\mathbf{y}_d^P = \mathbf{X}_{1d}^P \boldsymbol{\alpha} + \mathbf{X}_{2d}^P \mathbf{v}_d + \mathbf{e}_d^P. \tag{2.108}$$

La opción $\mathbf{X}_{1d}^P = \mathbf{X}_d^P \mathbf{Z}_d$ y $\mathbf{X}_{2d}^P = \mathbf{X}_d^P$ da el modelo de dos niveles (2.106). Este modelo cubre muchos de los modelos de áreas pequeñas considerados en la literatura.

Modelo mixto lineal general

Datta y Ghosh (1991), consideraron un modelo mixto lineal general el cual cubre a los modelos de nivel unidad univariantes como casos especiales:

$$\mathbf{y}^P = \mathbf{X}^P \boldsymbol{\beta} + \mathbf{Z}^P \mathbf{v} + \mathbf{e}^P. \tag{2.109}$$

Aquí $\mathbf{e}^P$ y $\mathbf{v}$ son independientes con $\mathbf{e}^P \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\boldsymbol{\Psi}^P)$ y $\mathbf{v} \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{D}(\boldsymbol{\lambda}))$, donde $\boldsymbol{\Psi}^P$ es una matriz definida positiva conocida y $\mathbf{D}(\boldsymbol{\lambda})$ es una matriz definida positiva la cual es estructuralmente conocida excepto para algunos parámetros $\boldsymbol{\lambda}$ típicamente involucrando razones de componentes de varianza de la forma σ_d^2/σ^2. Además, $\mathbf{X}^P$ y $\mathbf{Z}^P$ son matrices de diseño conocidas y $\mathbf{y}^P$ es el vector de $N \times 1$ de valores y de la población.

Podemos particionar (2.109), de forma similar a (2.95), como

$$\mathbf{y}^P = \begin{bmatrix} \mathrm{y} \\ \mathrm{y}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{X}^* \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}^* \end{bmatrix} \mathbf{v} + \begin{bmatrix} \mathbf{e} \\ \mathbf{e}^* \end{bmatrix}, \tag{2.110}$$

donde el asterisco denota unidades no muestreadas. El vector de totales de áreas pequeñas Y_d es de la forma $\mathbf{A}\mathrm{y}+\mathbf{C}\mathrm{y}^*$ con $\mathbf{A} = \oplus_{d=1}^{D} \mathbf{1}_{n_d}^T$ y $\mathbf{C} = \oplus_{d=1}^{D} \mathbf{1}_{N_d - n_d}^T$, donde $\oplus$ denota la suma directa, es decir, $\oplus_{d=1}^{D} \mathbf{A}_u = \mathrm{diag}(\mathbf{A}_1, \cdots, \mathbf{A}_D)$ por bloques.

Datta y Ghosh (1991), mencionaron un modelo de clasificación cruzada el cual es cubierto por el modelo general (2.109) pero no por el modelo "longitudinal" (2.108). Supongamos que las unidades en un área pequeña son clasificadas dentro de C grupos (por ejemplo, edad, clase socioeconómica) etiquetados $j = 1, \cdots, C$ y los tamaños de celda de subgrupo por área N_{dj} son conocidos. Un modelo de clasificación cruzada está dado entonces por

$$y_{djk} = \mathbf{x}_{djk}^T \boldsymbol{\beta} + v_d + a_j + u_{dj} + e_{djk}, \tag{2.111}$$

$$k = 1, \cdots, N_{dj}; j = 1, \cdots, C; d = 1, \cdots, D$$

donde $\{v_d\}, \{a_j\}$ y $\{u_{dj}\}$ son mutuamente independientes con $e_{djk} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, $v_d \overset{\text{iid}}{\sim} N(0, \lambda_1\sigma^2)$, $a_j \overset{\text{iid}}{\sim} N(0, \lambda_2\sigma^2)$ y $u_{dj} \overset{\text{iid}}{\sim} N(0, \lambda_3\sigma^2)$. Lui y Cumberland (1989), consideraron un modelo de la forma (2.111) con $\lambda_1 = \lambda_3 = 0$, es decir, v_d y u_{dj} son degeneradas en cero.

2.3. Modelos de efectos mixtos y estimadores EBLUP

En la sección anterior hemos presentado varios modelos de áreas pequeñas que pueden considerarse como casos especiales de un modelo mixto lineal general de efectos fijos y aleatorios. Además, las medias o totales de áreas pequeñas, pueden expresarse como combinaciones lineales de efectos aleatorios y fijos. Ahora consideraremos el conocido como "mejor predictor lineal insesgado" (BLUP) del estimador de tales parámetros, el cual puede obtenerse del enfoque frecuentista clásico, recurriendo a resultados generales sobre estimación BLUP. Los estimadores BLUP minimizan el ECM entre la clase de estimadores insesgados lineales y no dependen de la normalidad de los efectos aleatorios; pero dependen de las varianzas (o covarianzas) de los efectos aleatorios, los cuales pueden ser estimados por el método de momentos. Alternativamente, pueden usarse los métodos de máxima verosimilitud (MV) o máxima verosimilitud restringida (MVR) para estimar la varianza y componentes de covarianzas, asumiendo normalidad. Usando estos componentes estimados en el estimador BLUP obtenemos un estimador en dos etapas el cual se conoce como el estimador empírico BLUP o EBLUP (Harville, 1991).

A continuación presentamos resultados generales sobre estimación EBLUP y el problema de estimación del ECM de estimadores EBLUP, tomando en cuenta la variabilidad en la varianza estimada y componentes de covarianza.

2.3.1. Modelo mixto lineal general

Supongamos que los datos de la muestra obedecen al modelo mixto lineal general

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}. \tag{2.112}$$

Aquí $\mathbf{y}$ es el vector de $n \times 1$ observaciones de la muestra, $\mathbf{X}$ y $\mathbf{Z}$ son matrices conocidas de $n \times p$ y $n \times h$ de rango completo, y $\mathbf{v}$ y $\mathbf{e}$ son independientemente distribuídas con media $\mathbf{0}$ y matrices de covarianzas $\mathbf{G}$ y $\mathbf{R}$ dependiendo de la varianza de algunos parámetros $\boldsymbol{\delta}(\delta_1, \cdots, \delta_q)^T$. Asumimos que $\boldsymbol{\delta}$ pertenece a un subconjunto especificado de un q-espacio euclídeo tal que $\text{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{V}(\boldsymbol{\delta}) = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$ es no singular para todos los $\boldsymbol{\delta}$ pertenecientes al subconjunto, donde $\text{Var}(\mathbf{y})$ denota la matriz de varianzas y covarianzas de $\mathbf{y}$.

Estamos interesados en estimar una combinación lineal, $\mu = \mathbf{1}^T\boldsymbol{\beta} + \mathbf{m}^T\mathbf{v}$ de los parámetros de regresión $\boldsymbol{\beta}$ y la realización de $\mathbf{v}$, para vectores especificados de constantes $\mathbf{1}$ y $\mathbf{m}$. Un estimador lineal de μ es de la forma $\widehat{\mu} = \mathbf{a}^T\mathbf{y} + b$ para

a y b conocidos. El estimador es insesgado bajo el modelo para μ si

$$E(\widehat{\mu}) = E(\mu), \tag{2.113}$$

donde E denota la esperanza con respecto al modelo (2.112). El ECM de $\widehat{\mu}$ está dado por

$$\text{ECM}(\widehat{\mu}) = E(\widehat{\mu} - \mu)^2, \tag{2.114}$$

el cual se reduce a la varianza del error $\widehat{\mu} - \mu$:

$$\text{ECM}(\widehat{\mu}) = \text{Var}(\widehat{\mu} - \mu)$$

si $\widehat{\mu}$ es insesgado para μ. Valliant, Dorfman y Royall (2000, p.27), denotan $E(\widehat{\mu} - \mu)^2$ como la varianza del error (o predicción de la varianza) bajo el modelo. Estamos interesados en hallar el estimador BLUP que minimiza el ECM en la clase de los estimadores lineales insesgados de $\widehat{\mu}$.

Estimador BLUP

Para $\boldsymbol{\delta}$ conocido, el estimador BLUP de μ está dado por

$$\tilde{\mu}^H = t(\boldsymbol{\delta}, \mathbf{y}) = \mathbf{1}^T\tilde{\boldsymbol{\beta}} + \mathbf{m}^T\tilde{\mathbf{v}} = \mathbf{1}^T\tilde{\boldsymbol{\beta}} + \mathbf{m}^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \tag{2.115}$$

donde

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{2.116}$$

es el mejor estimador lineal insesgado (BLUE) de $\boldsymbol{\beta}$,

$$\tilde{\mathbf{v}} = \tilde{\mathbf{v}}(\boldsymbol{\delta}) = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \tag{2.117}$$

El superíndice H en $\tilde{\mu}$ fue puesto por Henderson, quien propuso el estimador (2.115); (ver Henderson, 1950). Una demostración de que (2.115) es el estimador BLUP está dada en Henderson (1963).

ECM del BLUP

El estimador BLUP $\mathbf{t}(\boldsymbol{\delta}, \mathbf{y})$ puede expresarse como

$$\mathbf{t}(\boldsymbol{\delta}, \mathbf{y}) = \mathbf{t}^*(\boldsymbol{\delta}, \boldsymbol{\beta}, \mathbf{y}) + \mathbf{d}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

donde $\mathbf{t}^*(\boldsymbol{\delta}, \boldsymbol{\beta}, \mathbf{y})$ es el estimador BLUP cuando $\boldsymbol{\beta}$ es conocido:

$$\mathbf{t}^*(\boldsymbol{\delta}, \boldsymbol{\beta}, \mathbf{y}) = \mathbf{1}^T\boldsymbol{\beta} + \mathbf{b}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.118}$$

con

$$\mathbf{b}^T = \mathbf{m}^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1},$$

y

$$\mathbf{d}^T = \mathbf{1}^T - \mathbf{b}^T\mathbf{X}.$$

Se sigue ahora que $\mathbf{t}^*(\boldsymbol{\delta}, \boldsymbol{\beta}, \mathbf{y}) - \mu$ y $\mathbf{d}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ son no correladas, dado que

$$E[(\mathbf{b}^T(\mathbf{Z}\mathbf{v} + \mathbf{e}) - \mathbf{m}^T\mathbf{v})(\mathbf{v}^T\mathbf{Z}^T + \mathbf{e}^T)\mathbf{V}^{-1}] = 0.$$

Por lo tanto,

$$\text{ECM}[t(\boldsymbol{\delta}, \mathbf{y})] = \text{ECM}[t^*(\boldsymbol{\delta}, \boldsymbol{\beta}, \mathbf{y})] + \text{Var}[\mathbf{d}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}), \quad (2.119)$$

donde

$$g_1(\boldsymbol{\delta}) = \text{Var}[t^*(\boldsymbol{\delta}, \boldsymbol{\beta}, \mathbf{y}) - \mu] = \mathbf{m}^T(\mathbf{G} - \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{G})\mathbf{m} \quad (2.120)$$

y

$$g_2(\boldsymbol{\delta}) = \mathbf{d}^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d}. \quad (2.121)$$

El segundo término, $g_2(\boldsymbol{\delta})$, en (2.119) cuenta para la variabilidad en el estimador $\tilde{\boldsymbol{\beta}}$.

Estimador EBLUP

El estimador BLUP $t(\boldsymbol{\delta}, \mathbf{y})$ dado por (2.115) depende de la varianza de los parámetros $\boldsymbol{\delta}$ los cuales son desconocidos en aplicaciones prácticas. Reemplazamos $\boldsymbol{\delta}$ por un estimador $\widehat{\boldsymbol{\delta}} = \boldsymbol{\delta}(\mathbf{y})$, obtenemos un estimador en dos etapas $\widehat{\mu}^H = t(\widehat{\boldsymbol{\delta}}, \mathbf{y})$, el cual se llama estimador EBLUP. Por conveniencia, también escribimos $t(\widehat{\boldsymbol{\delta}}, \mathbf{y})$ y $t(\boldsymbol{\delta}, \mathbf{y})$ como $t(\widehat{\boldsymbol{\delta}})$ y $t(\boldsymbol{\delta})$.

El estimador en dos etapas $t(\widehat{\boldsymbol{\delta}})$ permanece insesgado para μ, es decir $E[t(\widehat{\boldsymbol{\delta}}) - \mu] = 0$, si se cumple que

- $E[t(\widehat{\boldsymbol{\delta}})]$ es finito;
- $\widehat{\boldsymbol{\delta}}$ es cualquier traslación invariante del estimador de $\boldsymbol{\delta}$, es decir, $\widehat{\boldsymbol{\delta}}(-\mathbf{y}) = \widehat{\boldsymbol{\delta}}(\mathbf{y})$ y $\widehat{\boldsymbol{\delta}}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \widehat{\boldsymbol{\delta}}(\mathbf{y})$ para todo $\mathbf{y}$ y $\mathbf{b}$;
- Las distribuciones de $\mathbf{v}$ y $\mathbf{e}$ son ambas simétricas en torno a $\mathbf{0}$ (no necesariamente normal).

Kackar y Harville (1981), demostraron que los procedimientos estándar para estimar $\boldsymbol{\delta}$, producen estimadores invariantes aún con traslación; en particular, MV, MVR y los métodos de constantes fijas (también llamado método de Henderson). Consultar Searle, Casella y McCulloch (1992) y Rao (1997), para mayores detalles del método para el análisis de modelos de varianza (ANOVA),

el cual es un caso especial del modelo mixto lineal general (2.112). El modelo ANOVA está dado por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{v}_1 + \cdots + \mathbf{Z}_r\mathbf{v}_r + \mathbf{e}, \tag{2.122}$$

donde $\mathbf{v}_1, \cdots, \mathbf{v}_r$ y $\mathbf{e}$ son independientemente distribuidas con media $\mathbf{0}$ y matrices de covarianzas $\sigma_1^2\mathbf{I}_{h_1}, \cdots, \sigma_r^2\mathbf{I}_{h_r}$ y $\sigma_e^2\mathbf{I}_n$. Los parámetros $\boldsymbol{\delta} = (\sigma_0^2, \cdots, \sigma_r^2)^T$ con $\sigma_i^2 \geq 0\,(i = 1, \cdots, r)$ y $\sigma_0^2 = \sigma_e^2 > 0$ son las componentes de varianza. Nótese que $\mathbf{G}$ es ahora diagonal por bloques, con bloques $\sigma_i^2\mathbf{I}_{h_i}$, $\mathbf{R} = \sigma_e^2\mathbf{I}_n$ y $V = \sigma_e^2\mathbf{I}_n + \sum \sigma_i^2\mathbf{Z}_i\mathbf{Z}_i^T$, la cual es un caso especial de la matriz de covarianzas con estructura lineal: $\mathbf{V} = \sum \boldsymbol{\delta}_i\mathbf{H}_i$ para matrices simétricas conocida $\mathbf{H}_i$.

Estimación del ECM del EBLUP

Para aplicaciones prácticas, necesitamos un estimador de ECM$[t(\widehat{\boldsymbol{\delta}})]$ como una medida de variabilidad asociada con $t(\widehat{\boldsymbol{\delta}})$. Una solución sencilla es aproximar ECM$[t(\widehat{\boldsymbol{\delta}})]$ por ECM$[t(\boldsymbol{\delta})]$ y entonces sustituimos $\widehat{\boldsymbol{\delta}}$ por $\boldsymbol{\delta}$. El estimador resultante del ECM está dado por

$$\text{ecm}_N[t(\widehat{\boldsymbol{\delta}})] = g_1(\widehat{\boldsymbol{\delta}}) + g_2(\widehat{\boldsymbol{\delta}}). \tag{2.123}$$

Otro estimador de ECM se obtiene sustituyendo $\widehat{\boldsymbol{\delta}}$ por $\boldsymbol{\delta}$ en la aproximación siguiente del ECM:

$$\text{ecm}_1[t(\widehat{\boldsymbol{\delta}})] = g_1(\widehat{\boldsymbol{\delta}}) + g_2(\widehat{\boldsymbol{\delta}}) + g_3(\widehat{\boldsymbol{\delta}}). \tag{2.124}$$

Tenemos que $Eg_2(\widehat{\boldsymbol{\delta}}) \approx g_2(\boldsymbol{\delta})$ y $Eg_3(\widehat{\boldsymbol{\delta}}) \approx g_3(\boldsymbol{\delta})$ para el orden deseado de aproximación, pero $g_1(\widehat{\boldsymbol{\delta}})$ no es el estimador correcto de $g_1(\boldsymbol{\delta})$ porque su sesgo es generalmente del mismo orden que $g_2(\boldsymbol{\delta})$ y $g_3(\boldsymbol{\delta})$.

Para evaluar el sesgo de $g_1(\widehat{\boldsymbol{\delta}})$, hacemos un desarrollo de Taylor de $g_1(\widehat{\boldsymbol{\delta}})$ en torno a $\boldsymbol{\delta}$:

$$g_1(\widehat{\boldsymbol{\delta}}) = g_1(\boldsymbol{\delta}) + (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^T\nabla g_1(\boldsymbol{\delta}) + \frac{1}{2}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^T\nabla^2 g_1(\boldsymbol{\delta})(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = g_1(\boldsymbol{\delta}) + \Delta_1 + \Delta_2,$$

donde $\nabla g_1(\boldsymbol{\delta})$ es el vector de derivadas de primer orden de $g_1(\boldsymbol{\delta})$ con respecto a $\boldsymbol{\delta}$. Si $\widehat{\boldsymbol{\delta}}$ es insesgado para $\boldsymbol{\delta}$, entonces $E(\Delta_1) = 0$. En general, si $E(\Delta_1) \approx b_{\widehat{\boldsymbol{\delta}}}^T(\boldsymbol{\delta})\nabla g_1(\boldsymbol{\delta})$ es de orden menor que $E(\Delta_2)$, entonces

$$Eg_1(\widehat{\boldsymbol{\delta}}) \approx g_1(\boldsymbol{\delta}) + \frac{1}{2}\text{tr}[\nabla^2 g_1(\boldsymbol{\delta})\overline{\mathbf{V}}(\widehat{\boldsymbol{\delta}})], \tag{2.125}$$

donde $b_{\widehat{\boldsymbol{\delta}}}(\boldsymbol{\delta})$ es una aproximación del sesgo $E(\widehat{\boldsymbol{\delta}}) - \boldsymbol{\delta}$. Además, si la matriz de covarianzas $\mathbf{V}$ tiene una estructura lineal, (2.125) se reduce a

$$Eg_1(\widehat{\boldsymbol{\delta}}) \approx g_1(\boldsymbol{\delta}) - g_3(\boldsymbol{\delta}). \tag{2.126}$$

Se sigue ahora de (2.123), (2.124) y (2.126) que los sesgos de $\mathrm{ECM}_N[t(\widehat{\boldsymbol{\delta}})]$ y $\mathrm{ECM}_1[t(\widehat{\boldsymbol{\delta}})]$ son

$$B_N \approx -2g_3(\boldsymbol{\delta}), \quad B_1 \approx -g_3(\boldsymbol{\delta}).$$

Un estimador correcto de $\mathrm{ECM}[t(\widehat{\boldsymbol{\delta}})]$ para el orden deseado de aproximación está dado por

$$\mathrm{ecm}[t(\widehat{\boldsymbol{\delta}})] \approx g_1(\widehat{\boldsymbol{\delta}}) + g_2(\widehat{\boldsymbol{\delta}}) + 2g_3(\widehat{\boldsymbol{\delta}}), \tag{2.127}$$

notando que $E[g_1(\widehat{\boldsymbol{\delta}}) + g_3(\widehat{\boldsymbol{\delta}})] \approx g_1(\boldsymbol{\delta})$ de (2.126). Por consiguiente,

$$E\mathrm{ecm}[t(\widehat{\boldsymbol{\delta}})] \approx \mathrm{ECM}[t(\widehat{\boldsymbol{\delta}})].$$

La fórmula (2.127) se preserva para el estimador MVR, $\widehat{\boldsymbol{\delta}}_{\mathrm{RE}}$, y algunos estimadores de momentos.

Si $E(\Delta_1)$ es del mismo orden que $E(\Delta_2)$, como en el caso del estimador MV $\widehat{\boldsymbol{\delta}}_{\mathrm{MV}}$, entonces sustraemos un término extra $b^T_{\widehat{\boldsymbol{\delta}}}(\widehat{\boldsymbol{\delta}})\nabla g_1(\widehat{\boldsymbol{\delta}})$ de (2.127):

$$\mathrm{ecm}_*[t(\widehat{\boldsymbol{\delta}})] \approx g_1(\widehat{\boldsymbol{\delta}}) - \mathbf{b}^T_{\widehat{\boldsymbol{\delta}}}(\widehat{\boldsymbol{\delta}})\nabla g_1(\widehat{\boldsymbol{\delta}}) + g_2(\widehat{\boldsymbol{\delta}}) + 2g_3(\widehat{\boldsymbol{\delta}}). \tag{2.128}$$

El término $b^T_{\widehat{\boldsymbol{\delta}}}(\widehat{\boldsymbol{\delta}})\nabla g_1(\widehat{\boldsymbol{\delta}})$ deja fuera el caso especial de la matriz de covarianzas diagonal por bloques $\mathbf{V} = \mathbf{V}(\boldsymbol{\delta})$ y $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\delta}}_{\mathrm{MV}}$.

Prasad y Rao (1990) obtuvieron el estimador del ECM (2.127) para el caso especial cubierto por el modelo mixto lineal general con una estructura de covarianzas diagonal por bloques. Siguiendo a Prasad y Rao (1990), Harville y Jeske (1992), propusieron (2.127) para el modelo mixto lineal general (2.112), asumiendo $E(\widehat{\boldsymbol{\delta}}) = \boldsymbol{\delta}$, y refiriéndose a (2.127) como el estimador Prasad-Rao.

Das, Jiang y Rao (2001), proporcionan demostraciones de las aproximaciones (2.127) y (2.128) para los métodos MVR y MV, respectivamente.

2.3.2. Estructura de covarianzas diagonal por bloques

Estimador EBLUP

Un caso especial del modelo mixto lineal general (2.112) cubre muchos modelos de áreas pequeñas consideradas en la literatura. Para este modelo

$$\mathbf{y} = \mathrm{col}_{1\le d\le D}(\mathbf{y}_d) = (\mathbf{y}_1^T, \cdots, \mathbf{y}_D^T), \quad \mathbf{X} = \mathrm{col}_{1\le d\le D}(\mathbf{X}_d),$$

$$\mathbf{Z} = \mathrm{diag}_{1\le d\le D}(\mathbf{Z}_d), \quad \mathbf{v} = \mathrm{col}_{1\le d\le D}(\mathbf{v}_d), \quad \mathbf{e} = \mathrm{col}_{1\le d\le D}(\mathbf{e}_d),$$

donde D es el número de áreas pequeñas, $\mathbf{X}_d$ es de $n_d \times p$, $\mathbf{Z}_d$ es de $n_d \times h_d$ y $\mathbf{y}_d$ es un vector de $n_d \times 1$ ($\sum n_d = n, \sum h_d = h$). Además

$$\mathbf{R} = \mathrm{diag}_{1\le d\le D}(\mathbf{R}_d), \quad \mathbf{G} = \mathrm{diag}_{1\le d\le D}(\mathbf{G}_d)$$

así que $\mathbf{V}$ tiene una estructura diagonal por bloques:

$$\mathbf{V} = \text{diag}_{1\leq d\leq D}(\mathbf{V}_d)$$

con

$$\mathbf{V}_d = \mathbf{R}_d + \mathbf{Z}_d\mathbf{G}_d\mathbf{Z}_d^T.$$

El modelo, además, puede descomponerse en D submodelos

$$\mathbf{y}_d = \mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\mathbf{v}_d + \mathbf{e}_d, \quad d = 1, \cdots, D. \tag{2.129}$$

Estamos interesados en la estimación de combinaciones lineales $\mu_d = \mathbf{1}_d^T\boldsymbol{\beta} + \mathbf{m}_d^T\mathbf{v}_d, \quad d = 1, \cdots, D.$

Se sigue de (2.115) que el estimador BLUP de μ_d se reduce a

$$\tilde{\mu}_d^H = \mathbf{1}_d^T\tilde{\boldsymbol{\beta}} + \mathbf{m}_d^T\tilde{\mathbf{v}}_d, \tag{2.130}$$

donde

$$\tilde{\mathbf{v}}_d = \mathbf{G}_d\mathbf{Z}_d^T\mathbf{V}_d^{-1}(\mathbf{y}_d - \mathbf{X}_d\tilde{\boldsymbol{\beta}}), \tag{2.131}$$

y

$$\tilde{\boldsymbol{\beta}} = \left(\sum_d \mathbf{X}_d^T\mathbf{V}_d^{-1}\mathbf{X}_d\right)^{-1}\left(\sum_d \mathbf{X}_d^T\mathbf{V}_d^{-1}\mathbf{y}_d\right). \tag{2.132}$$

El estimador BLUP se extiende sin dificultad al caso del vector $\boldsymbol{\mu}_d = \mathbf{L}_d\boldsymbol{\beta} + \mathbf{M}_d\mathbf{v}_d$ para matrices especificadas $\mathbf{L}_d$ y $\mathbf{M}_d$. Está dada por

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_d^H &= \mathbf{t}_d(\boldsymbol{\delta}, \mathbf{y}) = \mathbf{L}_d\tilde{\boldsymbol{\beta}} + \mathbf{M}_d\tilde{\mathbf{v}}_d \\ &= \mathbf{L}_d\tilde{\boldsymbol{\beta}} + \mathbf{M}_d\mathbf{G}_d\mathbf{Z}_d^T\mathbf{V}_d^{-1}(\mathbf{y}_d - \mathbf{X}_d\tilde{\boldsymbol{\beta}}).\end{aligned} \tag{2.133}$$

El estimador EBLUP está dado por $\tilde{\boldsymbol{\mu}}_d^H = \mathbf{t}_d(\widehat{\boldsymbol{\delta}}, \mathbf{y})$. Puede obtenerse un estimador del ECM$(\widehat{\mu}_d^H)$ que se considera para la estimación de $\boldsymbol{\delta}$, pero los detalles se omiten aquí por simplicidad.

Se ha desarrollado una gran variedad de modelos EBLUP a nivel de área que resulta difícil incluirlos a todos. Presentamos en seguida algunos de ellos.

2.4. Modelos básicos EBLUP de nivel de área

Estos modelos cubren muchos modelos de áreas pequeñas usados en la práctica. En esta sección, aplicamos los resultados EBLUP para los modelos básico de nivel de área (tipo A) y básico de nivel unidad (tipo B), descritos anteriormente.

2.4.1. Modelo básico tipo A

En esta subsección describiremos la teorá concerniente al modelo básico de nivel de área (tipo A) (2.73) y la estimación EBLUP para el modelo mixto lineal general con estructura de covarianzas diagonal por bloques, a partir de los resultados previos.

Estimador EBLUP

El modelo básico de nivel de área es de la forma

$$\widehat{\theta}_d = \mathbf{z}_d^T\boldsymbol{\beta} + b_d v_d + e_d, \quad d = 1, \cdots, D, \tag{2.134}$$

donde $\mathbf{z}_d$ es un vector de $p \times 1$ covariables de nivel de área, $v_d \overset{\text{iid}}{\sim}(0, \sigma_v^2)$ e independiente del error de muestreo $e_d \overset{\text{ind}}{\sim}(0, \psi_d)$ con varianza conocida ψ_d, $\widehat{\theta}_d$ es un estimador directo del parámetro $\theta_d = g(\overline{Y}_d)$ en el área d, y b_d es una constante positiva conocida. El modelo (2.134) es un caso especial del modelo mixto lineal general com estructura de covarianzas diagonal por bloques, dada por (2.129). Tenemos

$$\mathbf{y}_d = \widehat{\theta}_d, \quad \mathbf{X}_d = \mathbf{z}_d^T, \quad \mathbf{Z}_d = b_d$$

y

$$\mathbf{v}_d = v_d, \quad \mathbf{e}_d = e_d, \quad \boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T.$$

Además,

$$\mathbf{G}_d = \sigma_v^2, \quad \mathbf{R}_d = \psi_d$$

así que

$$\mathbf{V}_d = \psi_d + \sigma_v^2 b_d^2.$$

También, $\mu_d = \theta_d = \mathbf{z}_d^T\boldsymbol{\beta} + b_d v_d$ así que $\mathbf{1}_d = \mathbf{z}_d$ y $\mathbf{m}_d = b_d$.

Haciendo las anteriores sustituciones en la fórmula general (2.133) para el estimador BLUP de μ_d conseguimos el estimador BLUP de θ_d como

$$\tilde{\theta}_d^H = \mathbf{z}_d^T\tilde{\boldsymbol{\beta}} + \gamma_d(\widehat{\theta}_d - \mathbf{z}_d^T\tilde{\boldsymbol{\beta}}) \tag{2.135}$$

$$= \gamma_d\widehat{\theta}_d + (1 - \gamma_d)\mathbf{z}_d^T\tilde{\boldsymbol{\beta}}, \tag{2.136}$$

donde

$$\gamma_d = \sigma_v^2 b_d^2/(\psi_d + \sigma_v^2 b_d^2) \tag{2.137}$$

y

$$\tilde{\beta} = \tilde{\beta}(\sigma_v^2) = \left[\sum_{d=1}^{D} \mathbf{z}_d \mathbf{z}_d^T/(\psi_d + \sigma_v^2 b_d^2)\right]^{-1} \left[\sum_{d=1}^{D} \mathbf{z}_d \widehat{\theta}_d/(\psi_d + \sigma_v^2 b_d^2)\right]$$

Es claro de (2.136) que el estimador BLUP, $\tilde{\theta}_d^H$, puede expresarse como un estimador combinado del estimador directo $\widehat{\theta}_d$ y el estimador sintético de regresión $\mathbf{z}_d^T\tilde{\boldsymbol{\beta}}$, donde los pesos $\gamma_d\,(0 \leq \gamma_d \leq 1)$, dados por (2.137) miden la incertidumbre en la modelanción de los θ_d, es decir, la varianza del modelo $\sigma_v^2 b_d^2$ con respecto a la varianza total $\psi_d + \sigma_v^2 b_d^2$. Así $\tilde{\theta}_d^H$ toma cuenta apropiada de la variación entre las áreas y la precisión del estimador directo. Si la varianza del modelo $\sigma_v^2 b_d^2$ es relativamente pequeña, entonces γ_d será pequeño y el mayor peso será ligado al estimador sintético. Similarmente, el mayor peso será ligado al estimador directo si la varianza de diseño ψ_d es relativamente pequeña o γ_d es grande. La fórmula (2.135) para $\tilde{\theta}_d^H$ sugiere que se ajusta al estimador sintético $\mathbf{z}_d^T\tilde{\boldsymbol{\beta}}$ para explicar a la incertidumbre del modelo.

Es importante notar que $\tilde{\theta}_d^H$ es válido para un diseño general de muestreo porque estamos modelando solamente los $\widehat{\theta}_d$ y no elementos individuales en la población, a diferencia de modelos tipo B, y el estimador directo $\widehat{\theta}_d$ usa los diseños ponderados. Además, $\tilde{\theta}_d^H$ es consistente bajo el diseño porque $\gamma_d \to 1$ cuando la varianza muestral $\psi_d \to 0$. El sesgo del diseño de $\tilde{\theta}_d^H$ está dado por

$$B_p(\tilde{\theta}_d^H) \approx (1 - \gamma_d)(\mathbf{z}_d^T\boldsymbol{\beta}^* - \theta_d), \tag{2.138}$$

donde $\boldsymbol{\beta}^* = E_2(\tilde{\beta})$ es la esperanza condicional de $\tilde{\beta}$ dado $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D)^T$. Se sigue de (2.138) que el sesgo del diseño respecto a θ_d tiende a cero cuando $\psi_d \to 0$ o $\gamma_d \to 1$. Notemos que $E_m(\mathbf{z}_d^T\boldsymbol{\beta}^*) = E_m(\theta_d)$ lo que implica que el sesgo promedio es cero cuando el modelo de enlace $\theta_d = \mathbf{z}_d^T\boldsymbol{\beta} + b_d v_d$ se conserva.

El estimador BLUP (2.136) depende de las componentes de varianza σ_v^2 las cuales son desconocidas en aplicaciones prácticas. Reemplazando σ_v^2 por un estimador $\widehat{\sigma}_v^2$, obtenemos un estimador EBLUP $\tilde{\theta}_d^H$:

$$\tilde{\theta}_d^H = \widehat{\gamma}_d\widehat{\theta}_d + (1 - \widehat{\gamma}_d)\mathbf{z}_d^T\widehat{\boldsymbol{\beta}}_d, \tag{2.139}$$

donde $\widehat{\gamma}_d$ y $\widehat{\beta}$ son los valores de γ_d y $\tilde{\boldsymbol{\beta}}_d$ cuando σ_v^2 es reemplazada por $\widehat{\sigma}_v^2$.

Fay y Herriot (1979), recomendaron el uso de un estimador imparcial EBLUP, $\tilde{\theta}_{d*}^H$, similar al estimador imparcial de James y Stein (J-S). Es decir:

(i) usar $\tilde{\theta}_d^H$ si $\tilde{\theta}_d^H$ cae el el intervalo $[\widehat{\theta}_d - c\sqrt{\psi_d},\ \widehat{\theta}_d + c\sqrt{\psi_d}\,]$ para una constante específica c (típicamdente $c = 1$);

(ii) usar $\widehat{\theta}_d^H - c\sqrt{\psi_d}$ si $\widehat{\theta}_d^H$ es menor que $\widehat{\theta}_d - c\sqrt{\psi_d}$;

(iii) usar $\widehat{\theta}_d^H + c\sqrt{\psi_d}$ si $\widehat{\theta}_d^H$ es mayor que $\widehat{\theta}_d + c\sqrt{\psi_d}$.

El estimador imparcial $\widehat{\theta}_{d*}^H$ (o $\widehat{\theta}_d^H$) es transformado a la escala original para obtener un estimador de la media $\overline{Y}_d$ del área d como $g^{-1}(\widehat{\theta}_{d*}^H)$ (o $g^{-1}(\widehat{\theta}_d^H)$).

Estimación de σ_v^2

Un método de estimación de momentos de σ_{vm}^2 puede obtenerse notando que

$$E\left[\sum_d (\widehat{\theta}_d - \mathbf{z}_d^T\tilde{\boldsymbol{\beta}})/(\psi_d + \sigma_v^2 b_d^2)\right] = E[h(\sigma_v^2)] = D - p$$

donde $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_v^2)$. Se sigue que $\widehat{\sigma}_{vm}^2$ se obtiene resolviendo

$$h(\sigma_v^2) = D - p$$

iterativamente y fijando $\widehat{\sigma}_{vm}^2 = 0$ cuando no existe solución positiva. Fay y Herriot (1979), sugirieron la siguiente solución iterativa: empezando con $\sigma_v^{2(0)} = 0$, se define

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + \frac{1}{h_*'(\sigma_v^{2(a)})}[D - p - h(\sigma_v^{2(a)})] \tag{2.140}$$

restringiendo $\sigma_v^{2(a+1)} \geq 0$, donde

$$h_*'(\sigma_v^2) = -\sum_d b_d^2(\widehat{\theta}_d - \mathbf{z}_d^T\tilde{\boldsymbol{\beta}})^2/(\psi_d + \sigma_v^2 b_d^2)^2$$

es una aproximación de la derivada de $h(\sigma_v^2)$. La convergencia de la iteración es rápida, generalmente requiere menos de 10 iteraciones.

Alternativamente, un estimador simple de momentos está dado por $\widehat{\sigma}_{vs}^2 = \text{máx}(\tilde{\sigma}_{vs}^2, 0)$, donde

$$\tilde{\sigma}_{vs}^2 = \frac{1}{D-p}\left[\sum_d (b_d^{-1}\widehat{\theta}_d - \tilde{\mathbf{z}}_d^T\widehat{\boldsymbol{\beta}}_{\text{WLS}})^2 - \sum_d \frac{\psi_d}{b_d^2}(1 - \tilde{h}_{dd})\right] \tag{2.141}$$

y

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} = \left(\sum_d \tilde{\mathbf{z}}_d\tilde{\mathbf{z}}_d^T\right)^{-1}\left(\sum_d \tilde{\mathbf{z}}\widehat{\theta}_d/b_d\right)$$

es un estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$. Si $b_i = 1$, entonces (2.141) se reduce a la fórmula de Prasad y Rao (1990). Ningún estimador

de momentos de σ_v^2 requiere normalidad y los dos conducen a estimadores consistentes cuando $D \to \infty$.

El algoritmo scoring para estimación MV de σ_v^2 se reduce a

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + [\mathcal{I}(\sigma_v^{2(a)})]^{-1} s(\tilde{\boldsymbol{\beta}}^{(a)}, \sigma_v^{2(a)})$$

donde

$$\mathcal{I}(\sigma_v^2) = \frac{1}{2} \sum_{d=1}^{D} \frac{b_d^4}{(\sigma_v^2 b_d^2 + \psi)^2} \tag{2.142}$$

y

$$s(\tilde{\boldsymbol{\beta}}, \sigma_v^2) = -\frac{1}{2} \sum_{d=1}^{D} \frac{b_d^2}{\sigma_v^2 b_d^2 + \psi} + \frac{1}{2} b_d^2 \frac{(\widehat{\theta}_d - \mathbf{z}_d^T \tilde{\boldsymbol{\beta}})^2}{(\sigma_v^2 b_d^2 + \psi)^2}.$$

similarmente, el algoritmo scoring para estimación MVR de σ_v^2 se reduce a

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + [\mathcal{I}_R(\sigma_v^{2(a)})]^{-1} s_R(\sigma_v^{2(a)}),$$

donde

$$\mathcal{I}_R(\sigma_v^2) = \frac{1}{2} \mathrm{tr}[\mathbf{PBPB}] \tag{2.143}$$

y

$$s_R(\sigma_v^2) = -\frac{1}{2} \mathrm{tr}[\mathbf{PB}] + \frac{1}{2} \mathbf{y}^T \mathbf{PBPy},$$

donde $\mathbf{B} = \mathrm{diag}(b_1^2, \cdots, b_D^2)$ y $\mathbf{P}$ se han definido antes; ver Cressie (1992).

El estimador EBLUP de $\tilde{\theta}_d^H$, basado en los momentos, el estimador MV o MVR de σ_v^2, permanecen insesgados bajo el modelo si los errores v_d y e_d están simétricamente distribuidos en torno a 0. En particular, $\tilde{\theta}_d^H$ es insesgado bajo el modelo para θ_d si v_d y e_d son normalmente distribuidos.

Para el caso especial de $b_d = 1$ e iguales varianzas muestrales $\psi_d = \psi$, el estimador BLUP (2.136) se reduce a

$$\tilde{\theta}_d^H = \gamma \widehat{\theta}_d + (1 - \gamma) \mathbf{z}_d^T \widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$$

con $1 - \gamma = \psi/(\psi + \sigma_v^2)$, donde $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ es el estimador de mínimos cuadrados de $\boldsymbol{\beta}$. Bajo normalidad, podemos obtener un estimador insesgado $1 - \gamma^*$ de $1 - \gamma$ notando que $S/(\psi + \sigma_v^2)$ es una variable χ^2 con $D - p$ grados de libertad, donde $S = \sum_d (\widehat{\theta}_d - \mathbf{z}_d^T \widehat{\boldsymbol{\beta}}_{\mathrm{LS}})^2$ es la suma de cuadrados residual. Tenemos

$$1 - \gamma^* = \psi(D - p - 2)/S,$$

y un estimador EBLUP de θ_d está dado por lo tanto por

$$\tilde{\theta}_d^H = \gamma^* \widehat{\theta}_d + (1 + \gamma^*) \mathbf{z}_d^T \widehat{\boldsymbol{\beta}}_{\mathrm{LS}}.$$

Este estimador es idéntico al estimador de James-Stein con valor inicial supuesto $\theta_d^0 = \mathbf{z}_d^T \widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$. Nótese que ambos estimadores simples de momentos, $\widehat{\sigma}_{vs}^2$ y $\widehat{\sigma}_{vm}^2$, conducen a

$$1 - \widehat{\gamma} = \psi(D-p)/S,$$

el cual es aproximadamente igual a $1-\gamma^*$ para D grande y p fijo.

2.4.2. Modelo básico tipo B

Consideraremos ahora el modelo básico de nivel unidad (tipo B) (2.95) y describiremos aquí la estimación EBLUP, usando el resultado general para el modelo mixto lineal general con estructura de covarianzas diagonal por bloques.

Como establecimos antes, podemos tomar la media del área pequeña d como $\mu_d = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + v_d$ si los tamaños de la población, N_d, de las áreas pequeñas son suficientemente grandes. En este caso, podemos usar la parte de la muestra del modelo (2.95), es decir, $y_{dj} = \mathbf{x}_d^T \beta + v_d + e_{dj}$, $j = 1, \cdots, n_d; d = 1, \cdots, D$ el cual puede escribirse en notación matricial como

$$\mathbf{y}_d = \mathbf{X}_d \boldsymbol{\beta} + v_d \mathbf{1}_{n_d} + \mathbf{e}_d, \quad d = 1, \cdots, D \tag{2.144}$$

para hacer inferencias sobre $\overline{Y}_d$, recurriendo al resultado general. En el caso de fracción de muestreo no despreciable, n_d/N_d, se manejará recurriendo al modelo poblacional (2.95).

Estimador BLUP

El modelo (2.144) es un caso especial del modelo general (2.129) con estructura de covarianzas diagonal por bloques. Tenemos

$$\mathbf{y}_d = \mathbf{y}_d, \quad \mathbf{X}_d = \mathbf{X}_d, \quad \mathbf{Z}_d = \mathbf{1}_{n_d},$$

$$\mathbf{v}_d = v_d, \quad \mathbf{e}_d = \mathbf{e}_d, \quad \boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T,$$

donde $\mathbf{y}_d$ es el vector de $n_d \times 1$ observaciones muestrales y_{dj} de el área d. Además,

$$\mathbf{G}_d = \sigma_v^2, \quad \mathbf{R}_d = \sigma_e^2 \mathrm{diag}_{1 \le j \le n_d}(k_{dj}^2)$$

de modo que

$$\mathbf{V}_d = \mathbf{R}_d + \sigma_v^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T.$$

También, $\mu_d = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + v_d$ así que $\mathbf{1}_d = \overline{\mathbf{X}}_d$ y $\mathbf{m}_d = 1$. La matriz $\mathbf{V}_d$ puede invertirse explícitamente como

$$\mathbf{V}_d^{-1} = \frac{1}{\sigma_e^2} \left[\mathrm{diag}_j(a_{dj}) - \frac{\gamma_d}{a_{d\cdot}} \mathbf{a}_d \mathbf{a}_d^T \right] \tag{2.145}$$

usando el siguiente resultado estándar sobre inversión de matrices:

$$(\mathbf{A}+\mathbf{u}\mathbf{v}^T)^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}/(1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}). \tag{2.146}$$

Aquí tenemos

$$a_{dj}=k_{dj}^{-2},\quad a_{d\cdot}=\sum_j a_{dj},\quad \mathbf{a}_d=(a_{d1},\cdots,a_{dn_d})^T$$

y

$$\gamma_d=\sigma_v^2/(\sigma_v^2+\sigma_e^2/a_{d\cdot}). \tag{2.147}$$

Haciendo las anteriores sustituciones en la fórmula general (2.130) y notando que $(\sigma_v^2/\sigma_e^2)(1-\gamma_d)=\gamma_d/a_{d\cdot}$, obtenemos el estimador EBLUP de μ_d como

$$\tilde{\mu}_d^H=\overline{\mathbf{X}}_d^T\tilde{\boldsymbol{\beta}}+\gamma_d(\overline{y}_{da}-\overline{\mathbf{x}}_{da}^T\tilde{\boldsymbol{\beta}}), \tag{2.148}$$

donde $\overline{y}_{da}$ y $\overline{\mathbf{x}}_{da}$ son medias ponderadas dadas por

$$\overline{y}_{da}=\sum_j a_{dj}y_{dj}/a_{d\cdot},\quad \overline{\mathbf{x}}_{da}=\sum_j a_{dj}\mathbf{x}_{dj}/a_{d\cdot},$$

y $\tilde{\boldsymbol{\beta}}$ es el BLUE de $\boldsymbol{\beta}$:

$$\tilde{\boldsymbol{\beta}}=\left(\sum_d \mathbf{X}_d^T\mathbf{V}_d^{-1}\mathbf{X}_d\right)^{-1}\left(\sum_d \mathbf{X}_d^T\mathbf{V}_d^{-1}\mathbf{y}_d\right), \tag{2.149}$$

donde

$$\mathbf{X}_d^T\mathbf{V}_d^{-1}\mathbf{X}_d=\mathbf{A}_d=\sigma_e^{-2}\left(\sum_j a_{dj}\mathbf{x}_{dj}\mathbf{x}_{dj}^T-\gamma_d a_{d\cdot}\overline{\mathbf{x}}_{da}\overline{\mathbf{x}}_{da}^T\right) \tag{2.150}$$

y

$$\mathbf{X}_d^T\mathbf{V}_d^{-1}\mathbf{y}_d=\sigma_e^{-2}\left(\sum_j a_{dj}\mathbf{x}_{dj}y_{dj}-\gamma_d a_{d\cdot}\overline{\mathbf{x}}_{da}\overline{y}_{da}\right). \tag{2.151}$$

El estimador BLUP (2.148) puede expresarse también como un estimador combinado del estimador de "regresión de encuesta" $\overline{y}_{da}+(\overline{\mathbf{X}}_d-\overline{\mathbf{x}}_{da})^T\tilde{\boldsymbol{\beta}}$ y el estimador de regresión sintético $\overline{\mathbf{X}}_d^T\tilde{\boldsymbol{\beta}}$:

$$\tilde{\mu}_d^H=\gamma_d[\overline{y}_{da}+(\overline{\mathbf{X}}_d-\overline{\mathbf{x}}_{da})^T\tilde{\boldsymbol{\beta}}]+(1-\gamma_d)\overline{\mathbf{X}}_d^T\tilde{\boldsymbol{\beta}}. \tag{2.152}$$

El peso γ_d mide la variación del modelo, σ_v^2, respecto a la varianza total $\sigma_v^2+\sigma_e^2/a_{d\cdot}$. Si la varianza del modelo es relativamente pequeña, entonces γ_d

será pequeña y se asignará mayor peso al componente sintético. Similarmente, el mayor peso corresponderá al estimador de regresión de encuesta cuando $a_{d.}$ aumenta.

El $\tilde{\boldsymbol{\beta}}$ BLUE y su matriz de covarianzas $(\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d)^{-1}$ pueden calcularse usando solamente mínimos cuadrados ordinarios transformando primero el modelo (2.144) con errores correlados $u_{dj} = v_d + e_{dj}$ a un modelo con errores no correlados u_{dj}^*. El modelo transformado está dado por

$$k_{dj}^{-1}(y_{dj} - \tau_d \overline{y}_{da}) = k_{dj}^{-1}(\mathbf{x}_{dj} - \tau_d \overline{\mathbf{x}}_{da})^T \boldsymbol{\beta} + u_{dj}^*, \tag{2.153}$$

donde $\tau_d = 1 - (1 - \gamma_d)^{1/2}$ y los u_{dj}^* tienen media cero y varianza constante σ_e^2 (Stukel y Rao, 1997). Si $k_{dj} = 1$ para todo (d, j), (2.153) se reduce al modelo transformado de Fuller y Battese (1973).

El estimador BLUP (2.152) depende de la razón de varianzas σ_v^2/σ_e^2, la cual es desconocida en la práctica. Reemplazando σ_v^2 y σ_e^2 por estimadores $\widehat{\sigma}_v^2$ y $\widehat{\sigma}_e^2$, obtenemos un estimador EBLUP.

$$\widehat{\mu}_d^H = \widehat{\gamma}_d[\overline{y}_{da} + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_{da})^T \widehat{\boldsymbol{\beta}}] + (1 - \widehat{\gamma}_d)\overline{\mathbf{X}}_d^T \widehat{\boldsymbol{\beta}} \tag{2.154}$$

donde $\widehat{\gamma}_d$ y $\widehat{\boldsymbol{\beta}}$ son los valores de γ_d y $\tilde{\boldsymbol{\beta}}$ cuando (σ_v^2, σ_e^2) es reemplazada por $(\widehat{\sigma}_v^2, \widehat{\sigma}_e^2)$.

Estimación de σ_v^2 y σ_e^2

Presentamos un método simple de estimación de las componentes de varianza σ_v^2 y σ_e^2. Involucra la presentación de regresión de mínimos cuadrados ordinarios y el método de momentos, dos métodos para obtener estimadores de σ_v^2 y σ_e^2 (Stukel y Rao, 1997). Fuller y Battese (1973), propusieron este método para el caso especial $k_{dj} = 1$ para todo (d, j).

Primero calculamos la suma de cuadrados residual SCR(1) con ν_1 grados de libertad haciendo la regresión a través del origen y de las desviaciones $k_{dj}^{-1}(y_{dj} - \overline{y}_{da})$ sobre las desviaciones $\mathbf{x}$ distintas de cero $k_{dj}^{-1}(\mathbf{x}_{dj} - \overline{\mathbf{x}}_{da})$ para aquellas áreas con $n_d > 1$. Esto lleva a un estimador insesgado de σ_e^2:

$$\widehat{\sigma}_{em}^2 = \nu_1^{-1}\text{SCR}(1) \tag{2.155}$$

donde $\nu_1 = n - D - p_1$ y p_1 es el número de las desviaciones de $\mathbf{x}$ distintas de cero. Calculamos en seguida la suma de cuadrados residual SCR(2) por regresión de y_{dj}/k_{dj} sobre $\mathbf{x}_{dj}/k_{dj}$. Un estimador insesgado de σ_v está dado entonces por

$$\tilde{\sigma}_{vm}^2 = \eta_1^{-1}[\text{SCR}(2) - (n - p)\widehat{\sigma}_e^2], \tag{2.156}$$

notando que

$$E[\text{SCR}(2)] = \eta_1 \sigma_v^2 + (n - p)\sigma_e^2,$$

donde

$$\eta_1 = \sum_d a_{d\cdot} \left[1 - a_{d\cdot} \mathbf{x}_{da}^T \left(\sum_d \sum_j a_{dj} \mathbf{x}_{dj} \mathbf{x}_{dj}^T \right)^{-1} \overline{\mathbf{x}}_{da} \right]. \tag{2.157}$$

Los estimadores $\tilde{\sigma}_{vm}^2$ y $\widehat{\sigma}_{em}^2$ son equivalentes a aquellos encontrados usando el método de "ajuste de constantes" atribuido a Henderson (1953). Sin embargo, el método anterior requiere regresión por mínimos cuadrados sobre $p_1 + D$ variables, en contraste con p_1 variables para el método de transformación, y así es computacionalmente más engorroso cuando el número de áreas pequeñas, D, aumenta.

Dado que $\tilde{\sigma}_{vm}^2$ puede tomar valores negativos, truncamos $\tilde{\sigma}_{vm}^2$ para poner cero siempre que sea negativo. El estimador truncado $\tilde{\sigma}_{vm}^2 = \text{máx}(\tilde{\sigma}_{vm}^2, 0)$ no es menos sesgado pero es consistente cuando D crece. Para el caso especial de $k_{dj} = 1$ para todo (d, j), Battese, Harter y Fuller (1988), propusieron un estimador alternativo de γ_d el cual es aproximadamente insesgado para γ_d.

Asumiendo normalidad de los errores v_d y e_d, pueden también emplearse MV o MVR . Una estimación sencilla de ECM,

$$\text{ecm}_N(\widehat{\mu}_d^H) = g_{1d}(\widehat{\sigma}_v^2, \widehat{\sigma}_e^2) + g_{2d}(\widehat{\sigma}_v^2, \widehat{\sigma}_e^2), \tag{2.158}$$

puede calcularse usando el procedimiento PROC MIXED en SAS, donde$(\widehat{\sigma}_v^2, \widehat{\sigma}_e^2)$ son los estimadores MV o MVR de $g_{3d}(\sigma_v^2, \sigma_e^2)$.

Fracciones de muestreo no despreciables

Si la fracción de muestreo $f_d = n_d/N_d$ no es despreciable, entonces no podemos tomar la media de área pequeña $\overline{Y}_d$ como $\overline{\mathbf{X}}\boldsymbol{\beta} + v_d$. Sin embargo, podemos escribir $\overline{Y}_d$ como

$$\overline{Y}_d = f_d \overline{y}_d + (1 - f_d)\overline{y}_d^*,$$

donde $\overline{y}_d$ es la media muestral y $\overline{y}_d^*$ es la media de los valores no muestreados y_{dj}^*, de el área d. Bajo el modelo poblacional (2.95), reemplazamos las y_{dj}^* no observadas por su estimador $\mathbf{x}_{dj}^{*T}\tilde{\boldsymbol{\beta}} + \tilde{v}_d$, donde $\mathbf{x}_{dj}^*$ son los valores $\mathbf{x}$ asociados con y_{dj}^*. El estimador BLUP resultante de $\overline{Y}_d$ está dado por

$$\tilde{\overline{Y}}_d^H = f_d \overline{y}_d + (1 - f_d)(\overline{\mathbf{x}}_d^{*T}\tilde{\boldsymbol{\beta}} + \tilde{v}_d) = f_d \overline{y}_d + (1 - f_d)\tilde{\overline{y}}_d^{*H}, \tag{2.159}$$

donde $\overline{\mathbf{x}}_d^*$ es la media de los valores no muestreados $\mathbf{x}_{dj}^*$ y

$$\tilde{\overline{y}}_d^{*H} = \gamma_d[\overline{y}_{da} + (\overline{\mathbf{x}}_d^* - \overline{\mathbf{x}}_{da})^T\tilde{\boldsymbol{\beta}}] + (1 - \gamma_d)\overline{\mathbf{x}}_d^{*T}\tilde{\boldsymbol{\beta}}. \tag{2.160}$$

Nótese que $\overline{\mathbf{x}}_d^* = (N_d\overline{\mathbf{X}}_d - n_d\overline{\mathbf{x}}_d)/(N_d - n_d)$ puede calcularse conociendo solamente la media de la población $\overline{\mathbf{X}}_d$. La propiedad BLUP de $\tilde{\overline{Y}}_d^H$ se establece fácilmente mostrando que $\mathrm{Cov}(\mathbf{b}^T\mathbf{y}, \tilde{\overline{Y}}_d^H - \overline{Y}_d) = 0$ para toda función lineal cero $\mathbf{b}^T\mathbf{y}$, es decir, $E(\mathbf{b}^T\mathbf{y}) = 0$ (Stukel, 1991).

Algunas Extensiones EBLUP para modelos de áreas pequeñas

Existen varias extensiones de los modelos EBLUP para áreas pequeñas. En seguida presentaremos un breve resumen de ellas.

Modelo Multivariante de Fay-Herriot: El modelo multivariante de Fay-Herriot está dado por (2.77). Este modelo puede también expresarse como un caso especial del modelo general (2.129) con estructura de covarianzas diagonal por bloques. Tenemos

$$\mathbf{y}_d = \widehat{\boldsymbol{\theta}}_d, \quad \mathbf{X}_d = \mathbf{Z}_d, \quad \mathbf{Z}_d = \mathbf{I}_r$$

y

$$\mathbf{v}_d = \mathbf{v}_d, \quad \mathbf{e}_d = \mathbf{e}_d, \quad \boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T,$$

donde $\widehat{\boldsymbol{\theta}}_d = (\widehat{\theta}_{d1}, \cdots, \widehat{\theta}_{dr})^T$ es un vector de r estimadores directos para el área d. Además,

$$\mathbf{G}_d = \boldsymbol{\Sigma}_v, \quad \mathbf{R}_d = \boldsymbol{\Psi}_d,$$

para que

$$\mathbf{V}_d = \boldsymbol{\Psi}_d + \boldsymbol{\Sigma}_v.$$

También, $\boldsymbol{\mu}_d = \boldsymbol{\theta}_d = \mathbf{L}_d\boldsymbol{\beta} + \mathbf{M}_d\mathbf{v}_d$, con $\mathbf{L}_d = \mathbf{Z}_d$ y $\mathbf{M}_d = \mathbf{I}_r$.

Haciendo la anterior sustitución en la fórmula general (2.133), obtenemos el estimador BLUP de $\boldsymbol{\theta}_d$ como

$$\tilde{\boldsymbol{\theta}}_d^H = \boldsymbol{\Sigma}_v(\boldsymbol{\Psi}_d + \boldsymbol{\Sigma}_v)^{-1}\widehat{\boldsymbol{\theta}}_d + \boldsymbol{\Psi}_d(\boldsymbol{\Psi}_d + \boldsymbol{\Sigma}_v)^{-1}\mathbf{Z}_d\tilde{\boldsymbol{\beta}}, \tag{2.161}$$

donde

$$\tilde{\boldsymbol{\beta}} = \left[\sum_{d=1}^{D}\mathbf{Z}_d^T(\boldsymbol{\Psi}_d + \boldsymbol{\Sigma}_v)^{-1}\mathbf{Z}_d\right]^{-1}\left[\sum_{d=1}^{D}\mathbf{Z}_d^T(\boldsymbol{\Psi}_d + \boldsymbol{\Sigma}_v)^{-1}\widehat{\boldsymbol{\theta}}_d\right]. \tag{2.162}$$

El estimador (2.161) es una extensión natural del estimador BLUP univariante (2.136).

El estimador EBLUP $\tilde{\boldsymbol{\theta}}_d^H$ se obtiene sustituyendo un estimador $\widehat{\boldsymbol{\Sigma}}_v$ para $\boldsymbol{\Sigma}_v$ en (2.161).

Errores de Muestreo Correlados: El modelo Fay-Herriot con errores de muestreo correlados está dado por (2.79). Es un caso especial del modelo mixto lineal general (2.112) con $\mathbf{R} = \mathbf{\Psi}$, $\mathbf{Z} = \mathbf{I}, \mathbf{G} = \sigma_v^2 \mathbf{I}_D, \mathbf{X} = \mathbf{Z}$ y $\mathbf{y} = \widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \cdots, \widehat{\theta}_D)^T$. Usando estos valores en el estimador BLUP general con $\mathbf{L} = \mathbf{Z}$ y $\mathbf{M} = \mathbf{I}$, obtenemos el estimador BLUP de $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D)^T$ como

$$\tilde{\boldsymbol{\theta}}^H = \mathbf{Z}\tilde{\boldsymbol{\beta}} + \sigma_v^2 \mathbf{V}^{-1}(\widehat{\boldsymbol{\theta}} - \mathbf{Z}\tilde{\boldsymbol{\beta}}), \tag{2.163}$$

donde

$$\tilde{\boldsymbol{\beta}} = (\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z})^{-1} (\mathbf{Z}^T \mathbf{V}^{-1} \widehat{\boldsymbol{\beta}})$$

y

$$\mathbf{V} = \mathbf{\Psi} + \sigma_v^2 \mathbf{I}.$$

El estimador EBLUP $\widehat{\boldsymbol{\theta}}^H$ se obtiene asumiendo $\mathbf{\Psi}$ conocida y sustituyendo un estimador de $\widehat{\sigma}_v^2$ para σ_v^2.

Isaki, Tsay y Fuller (2000) relajaron el supuesto de una matriz de covarianzas muestral conocida $\mathbf{\Psi}$. Reemplazaron $\mathbf{\Psi}$ en el estimador EBLUP $\widehat{\boldsymbol{\theta}}^H$ por

$$\widehat{\mathbf{\Psi}}_\phi = \phi \widehat{\mathbf{\Psi}}_d + (1 - \phi) \widehat{\mathbf{\Psi}}, \tag{2.164}$$

donde $\widehat{\mathbf{\Psi}}$ es un estimador basado en la muestra de $\mathbf{\Psi}$ con elementos en la diagonal $\widehat{\psi}_{dd}$, $d = 1, \cdots, D$, $\widehat{\mathbf{\Psi}}_d = (\psi_{dd}, d = 1, \cdots, D)$ y ϕ es una constante especificada, $0 \leq \phi \leq 1$.

Series de tiempo y modelos transversales

Modelo de Rao-Yu: El modelo dado por (2.80) y (2.81) con AR(1) o modelo de salto aleatorio sobre los errores u_{dt} proporciona una extensión del modelo básico tipo A para manejar series de tiempo y datos transversales. Notando los estimadores directos $\widehat{\theta}_{dt}$ como $\widehat{\boldsymbol{\theta}}_d = (\widehat{\theta}_{d1}, \cdots, \widehat{\theta}_{dT})^T, d = 1, \cdots, D$, podemos escribir el modelo en la forma (2.129) con estructura de covarianzas diagonal por bloques. Tenemos

$$\mathbf{y}_d = \widehat{\boldsymbol{\theta}}_d, \quad \mathbf{X}_d = (\mathbf{z}_{d1}, \cdots, \mathbf{z}_{dT})^T, \quad \mathbf{Z}_d = (\mathbf{1}_T, \mathbf{I}_T),$$

$$\mathbf{v}_d^T = (v_d, \mathbf{u}_d^T), \quad \mathbf{e}_d = (\mathbf{e}_{d1}, \cdots, \mathbf{e}_{dT})^T$$

y $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, donde $\mathbf{1}_T$ es el vector de 1s, e $\mathbf{I}_T$ es la matriz identidad de orden T. Además,

$$\mathbf{G}_d = \begin{bmatrix} \sigma_v^2 & \mathbf{0}^T \\ \mathbf{0} & \sigma^2 \mathbf{\Lambda} \end{bmatrix}, \quad \mathbf{R}_d = \mathbf{\Psi}_d$$

donde $\mathbf{\Lambda}_d = \mathbf{\Lambda}$ es la matriz de covarianzas de $T \times T$ de $\mathbf{u}_d = (u_{d1}, \cdots, u_{dT})^T$ con (t,s)-ésimo elemento $\rho^{|t-s|}/(1-\rho^2)$ para el modelo AR(1) $u_{dt} = \rho u_{d,t-1} + \epsilon_{dt}$, $|\rho| < 1$; y el (t,s)-ésimo elemento mín(t,s) para el modelo de salto aleatorio $u_{dt} = u_{d,t-1} + \epsilon_{dt}$. Podemos escribir $\mathbf{V}_d$ como

$$\mathbf{V}_d = \mathbf{\Psi}_d + \sigma^2 \mathbf{\Lambda} + \sigma_v^2 \mathbf{J}_T$$

donde $\mathbf{J}_T = \mathbf{1}_T \mathbf{J}_T^T$ denota una matriz con todos sus elementos iguales a uno. Además, los parámetros de áreas pequeñas θ_{dT} para el valor actual T puede expresarse como $\theta_{dT} = \mu_d = \mathbf{1}_i^T \boldsymbol{\beta} + \mathbf{m}_d^T \mathbf{v}_d$ con $\mathbf{1}_d = \mathbf{z}_{dT}$ y $\mathbf{m}_d = (1, 0, \cdots, 0, 1)^T$.

Haciendo la anterior sustitución en la fórmula general BLUP (2.130), obtenemos el estimador BLUP de θ_{dT} como

$$\tilde{\theta}_{dT}^H = \mathbf{z}_{dT}^T \tilde{\boldsymbol{\beta}} + (\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\lambda}_T)^T \mathbf{V}_d^{-1} (\widehat{\theta}_d - \mathbf{X}_d \tilde{\boldsymbol{\beta}}), \tag{2.165}$$

donde $\boldsymbol{\lambda}_T^T$ es la T-ésima fila de $\mathbf{\Lambda}$ y

$$\tilde{\boldsymbol{\beta}} = \left(\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \mathbf{X}_d \right)^{-1} \left(\sum_d \mathbf{X}_d^T \mathbf{V}_d^{-1} \widehat{\boldsymbol{\theta}}_d \right).$$

El estimador BLUP (2.165) puede expresarse también como una combinación ponderada del estimador directo de $\widehat{\theta}_{dT}$, del estimador sintético $\mathbf{z}_{dT}^T \tilde{\boldsymbol{\beta}}$ y los residuos $\widehat{\theta}_{dt} - \mathbf{z}_{dT}^T \tilde{\boldsymbol{\beta}}$, $t = 1, \cdots, T-1$:

$$\tilde{\theta}_{dT}^H = w_{dT}^* \widehat{\theta}_{dT} + (1 - w_{dT}^*) \mathbf{z}_{dT}^T \tilde{\boldsymbol{\beta}} + \sum_{t=1}^{T-1} w_{dt}^* (\widehat{\theta}_{dt} - \mathbf{z}_{dt}^T \tilde{\boldsymbol{\beta}}), \tag{2.166}$$

donde $(w_{d1}^*, \cdots, w_{dT}^*) = (\sigma_v^2 \mathbf{1}_T + \sigma^2 \mathbf{\Lambda}_T)^T \mathbf{V}_d^T$; $d = 1, \cdots, D$.

Reemplazando σ^2 y σ_v^2 por sus estimadores $\widehat{\sigma}^2(\rho)$ y $\widehat{\sigma}_v^2(\rho)$ en (2.165), obtenemos el estimador EBLUP $\widehat{\theta}_{dT}^H(\rho)$ bajo el modelo AR(1) con ρ conocido. De forma similar se obtiene el estimador EBLUP, $\widehat{\theta}_{dT}^H$, bajo el modelo de salto aleatorio. El estimador EBLUP no requiere normalidad de los errores; solamente se necesitan errores simétricamente distribuidos.

Los estimadores del ECM de $\widehat{\theta}_{dT}^H(\rho)$ y $\widehat{\theta}_{dT}^H$ para términos de orden $o(D^{-1})$, pueden obtenerse del resultado general obtenido antes.

Modelos de Espacio de Estado: El modelo de nivel de área dado por (2.87)-(2.89), de estimación indirecta de dominios, permiten a los coeficientes del modelo variar sobre el tiempo y sobre la sección. Es un caso especial del modelo general de espacio de estado el cual puede expresarse en la forma

$$\mathbf{y}_t = \mathbf{Z}_t \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t; \quad E(\boldsymbol{\varepsilon}_t) = \mathbf{0}, \quad E(\varepsilon_t, \varepsilon_t^T) = \mathbf{\Sigma}_t \tag{2.167}$$

$$\boldsymbol{\alpha}_t = \mathbf{H}\boldsymbol{\alpha}_{t-1} + \mathbf{A}\boldsymbol{\eta}_t; \quad E(\boldsymbol{\eta}_t) = \mathbf{0}, \quad E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t^T) = \boldsymbol{\Gamma} \tag{2.168}$$

donde $\boldsymbol{\varepsilon}_t$ y $\boldsymbol{\eta}_t$ son incorreladas sobre el tiempo. Este modelo es un caso especial del modelo mixto lineal general pero, la forma espacio de estado permite la actualización de las estimaciones sobre el tiempo, usando las ecuaciones del filtro de Kalman (2.170)y (2.171), y suavizando el estimador anterior cuando los nuevos datos están disponibles, usando un algoritmo apropiado de suavizado. El vector $\boldsymbol{\alpha}_t$ es conocido como el *vector de estado*, (2.168) como la *ecuación de transición* y (2.167) como la *ecuación de medida*.

Sea $\tilde{\boldsymbol{\alpha}}_{t-1}$ el estimador BLUP de $\boldsymbol{\alpha}_{t-1}$ basado en todos los datos observados al tiempo $t-1$, tal que $\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathbf{H}\tilde{\boldsymbol{\alpha}}_{t-1}$ es el BLUP de $\boldsymbol{\alpha}_t$ al tiempo $t-1$. Además,

$$\mathbf{P}_{t|t-1} = \mathbf{H}\mathbf{P}_{t-1}\mathbf{H}^T + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$$

es la matriz de covarianzas de los errores de predicción $\tilde{\boldsymbol{\alpha}}_{t|t-1} - \boldsymbol{\alpha}_t$, donde $\mathbf{P}_{t-1} = E(\tilde{\boldsymbol{\alpha}}_{t-1} - \boldsymbol{\alpha}_{t-1})(\tilde{\boldsymbol{\alpha}}_{t-1} - \boldsymbol{\alpha}_{t-1})^T$ es la matriz de covarianzas de los errores de predicción al tiempo $t-1$. Este resultado se sigue sin dificultad de (2.168). A veces t, el predictor de $\boldsymbol{\alpha}_t$ y su matriz de covarianzas son actualizadas usando los nuevos datos $(\mathbf{y}_t, \mathbf{Z}_t)$. Tenemos

$$\mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathbf{Z}_t(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}) + \boldsymbol{\varepsilon}_t, \tag{2.169}$$

el cual tiene la forma del modelo lineal mixto (2.112) con $\mathbf{y} = \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}, \mathbf{Z} = \mathbf{Z}_t, \mathbf{v} = \boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}, \mathbf{G} = \mathbf{P}_{t|t-1},\ \mathbf{X}\boldsymbol{\beta}$ ausente y $\mathbf{V} = \mathbf{F}_t$, donde

$$\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t^T + \boldsymbol{\Sigma}_t.$$

Por lo tanto, el estimador BLUP $\tilde{\mathbf{v}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{y}$ se reduce a

$$\tilde{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{Z}_t^T\mathbf{F}_t^{-1}(\mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}). \tag{2.170}$$

Además, se sigue de (2.120) que la matriz de covarianzas de los errores de predicción $\tilde{\mathbf{v}} - \mathbf{v}$ son $\mathbf{G} - \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}$, la cual en el caso de los errores de predicción $\tilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t$ se reduce a

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{Z}_t^T\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}. \tag{2.171}$$

El modelo general de espacio de estado, (2.167) y (2.168), cubren ambos modelos, el modelo de nivel de área y el modelo de nivel unidad.

El estimador BLUP $\tilde{\boldsymbol{\alpha}}_t$ incluye parámetros desconocidos $\boldsymbol{\delta}$ que especifican las matrices de covarianzas $\boldsymbol{\Sigma}_t$ y $\boldsymbol{\Gamma}$. El vector $\boldsymbol{\delta}$ puede contener también elementos desconocidos de la matriz de transición. Asumiendo normalidad de los errores $\boldsymbol{\varepsilon}_t$ y $\boldsymbol{\eta}_t$, pueden obtenerse estimadores MV de estos parámetros expresando la *log* verosimilitud en la forma

$$\log L = \text{const.} - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1})^T\mathbf{F}_t^{-1}(\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}), \tag{2.172}$$

donde $\tilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}_t\tilde{\alpha}_{t|t-1}$ es el BLUP de $\mathbf{y}_t$ al tiempo $t-1$ y $\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}$ es el vector de errores de predicción. La representación (2.172) se sigue escribiendo L como

$$L = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{Y}_{t-1}),$$

donde $p(\mathbf{y}_t|\mathbf{Y}_{t-1})$ es la distribución condicional de $\mathbf{y}_t$ dado $\mathbf{Y}_{t-1} = \{\mathbf{y}_{t-1}, \cdots, \mathbf{y}_1\}$ y notando que $p(\mathbf{y}_t|\mathbf{Y}_{t-1})$ es normal con media $\tilde{\mathbf{y}}_{t|t-1}$ y matriz de covarianzas $\mathbf{F}_t$. El estimador MV de $\boldsymbol{\delta}$ y los parámetros de la especificación inicial pueden ser obtenidos usando el método scoring con una longitud de paso variable (Pfeffermann y Burck, 1990). Resolvemos el siguiente conjunto de ecuaciones iterativamente:

$$\boldsymbol{\delta}^{(a+1)} = \boldsymbol{\delta}^{(a)} + r_d[\mathcal{I}(\boldsymbol{\delta}^{(a)})]^{-1}\mathbf{s}(\boldsymbol{\delta}^{(a)}).$$

Aquí, $\boldsymbol{\delta}^{(a)}$ es el valor del estimador en la a-ésima iteración, $\mathbf{s}(\boldsymbol{\delta})$ es el vector score consistente de los elementos $\partial L/\partial\delta_d, \mathcal{I}(\boldsymbol{\delta})$ es la matriz de información y r_d es una variable de longitud de paso introducida para garantizar que la verosimilitud sea no decreciente, es decir, $L(\boldsymbol{\delta}^{(a+1)}) \geq L(\boldsymbol{\delta}^{(a)})$ en cada iteración a. El método de momentos también puede usarse para estimar los parámetros del modelo $\boldsymbol{\delta}$ sin el supuesto de normalidad. Singh, Mantel y Thomas (1994), presentaron tales estimadores para casos especiales, usando un método de momentos análogo al método de Fay y Herriot (1979), para datos transversales.

El estimador EBLUP $\widehat{\boldsymbol{\alpha}}_t$ se obtiene sustituyendo un estimador $\widehat{\boldsymbol{\delta}}$ por $\boldsymbol{\delta}$ en $\tilde{\boldsymbol{\alpha}}_t$. El estimador EBLUP resultante de $\boldsymbol{\theta}_t = \mathbf{Z}_t\boldsymbol{\alpha}_t$ es

$$\widehat{\boldsymbol{\theta}}_t^H = \mathbf{Z}_t\widehat{\boldsymbol{\alpha}}_t. \tag{2.173}$$

Pfeffermann y Tiller (2001), establecieron la validez del estimador bootstrap del ECM para términos de segundo orden, bajo ciertas condiciones de regularidad.

Modelos espaciales: Los modelos espaciales discutidos antes son similares al modelo de Fay-Herriot excepto por la correlación espacial entre los efectos aleatorios de área pequeña v_d. Por lo tanto, $\mathbf{G} = \sigma_v^2\mathbf{I}_D$ se cambia por $\mathbf{G} = \boldsymbol{\Gamma}(\boldsymbol{\delta})$, donde la matriz de covarianzas $\boldsymbol{\Gamma}(\boldsymbol{\delta})$ puede tomar la forma $\boldsymbol{\Gamma}(\boldsymbol{\delta}) = \sigma_v^2(\mathbf{I} - \rho\mathbf{Q})^{-1}\mathbf{B}$ como se ha especificado en (2.91) o $\boldsymbol{\Gamma}(\boldsymbol{\delta}) = \sigma_v^2(\delta_1\mathbf{I} + \delta_2\mathbf{D})$ o $\boldsymbol{\Gamma}(\boldsymbol{\delta}) = \sigma_v^2[\delta_1\mathbf{I} - \delta_2\mathbf{D}(\delta_3)]$ como se ha usado en la literatura geoestadística, donde $\mathbf{D}$ y $\mathbf{D}(\delta_3)$ están definidas en la subsección 2.2.2.

Los modelos espaciales son casos especiales del modelo mixto lineal general (2.112). Por lo tanto, el estimador BLUP de $\theta_d = \mathbf{z}_d^T\boldsymbol{\beta} + b_dv_d$ puede obtenerse de la fórmula general (2.115), para un $\boldsymbol{\delta}$ especificado. En la práctica, $\boldsymbol{\delta}$ es desconocido y necesitamos reemplazarlo por un estimador $\widehat{\boldsymbol{\delta}}$ para obtener el

estimador EBLUP de θ_d. Los estimadores MV y MVR de $\boldsymbol{\delta}$ pueden obtenerse del resultado general obtenido antes.

Modelo de regresión multivariante de error anidado: La parte de la muestra del modelo de regresión multivariante con error anidado (2.99) puede expresarse como un caso especial del modelo mixto general (2.129) con estructura de covarianzas diagonal por bloques. Tenemos

$$\mathbf{y}_d^T = (\mathbf{y}_{d1}^T, \cdots, \mathbf{y}_{dn_d}^T), \quad \mathbf{e}_d^T = (\mathbf{e}_{d1}^T, \cdots, \mathbf{e}_{dn_d}^T)$$

$$\mathbf{v}_d = \mathbf{v}_d, \quad \boldsymbol{\beta} = \text{vec}(\mathbf{B}),$$
$$\mathbf{X}_d^T = [(\mathbf{I}_r \otimes \mathbf{x}_{d1}^T)^T, \cdots, (\mathbf{I}_r \otimes \mathbf{x}_{dn_d}^T)^T]$$
$$\mathbf{Z}_d = \mathbf{I}_{n_d} \otimes \mathbf{I}_r, \quad d = 1, \cdots, D,$$

donde el operador $\otimes$ denota el producto directo, $\mathbf{I}_{n_d}$ es el vector de n_d unos, vec($\mathbf{B}$) es el vector de $pr \times 1$ obtenido colocando las columnas de la matriz $\mathbf{B}$ de $r \times p$ uno abajo la otra empezando con la primera columna, y n_d es el número de observaciones multivariantes y_{dj} de la muestra. Además,

$$\mathbf{G}_d = \boldsymbol{\Sigma}_v, \quad \mathbf{R}_d = \mathbf{I}_{n_d} \otimes \boldsymbol{\Sigma}_e$$

para que

$$\mathbf{V}_d = (\mathbf{J}_{n_d} \otimes \boldsymbol{\Sigma}_v) + (\mathbf{I}_{n_d} \otimes \boldsymbol{\Sigma}_e),$$

donde $\mathbf{J}_{n_d} = \mathbf{1}_{n_d}\mathbf{1}_{n_d}^T$ y $\mathbf{I}_{n_d}$ es la matriz identidad de orden n_d. Los parámetros de interés son los vectores de medias de áreas pequeñas $\overline{\mathbf{Y}}_d \approx \mathbf{B}\overline{\mathbf{X}}_d + \mathbf{v}_d = (\mathbf{I}_r \otimes \overline{\mathbf{X}}_d^T)\boldsymbol{\beta} + \mathbf{v}_d$ si el tamaño de la población N_d es grande. Por lo tanto, $\overline{\mathbf{Y}}_d$ es de la forma $\boldsymbol{\mu}_d = \mathbf{L}_d\boldsymbol{\beta} + \mathbf{M}_d\mathbf{v}_d$ con $\mathbf{L}_d = \mathbf{I}_r \otimes \overline{\mathbf{X}}_d^T$ y $\mathbf{M}_i = \mathbf{I}_r$.

Haciendo la anterior sustitución en la fórmula general (2.133) obtenemos el estimador BLUP $\tilde{\boldsymbol{\mu}}_d^H$ de $\boldsymbol{\mu}_d = \mathbf{B}^T\overline{\mathbf{X}}_d + \mathbf{v}_d$. El estimador EBLUP $\widehat{\boldsymbol{\mu}}_d^H$ se obtiene de la sustitución apropiada de los estimadores $\widehat{\boldsymbol{\Sigma}}_v$ y $\widehat{\boldsymbol{\Sigma}}_e$ para $\boldsymbol{\Sigma}_v$ y $\boldsymbol{\Sigma}_e$.

Modelo lineal de varianzas con error aleatorio: Introdujimos antes un modelo de efectos aleatorios simple con varianza de errores aleatorios, σ_{ed}^2. Asumiendo igualdad de tamaños de muestra de área pequeña, $n_d = \overline{n}$, la parte de la muestra de este modelo puede escribirse como

$$y_{dj} = \beta + v_d + e_{dj}, \quad j = 1, \cdots, \overline{n}; \quad d = 1, \cdots, D$$

$$v_d \overset{\text{iid}}{\sim}(0, \sigma_v^2), \quad e_{dj}|\sigma_{ed}^2 \overset{\text{iid}}{\sim}(0, \sigma_{ed}^2) \quad \text{para cada} \quad d, \tag{2.174}$$
$$\sigma_{ed}^2 \overset{\text{iid}}{\sim}(\sigma_e^2, \delta_e).$$

No se imponen supuestos a las distribuciones de $v_d, e_{dj}|\sigma^2_{ed}$ y σ^2_{ed}.

El estimador EBLUP de la media de área pequeña $\mu_d = \mu + v_d$ bajo (2.174) es idéntico al estimador EBLUP obtenido bajo el supuesto de igualdad de varianza del error $\sigma^2_{ed} = \sigma^2_e$ (es decir, $\delta_e = 0$). Está dado por

$$\widehat{\mu}^H_d = \overline{y}_d + \widehat{\gamma}(\overline{y}_d - \overline{y}), \tag{2.175}$$

donde $\overline{y}_d$ y $\overline{y}$ es la media de la muestra del área d y la media muestral total, y $\widehat{\gamma} = \widehat{\sigma}^2_v/(\widehat{\sigma}^2_v + \widehat{\sigma}^2_e/\overline{n})$ con $D(\overline{n}-1)\widehat{\sigma}^2_e = \sum_d \sum_j (y_{dj} - \overline{y}_d)^2$ y $(D-1)(\widehat{\sigma}^2_v + \widehat{\sigma}^2_e/\overline{n}) = \sum_d (\overline{y}_d - \overline{y})^2$.

Modelo de regresión de error anidado doble: Hemos descrito un modelo de regresión de error anidado doble apropiado para muestreo en dos etapas dentro de áreas: se selecciona una muestra, s_d, de m_d unidades primarias (conglomerados) de el área d y si el j-ésimo conglomerado es seleccionado entonces se selecciona una submuestra, s_{dj}, de n_{dj} elementos del j-ésimo conglomerado y se observan los valores asociados $(y_{djl}, \mathbf{x}_{djl}); l = 1, \cdots, n_{dj}$. Asumimos que la muestra obedece el modelo bivariado dado por (2.100) así que

$$y_{djl} = \mathbf{x}^T_{djl}\boldsymbol{\beta} + v_d + u_{dj} + e_{djl}; l = 1, \cdots, n_{dj}; j = 1, \cdots, m_d; d = 1, \cdots, D, \tag{2.176}$$

donde $v_d \overset{\text{iid}}{\sim}(0, \sigma^2_v), u_{dj} \overset{\text{iid}}{\sim}(0, \sigma^2_u)$ y $e_{djl} = k_{djl}\tilde{e}_{djl}$ con $\tilde{e}_{djl} \overset{\text{iid}}{\sim}(0, \sigma^2_e)$, y constantes conocidas k_{djl}. Además, $\{v_d\}$, $\{u_{dj}\}$ y $\{\tilde{e}_{djl}\}$ son mutuamente independientes.

El modelo bivariado (2.176) es un caso especial del modelo mixto lineal general (2.129) con estructura de covarianzas diagonal por bloques. Fijando $\text{col}_{1\le d\le t}(\mathbf{a}_d) = (\mathbf{a}^T_1, \cdots, \mathbf{a}^T_t)^T$ y $\text{diag}_{1\le d\le t}(\mathbf{A}_d) = \text{blockdiag}(\mathbf{A}_1, \cdots, \mathbf{A}_t)$, tenemos

$$\mathbf{y}_d = \text{col}_{1\le j\le m_d}(\mathbf{y}_{dj}), \quad \mathbf{y}_{dj} = \text{col}_{1\le l\le n_{dj}}(\mathbf{y}_{djl}),$$

$$\mathbf{X}_d = \text{col}_{1\le j\le m_d}(\text{col}_{1\le l\le n_{dj}}(\mathbf{x}^T_{djl})), \quad \mathbf{Z}_d = (\mathbf{z}_d|\mathbf{Z}_{2d}),$$

$$\mathbf{v}_d = \begin{pmatrix} v_d \\ \mathbf{v}_{2d} \end{pmatrix}, \quad \boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T,$$

y $\mathbf{e}_d$ se obtiene de $\mathbf{y}_d$ cambiando y_{djl} por e_{djl}, donde

$$\mathbf{z}_d = \text{col}_{1\le j\le m_d}(\text{col}_{1\le k\le n_{dj}}(1)) = \text{col}_{1\le j\le m_d}(\mathbf{z}_{dj}),$$

$$\mathbf{Z}_{2d} = \text{diag}_{1\le j\le m_d}(\mathbf{z}_{dj}), \quad \mathbf{v}_{2d} = \text{col}_{1\le j\le m_d}(u_{dj}).$$

Además,

$$\mathbf{G}_d = \begin{bmatrix} \sigma^2_v & \mathbf{0}^T \\ \mathbf{0} & \sigma^2_u\mathbf{I}_{m_d} \end{bmatrix}, \quad \mathbf{R}_d = \text{diag}_{1\le j\le m_d}(\mathbf{R}_{dj})$$

y $\mathbf{R}_{dj} = \text{diag}_{1\leq l\leq n_{dj}}(\sigma_e^2 k_{djl}^2)$, donde $\mathbf{I}_b$ es la matriz identidad de orden b y $\mathbf{0}$ es un vector de ceros. Usando una forma explícita para $\mathbf{V}_d^{-1}$, la fórmula general (2.132) para $\tilde{\boldsymbol{\beta}}$ se reduce a

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{b}, \tag{2.177}$$

$$\sigma_e^2\mathbf{A} = \sum_d\sum_j\sum_l a_{djl}\mathbf{x}_{djl}\mathbf{x}_{djl}^T - \sum_d\sum_j a_{dj\cdot}\gamma_{dj}\overline{\mathbf{x}}_{dja}\overline{\mathbf{x}}_{dja}^T$$

$$-(\sigma_e^2/\sigma_u^2)\sum_d \gamma_d\left(\sum_j \gamma_{dj}\right)\overline{\mathbf{x}}_{d\gamma}\overline{\mathbf{x}}_{d\gamma}^T,$$

y

$$\sigma_e^2\mathbf{b} = \sum_d\sum_j\sum_l a_{djl}\mathbf{x}_{djl}y_{djl} - \sum_d\sum_j a_{dj\cdot}\gamma_{dj}\overline{\mathbf{x}}_{dja}\overline{y}_{dja}$$

$$-(\sigma_e^2/\sigma_u^2)\sum_d \gamma_d\left(\sum_j \gamma_{dj}\right)\overline{\mathbf{x}}_{d\gamma}\overline{y}_{d\gamma},$$

donde $a_{djl} = k_{djl}^{-2}, a_{dj\cdot} = \sum_l a_{djl}, \overline{y}_{d\gamma} = \sum_j \gamma_{dj}\overline{y}_{dja}/(\sum_j \gamma_{dj}), \overline{\mathbf{x}}_{d\gamma} = \sum_j \gamma_{dj}\overline{\mathbf{x}}_{dja}/(\sum_j \gamma_{dj})$ con $\overline{y}_{dja} = \sum_l a_{djl}y_{djl}/a_{dj\cdot}, \overline{\mathbf{x}}_{dja} = \sum_l a_{djl}\mathbf{x}_{djl}/a_{dj\cdot}$, y

$$\gamma_{dj} = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/a_{dj\cdot}), \quad \gamma_d = \sigma_v^2/[\sigma_v^2 + \sigma_u^2/(\sum_j \gamma_{dj})] \tag{2.178}$$

ver Stukel y Rao (1999). Además, la fórmula general (2.131) para el estimador BLUP de $\mathbf{v}_d$ se reduce a $\tilde{\mathbf{v}} = (\tilde{v}_d, \mathbf{v}_{2d}^T)$ con elementos

$$\tilde{v}_d = \gamma_d(\overline{y}_{d\gamma} - \overline{\mathbf{x}}_{d\gamma}^T\tilde{\boldsymbol{\beta}}) \tag{2.179}$$

y

$$\tilde{u}_{dj} = \gamma_{dj}(\overline{y}_{dja} - \overline{\mathbf{x}}_{dja}^T\tilde{\boldsymbol{\beta}}) - \gamma_d\gamma_{dj}(\overline{y}_{d\gamma} - \overline{\mathbf{x}}_{d\gamma}^T\tilde{\boldsymbol{\beta}}). \tag{2.180}$$

El estimador BLUP de la media de área pequeña $\overline{Y}_d$ está dada por

$$\tilde{\overline{Y}}_d^H = \frac{1}{N_d}\left[\sum_{j\in s_d}\sum_{k\in s_{dj}} y_{djk} + \sum_{j\in s_d}\left(\sum_{l\in\overline{s}_{dj}} y_{djl}^*\right) + \sum_{j\in\overline{s}_d}\left(\sum_{l=1}^{N_{dj}} y_{djl}^{**}\right)\right], \tag{2.181}$$

donde $\overline{s}_{dj}$ es el conjunto de elementos no muestreados en el j-ésimo conglomerado muestreado, $\overline{s}_d$ es el conjunto de conglomerados no muestreados, N_{dj} es el número de elementos en el j-ésimo conglomerado de el área d ($j = 1, \cdots, M_d$) y $N_d = \sum_j N_{dj}$ es el número total de elementos en el área d. Además,

$$y_{djl}^* = \mathbf{x}_{djl}^T\tilde{\boldsymbol{\beta}} + \tilde{v}_d + \tilde{u}_{dj} \tag{2.182}$$

y

$$y^{**}_{djl} = \mathbf{x}^T_{djl}\tilde{\boldsymbol{\beta}} + \tilde{v}_d. \tag{2.183}$$

La propiedad BLUP de $\overset{\sim}{\overline{Y}}{}^H_d$ se establece mostrando que $\mathrm{Cov}(\mathbf{a}^T\mathbf{y}, \overset{\sim}{\overline{Y}}{}^H_d - \overline{Y}_d) = 0$ para toda función cero $\mathbf{a}^T\mathbf{y} = \sum_d \mathbf{a}^T_d\mathbf{y}_d$, es decir, $E(\mathbf{a}^T\mathbf{y}) = 0$; ver Stukel (1991, capítulo 3). La fórmula (2.181) para $\overset{\sim}{\overline{Y}}{}^H_d$ muestra que $\overset{\sim}{\overline{Y}}{}^H_d$ se obtiene sin dificultad de $\tilde{\boldsymbol{\beta}}, \tilde{v}_d$ y $\tilde{u}_{dj}$ dados por (2.177), (2.179) y (2.180), respectivamente.

Si el número de unidades primarias, N_d, es grande, entonces $\overline{Y}_d$ puede aproximarse como $\overline{Y}_d \approx \boldsymbol{\mu}_d = \overline{\mathbf{X}}^T_d\boldsymbol{\beta} + v_d$, donde $\overline{\mathbf{X}}$ es el vector de medias de x conocidas de la población. En este caso, el estimador BLUP de μ_d está dado por

$$\tilde{\boldsymbol{\mu}}^H_d = \gamma_d[\overline{y}_{d\gamma} + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_{d\gamma})^T\tilde{\boldsymbol{\beta}}] + (1-\gamma_d)\overline{\mathbf{X}}^T_d\tilde{\boldsymbol{\beta}}. \tag{2.184}$$

La fórmula (2.184) muestra que el estimador BLUP de μ_d es un estimador combinado del estimador de regresión $\overline{y}_{d\gamma} + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_{d\gamma})^T\tilde{\boldsymbol{\beta}}$ y el estimador de regresión sintético $\overline{\mathbf{X}}^T_d\tilde{\boldsymbol{\beta}}$. Nótese que $\tilde{\boldsymbol{\mu}}^H_d$ depende de los valores x poblacionales solamente a través de la media $\overline{\mathbf{X}}_d$.

Modelo de dos niveles: Hemos introducido un modelo de dos niveles que integra efectivamente el uso de unidades de nivel y covariables de nivel de área en un solo modelo (2.106). Asumimos que la muestra obedece al modelo de dos niveles para que

$$\mathbf{y}_d = \tilde{\mathbf{X}}_d\tilde{\mathbf{Z}}_d\boldsymbol{\alpha} + \tilde{\mathbf{X}}_d\mathbf{v}_d + \mathbf{e}_d; \quad d = 1, \cdots, D \tag{2.185}$$

donde $(\mathbf{y}_d, \tilde{\mathbf{X}}_d, \mathbf{e}_d)$ corresponde a la parte de la muestra del modelo de la población (2.106); usamos $\tilde{\mathbf{X}}_d$ aquí para denotar la parte de la muestra de $\mathbf{X}^P_d$ para evitar confundir con $\mathbf{X}_d$ del modelo mixto lineal general. El modelo (2.185) es un caso especial del modelo mixto lineal general (2.129) con estructura de covarianzas diagonal por bloques. Tenemos

$$\mathbf{y}_d = \mathbf{y}_d = \mathrm{col}_{1\le j\le n_d}(y_{dj}), \quad \mathbf{X}_d = \tilde{\mathbf{X}}_d\tilde{\mathbf{Z}}_d, \quad \mathbf{Z}_d = \tilde{\mathbf{X}}_d,$$

$$\boldsymbol{\beta} = \boldsymbol{\alpha}, \quad \mathbf{v}_d = \mathbf{v}_d, \quad \mathbf{e}_d = \mathbf{e}_d,$$

$$\mathbf{G}_d = \boldsymbol{\Sigma}_v, \quad \mathbf{R}_d = \sigma^2_e\mathrm{diag}_{1\le j\le n_d}(k^2_{dj}),$$

donde n_d es el tamaño de muestra en el área d $(d = 1, \cdots, D)$ y las k_{dj} son constantes conocidas.

Si N_d es grande, podemos expresar la media $\overline{Y}_d$ como $\overline{\mathbf{X}}^T_d\mathbf{Z}_d\boldsymbol{\alpha} + \overline{\mathbf{X}}^T_d\mathbf{v}_d$, donde $\overline{\mathbf{X}}_d$ es el vector de medias de x conocidas de la población. Si ahora se sigue que el estimador BLUP, $\tilde{\mu}^H_d$, y su ECM son fácilmente obtenidos de (2.130) y (2.132) usando $\mathbf{1}^T_d = \overline{\mathbf{X}}^T_d\mathbf{Z}_d$ y $\mathbf{m}^T_d = \overline{\mathbf{X}}^T_d$.

El estimador BLUP, $\tilde{\mu}_d^H$, depende de los $p(p+1)/2$ distintos elementos de la matriz $\mathbf{\Sigma}_v$ de $p \times p$ y la varianza del error σ_e^2. Moura y Holt (1999), usaron mínimos cuadrados generalizados iterados restringidos (RIGLS) para estimar los anteriores parámetros, $\boldsymbol{\delta}$. Goldstein (1989), propuso este método para modelos multi nivel. El estimador EBLUP $\widehat{\mu}_d^H$, se obtiene del estimador BLUP reemplazando $\boldsymbol{\delta}$ por el estimador RIGLS $\widehat{\boldsymbol{\delta}}$.

Cuando las fracciones de muestreo, $f_d = n_d/N_d$, no son pequeñas, el estimador EBLUP de $\overline{Y}_d$ y su estimador del ECM se obtienen mediante la metodología descrita para fracciones de muestreo no despreciables. Para mayores detalles consultar Moura y Holt (1999).

2.5. Estimación bayesiana de áreas pequeñas

El método EBLUP es aplicable a modelos lineales mixtos que cubren muchas aplicaciones de estimación de áreas pequeñas. La normalidad de los efectos aleatorios y de los errores no es necesaria para la estimación puntual, pero si es necesario el supuesto de normalidad para obtener el estimador del ECM exacto.

Los modelos lineales mixtos se diseñan para variables continuas, y, pero no son apropiados para manejar datos binarios o de frecuencias. Los métodos de bayes empírico (BE) y bayes jerárquico (BJ) son aplicables más generalmente en el sentido de ocuparse de modelos para datos binarios y de frecuencias así como los modelos lineales mixtos normales. En el último caso, los estimadores BE y EBLUP son idénticos.

2.5.1. Método empírico de bayes

En esta subsección, describiremos el método empírico de bayes (BE) en el contexto de la estimación de áreas pequeñas. La aproximación de BE puede resumirse como sigue:

(i) Obtener la densidad a posteriori, $f(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\lambda})$, de los parámetros de interés (aleatorios) del área pequeña, $\boldsymbol{\mu}$, dados los datos $\mathbf{y}$, usando la densidad condicional, $f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\lambda})$, de $\mathbf{y}$ dado $\boldsymbol{\mu}$ y la densidad $f(\boldsymbol{\mu}|\boldsymbol{\lambda}_2)$ de $\boldsymbol{\mu}$, donde $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T)^T$ denota el vector de parámetros del modelo.

(ii) Estimar los parámetros del modelo, $\boldsymbol{\lambda}$, de la densidad marginal, $f(\mathbf{y}|\boldsymbol{\lambda})$, de $\mathbf{y}$.

(iii) Usar la densidad a posteriori estimada, $f(\boldsymbol{\mu}|\mathbf{y}, \widehat{\boldsymbol{\lambda}})$, para hacer inferencias acerca de $\boldsymbol{\mu}$, donde $\widehat{\boldsymbol{\lambda}}$ es un estimador de $\boldsymbol{\lambda}$.

La densidad de $\boldsymbol{\mu}$ se interpreta a menudo como densidad a priori sobre $\boldsymbol{\mu}$, pero realmente es una parte del modelo postulado sobre $(\mathbf{y}, \boldsymbol{\mu})$ y puede validarse de los datos, a diferencia de a prioris subjetivas sobre parámetros del modelo, $\boldsymbol{\lambda}$, usados en la aproximación BJ. En este sentido la aproximación BE es esencialmente frecuentista, y las inferencias de BE hacen referencia a promediar sobre la distribución conjunta de $\mathbf{y}$ y $\boldsymbol{\mu}$. A veces se escoge una densidad a priori para $\boldsymbol{\lambda}$, pero solamente para obtener estimadores y medidas de variabilidad asociadas, así como intervalos de confianza con buenas propiedades frecuentistas (Morris, 1983b).

En la aproximación de bayes empírica paramétrica (BEP), se asume una forma paramétrica, $f(\boldsymbol{\mu}, \boldsymbol{\lambda}_2)$, para la densidad de $\boldsymbol{\mu}$. Por otro lado, los métodos de bayes empíricos no paramétricos (BENP) no especifican la forma de

la distribución (a priori) de $\boldsymbol{\mu}$. Se usa máxima verosimilitud no paramétrica para estimar la distribución (a priori) de $\boldsymbol{\mu}$ (Laird, 1978). También se usan representaciones semi-no paramétricas (SNP) de la densidad de $\boldsymbol{\mu}$ para hacer inferencias de BE.

Modelo básico de nivel de área

Asumiendo normalidad, el modelo básico de nivel de área (2.134) puede expresarse como un modelo jerárquico de dos etapas:

(i) $\widehat{\theta}_d|\theta_d \overset{\text{ind}}{\sim} N(\theta_d, \psi_d), \quad d = 1, \cdots, D;$

(ii) $\theta_d \overset{\text{ind}}{\sim} N(\mathbf{z}_d^T\boldsymbol{\beta}, b_d^2\sigma_d^2), \quad d = 1, \cdots, D$, donde $\boldsymbol{\beta}$ es el vector de $p \times 1$ parámetros de regresión.

En el enfoque bayesiano, los parámetros del modelo $\boldsymbol{\beta}$ y σ_v^2 son aleatorios, y el modelo jerárquico de dos etapas es llamado "modelo jerárquico condicionalmente independiente" (MJCI) porque los pares $(\widehat{\theta}_d, \theta_d)$ son independientes en las d áreas, condicionalmente sobre $\boldsymbol{\beta}$ y σ_v^2 (Kass y Steffey, 1989).

Estimador BE: El estimador "óptimo" del valor realizado de θ_d está dado por la esperanza condicional de θ_d dado $\widehat{\theta}_d$, $\boldsymbol{\beta}$ y σ_v^2:

$$E(\theta_d|\widehat{\theta}_d, \boldsymbol{\beta}, \sigma_v^2) = \widehat{\theta}_d^B = \gamma_d\widehat{\theta}_d + (1-\gamma_d)\mathbf{z}_d^T\boldsymbol{\beta}, \tag{2.186}$$

donde $\gamma_d = b_d^2\sigma_v^2/(b_d^2\sigma_v^2 + \psi_d)$. El resultado (2.186) se sigue de la distribución a posteriori (o condicional) de θ_d dado $\widehat{\theta}_d$, $\boldsymbol{\beta}$ y σ_v^2:

$$\widehat{\theta}_d, \boldsymbol{\beta}, \sigma_v^2 \overset{\text{ind}}{\sim} N(\widehat{\theta}_d^B, g_{1d}(\sigma_v^2) = \gamma_d\psi_d). \tag{2.187}$$

El estimador $\widehat{\theta}_d^B = \widehat{\theta}_d^B(\boldsymbol{\beta}, \sigma_v^2)$ es el estimador de "bayes" bajo pérdida de error cuadrático medio y es óptimo en el sentido que su ECM, $\text{ECM}(\widehat{\theta}_d^B) = (\widehat{\theta}_d^B - \theta_d)^2$ es más pequeño que el ECM de cualquier otro estimador de θ_d, lineal o no lineal en los $\widehat{\theta}_d$. Puede ser más apropiado nombrar a $\widehat{\theta}_d^B$ como el estimador de la mejor predicción (MP) de θ_d porque se obtiene de la distribución condicional (2.187) sin asumir una distribución a priori sobre los parámetros del modelo (Jiang, Lahiri y Wan, 2002).

El estimador de bayes $\widehat{\theta}_d^B$ depende de los parámetros del modelo $\boldsymbol{\beta}$ y σ_v^2, los cuales se estiman de la distribución marginal: $\theta_d \overset{\text{ind}}{\sim} N(\mathbf{z}_d^T\boldsymbol{\beta}, b_d^2\sigma_v^2 + \psi_d)$ usando MV o MVR. Denotando los estimadores como $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}_v^2$ obtenemos el estimador BE o el BP empírico (EBP) de θ_d de $\widehat{\theta}_d^B$ sustituyendo $\widehat{\boldsymbol{\beta}}$ por $\boldsymbol{\beta}$ y $\widehat{\sigma}_v^2$ por σ_v^2:

$$\widehat{\theta}_d^{BE} = \widehat{\theta}_d^B(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_v^2) = \widehat{\gamma}_d\widehat{\theta}_d + (1-\widehat{\gamma}_d)\mathbf{z}_d^T\widehat{\boldsymbol{\beta}}. \tag{2.188}$$

El estimador BE, $\widehat{\theta}_d^{BE}$, es idéntico al estimador EBLUP $\widehat{\theta}_d^H$ dado por (2.139). Nótese que $\widehat{\theta}_d^{BE}$ es también la media de la densidad a posteriori estimada, $f(\theta_d|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}_d^2)$, de θ_d, es decir, $N(\widehat{\theta}_d^{BE}, \widehat{\gamma}_d\psi_d)$.

Morris (1983b), estudió el caso especial de igualdad de varianzas muestrales $\psi_d = \psi$ y $b_d = 1$ para todo d, que resulta ser idéntico al estimador de James-Stein. Para el caso de varianzas desiguales ψ_d y $b_d = 1$, utilizó la constante multiplicativa $(D-p-2)/(D-p)$ en el estimador BE (2.188); es decir, reemplazó $1-\widehat{\gamma}$ por

$$1-\gamma^* = [(D-p-2)/(D-p)]\psi/(\psi+\widehat{\sigma}_v^2), \tag{2.189}$$

donde $\widehat{\sigma}_v^2$ es el estimador MVR de σ_v^2. Propuso también un estimador alternativo de σ_v^2. Es similar al estimador de momentos de Fay-Herriot. Se obtiene resolviendo

$$\sigma_v^2 = \left(\sum_d \alpha_d\right)^{-1}\left[\sum_d \alpha_d\left\{\frac{m}{D-p}(\widehat{\theta}_d - \mathbf{z}_d^T\tilde{\boldsymbol{\beta}})^2 - \psi_d\right\}\right] \tag{2.190}$$

iterativamente para σ_v^2, donde $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_v^2)$ es el estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$ y $\alpha_d = 1/(\sigma_v^2+\psi_d)$. Si la solución, $\tilde{\sigma}_v^2$, es negativa, tomamos $\widehat{\sigma}_v^2 = \text{máx}(\tilde{\sigma}_v^2, 0)$. Si α_d se reemplaza por α_d^2 en (2.190), entonces la solución resultante es aproximadamente igual al estimador de MVR de σ_v^2.

Una ventaja del estimador BE (o EBP) es que puede aplicarse para hallar el estimador BE de cualquier función $\phi_d = h(\theta_d)$; en particular, $\overline{Y}_d = g^{-1}(\theta_d) = h(\theta_d)$. El estimador BE se obtiene del estimador de bayes $\widehat{\phi}_d^B = E(\phi_d|\widehat{\theta}_d, \boldsymbol{\beta}, \sigma_v^2)$ sustituyendo $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_v^2)$ por $(\boldsymbol{\beta}, \sigma_v^2)$. El cálculo del estimador BE $\widehat{\phi}_d^{BE}$ podría requerir el uso de integración de Monte Carlo o numérica. Por ejemplo, $\{\theta_d^{(r)}, r = 1, \cdots, R\}$ puede simularse de la densidad posterior estimada, es decir, $N(\widehat{\theta}_d^{BE}, \widehat{\gamma}_d\psi_d)$, para obtener una aproximación de Monte Carlo:

$$\widehat{\phi}_d^{BE} \approx \frac{1}{R}\sum_{r=1}^{R} h(\theta_d^{(r)}). \tag{2.191}$$

El cálculo de (2.191) puede simplificarse reescribiendo (2.191) como

$$\widehat{\phi}_d^{BE} \approx \frac{1}{R}\sum_{r=1}^{R} h\left(\widehat{\theta}_d^{BE} + z_d^{(r)}\sqrt{\widehat{\gamma}_d\psi_d}\right), \tag{2.192}$$

donde $\{z_d^{(r)}, r = 1, \cdots, R\}$ son generados de $N(0,1)$.

La aproximación (2.192) será buena si el número de muestras simuladas, R, es grande.

Jiang, Lahiri y Wan (2002), propusieron un método Jackknife de estimación del ECM de los estimadores BE. Este método es más general que el método bootstrap paramétrico para $\widehat{\theta}_d^H = \widehat{\theta}_d^{BE}$, pero ilustraremos su uso aquí para estimar $\text{MSE}(\widehat{\theta}_d^{BE})$.

Los pasos jackknife para estimar los dos términos, M_{2d} y M_{1d}, son como sigue. Sea $\widehat{\theta}_d^{BE} = k_d(\widehat{\theta}_d, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}_v^2)$ el estimador BE de θ_d expresado como una función del estimador directo $\widehat{\theta}_d$ y los estimadores de los parámetros $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}_v^2$.

Paso 1. Calcular los estimadores $\widehat{\boldsymbol{\beta}}_{-l}$ y $\widehat{\sigma}_{v,-l}^2$ eliminando el l-ésimo conjunto de datos de área $(\widehat{\theta}_l, \mathbf{z}_l)$ del conjunto total de datos $\{(\widehat{\theta}_d, \mathbf{z}_d); d = 1, \cdots, D\}$. Este cálculo se hace para cada l para obtener D estimadores de $\boldsymbol{\beta}$ y σ_v^2: $\{(\widehat{\boldsymbol{\beta}}_{-l}, \widehat{\sigma}_{v,-l}^2); l = 1, \cdots, D\}$, los cuales, a su vez, proporcionan D estimadores de θ_d: $\{\widehat{\theta}_{d,-l}^{BE}; l = 1, \cdots, D\}$, donde $\widehat{\theta}_{d,-l}^{BE} = k_d(\widehat{\theta}_d, \widehat{\boldsymbol{\beta}}_{-l}, \widehat{\sigma}_{v,-l}^2)$.

Paso 2. Calcular el estimador de M_{2d} como

$$\widehat{M}_{2d} = \frac{D-1}{D} \sum_{l=1}^{D} (\widehat{\theta}_{d,-l}^{BE} - \widehat{\theta}_d^{BE})^2. \tag{2.193}$$

Paso 3. Calcular el estimador de M_{1d} como

$$\widehat{M}_{1d} = g_{1d}(\widehat{\sigma}_v^2) - \frac{D-1}{D} \sum_{l=1}^{D} [g_{1d}(\widehat{\sigma}_{v,-l}^2) - g_{1d}(\widehat{\sigma}_v^2)]. \tag{2.194}$$

El estimador $\widehat{M}_{1d}$ corrige el sesgo de $g_{1d}(\widehat{\sigma}_v^2)$.

Paso 4. Calcular el estimador jackknife de $\text{ECM}(\widehat{\theta}_d^{BE})$ como

$$\text{ecm}_J(\widehat{\theta}_d^{BE}) = \widehat{M}_{1d} + \widehat{M}_{2d}. \tag{2.195}$$

El método jackknife es aplicable al estimador BE de cualquier función $\phi_d = h(\theta_d)$, en particular, $\overline{Y}_d = g^{-1}(\theta_d) = h(\theta_d)$. Sin embargo, el cálculo de $\text{ECM}_J(\widehat{\phi}_d^{BE}) = \widehat{M}_{1d} + \widehat{M}_{2d}$ podría requerir integración de Monte Carlo o numérica repetida para obtener $\widehat{M}_{2d} = (D-1)D^{-1} \sum_l (\widehat{\phi}_{d,-l}^{BE} - \widehat{\phi}_d^{BE})^2$ y el estimador del sesgo corregido, $\widehat{M}_{1d}$, de $E(\widehat{\phi}_d^B - \phi_d)^2$.

Modelos lineales mixtos

Estimación BE: Asumiendo normalidad, el modelo lineal mixto (2.129) con estructura de covarianzas diagonal por bloques puede expresarse como

$$\mathbf{y}_d | \mathbf{v}_d \overset{\text{ind}}{\sim} N(\mathbf{X}_d \boldsymbol{\beta} + \mathbf{Z}_d \mathbf{v}_d, \mathbf{R}_d) \tag{2.196}$$

$$\mathbf{v}_d \overset{\text{ind}}{\sim} N(\mathbf{0}, \mathbf{G}_d); \quad d = 1, \cdots, D,$$

donde $\mathbf{G}_d$ y $\mathbf{R}_d$ dependen de la varianza de los parámetros $\boldsymbol{\delta}$. El estimador de bayes o mejor predictor (BP) de $\mu_d = \mathbf{1}_d^T\boldsymbol{\beta} + \mathbf{m}_d^T\mathbf{v}_d$ está dado por la esperanza condicional de μ_d dado $\mathbf{y}_d$, $\boldsymbol{\beta}$ y $\boldsymbol{\delta}$:

$$\widehat{\mu}_d^B = \widehat{\mu}_d^B(\boldsymbol{\beta}, \boldsymbol{\delta}) = E(\mu_d|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\delta}) = \mathbf{1}_d^T\boldsymbol{\beta} + \mathbf{m}_d^T\widehat{\mathbf{v}}_d^B, \tag{2.197}$$

donde

$$\widehat{\mathbf{v}}_d^B = E(\mathbf{v}_d|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\delta}) = \mathbf{G}_d\mathbf{Z}_d^T\mathbf{V}_d^{-1}(\mathbf{y}_d - \mathbf{X}_d\boldsymbol{\beta}) \tag{2.198}$$

y $\mathbf{V}_d = \mathbf{R}_d + \mathbf{Z}_d\mathbf{G}_d\mathbf{Z}_d^T$. El resultado (2.197) se sigue de la distribución a posteriori de μ_d dado $\mathbf{y}_d$:

$$\mu_d|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\delta} \overset{\text{ind}}{\sim} N(\widehat{\mu}_d^B, g_{1d}(\boldsymbol{\delta})). \tag{2.199}$$

El estimador $\widehat{\mu}_d^B$ depende de los parámetros del modelo $\boldsymbol{\beta}$ y $\boldsymbol{\delta}$ los cuales son estimados de la distribución marginal

$$\mathbf{y}_d \overset{\text{ind}}{\sim} N(\mathbf{X}_d\boldsymbol{\beta}, \mathbf{V}_d), \quad d = 1, \cdots, D \tag{2.200}$$

usando MV o MVR. Denotando los estimadores como $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\delta}}$, obtenemos el estimador BE o BP empírico de μ_d de $\widehat{\mu}_d^B$ sustituyendo $\widehat{\boldsymbol{\beta}}$ por $\boldsymbol{\beta}$ y $\widehat{\boldsymbol{\delta}}$ por $\boldsymbol{\delta}$:

$$\widehat{\mu}_d^{BE} = \widehat{\mu}_d^{BE}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}}) = \mathbf{1}_d^T\widehat{\boldsymbol{\beta}} + \mathbf{m}_d^T\widehat{\mathbf{v}}_d^B(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}}). \tag{2.201}$$

El estimador $\widehat{\mu}_d^{BE}$ es idéntico al estimador EBLUP. Observe que $\widehat{\mu}_d^{BE}$ es también la media de la densidad a posteriori estimada, $f(\mu_d|\mathbf{y}_d, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}})$, de μ_d, es decir, $N(\widehat{\mu}_d^{BE}, g_{1d}(\widehat{\boldsymbol{\delta}}))$.

Estimador de bayes lineal empírico

Los métodos BE están basados en supuestos distribucionales sobre $\widehat{\theta}_d|\theta_d$ y θ_d. Los métodos de bayes lineales empíricos (BLE) evitan supuestos distribucionales especificando solamente el primero y segundo momentos pero restringiendo a la clase de estimadores lineales, como en el caso de EBLUP para los modelos lineales mixtos. Maritz y Lwin (1989), proporcionan un excelente resumen de métodos lineales de bayes (BL).

Estimación de BL: Asumimos un modelo de dos fases de la forma $\widehat{\theta}_d|\theta_d \overset{\text{ind}}{\sim}(\theta_d, \psi_d(\theta_d))$ y $\theta_d \overset{\text{ind}}{\sim}(\mu_d, \sigma_d^2), d = 1, \cdots, D$, entonces, tenemos $\widehat{\theta}_d \overset{\text{ind}}{\sim}(\mu_d, \psi_d + \sigma_d^2)$ incondicionalmente, donde $\psi_d = E(\psi_d(\theta_d))$. Este resultado se obtiene notando que $E(\widehat{\theta}_d) = E\{E(\widehat{\theta}_d|\theta_d)\} = E(\theta_d) = \mu_d$ y $V(\widehat{\theta}_d) = E\{V(\widehat{\theta}_d|\theta_d)\} +$

$V\{E(\widehat{\theta}_d|\theta_d)\} = E(\psi_d(\theta_d)) + V(\theta_d) = \psi_d + \sigma_d^2$. Consideremos un estimador de la clase lineal de estimadores de θ_d realizado de la forma $a_d\widehat{\theta}_d + b_d$ y minimicemos el ECM no condicional, $E(a_d\widehat{\theta}_d + b_d - \theta_d)^2$ con respecto a las constantes a_d y b_d. El estimador óptimo, llamado el estimador BL, está dado por

$$\widehat{\theta}_d^{BL} = \mu_d + \gamma_d(\widehat{\theta}_d - \mu_d) = \gamma_d\widehat{\theta}_d + (1-\gamma_d)\mu_d, \tag{2.202}$$

donde $\gamma_d = \sigma_d^2/(\psi_d + \sigma_d^2)$; ver Griffin y Krutchkoff (1971) y Hartigan (1969). El estimador BL (2.202) incluye $2D$ parámetros (μ_d, σ_d^2). En la práctica, necesitamos asumir que μ_d y σ_d^2 dependen de un conjunto fijo de parámetros $\boldsymbol{\lambda}$. El ECM de $\widehat{\theta}_d^{BL}$ está dado por

$$\text{ECM}(\widehat{\theta}_d^{BL}) = E(\widehat{\theta}_d^{BL} - \theta_d) = \gamma_i\psi_d. \tag{2.203}$$

BL restringido

Consideremos el modelo de dos fases (i) $\widehat{\theta}_d|\theta_d \overset{\text{ind}}{\sim}(\theta_d, \psi_d(\theta_d))$ y (ii) $\theta_d \overset{\text{iid}}{\sim}(\mu, \sigma^2)$ y $\psi_d = E(\psi_d(\theta_d))$. Consideremos un estimador de la clase lineal del θ_d realizado de la forma $a_d\widehat{\theta}_d + b_d$ y determinemos las constantes a_d y b_d de forma que:

$$E(a_d\widehat{\theta}_d + b_d) = \mu, \tag{2.204}$$

$$E(a_d\widehat{\theta}_d + b_d - \mu) = \sigma^2. \tag{2.205}$$

Notando que $E(\widehat{\theta}_d) = \mu$, obtenemos $b_d = \mu - (1-a_d)$ de (2.204), y (2.205) se reduce a

$$a_d^2E(\widehat{\theta}_d - \mu)^2 = a_d^2(\sigma^2 + \psi_d) = \sigma^2 \tag{2.206}$$

o $a_d = \gamma_d^{1/2}$, donde $\gamma_d = \sigma^2/(\sigma^2 + \psi_d)$. El estimador resultante es un estimador restringido BL (CLB):

$$\widehat{\theta}_d^{CLB} = \gamma_d^{1/2}\widehat{\theta}_d + (1-\gamma_d^{1/2})\mu \tag{2.207}$$

(Spjøtvol y Thomsen, 1987). El método de momentos puede usarse para estimar los parámetros del modelo. El estimador resultante es un estimador empírico CLB y su varianza puede estimarse usando el método jackknife.

2.5.2. Método de bayes jerárquico (BJ)

En la aproximación de bayes jerárquica (BJ), se especifica una distribución a priori subjetiva $f(\boldsymbol{\lambda})$ sobre los parámetros del modelo $\boldsymbol{\lambda}$ y se obtiene la distribución a posteriori $f(\boldsymbol{\mu}|\mathbf{y})$ del parámetro de interés de área pequeña (aleatoria) $\boldsymbol{\mu}$, dado el dato $\mathbf{y}$. El modelo de dos fases, $f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\lambda}_1)$ y $f(\boldsymbol{\mu}|\boldsymbol{\lambda}_2)$,

se combina con la a priori subjetiva sobre $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T)^T$, usando el teorema de bayes, para llegar a la a posteriori $f(\boldsymbol{\mu}|\mathbf{y})$. Las inferencias están basadas en $f(\boldsymbol{\mu}|\mathbf{y})$; en particular, un parámetro de interés, digamos $\phi = h(\boldsymbol{\mu})$ se estima por su media a posteriori $\widehat{\phi}^{BJ} = E[h(\boldsymbol{\mu})|\mathbf{y}]$, y la varianza a posteriori $V[h(\boldsymbol{\mu})|\mathbf{y}]$ se usa como una medida de precisión del estimador, con tal de que sean finitos.

La aproximación BJ es fácil, y las inferencias son claras y "exactas", pero requieren de la especificación de una a priori subjetiva $f(\boldsymbol{\lambda})$ sobre el parámetro del modelo $\boldsymbol{\lambda}$. Las a priori sobre $\boldsymbol{\lambda}$ pueden ser informativas o "difusas". Las a prioris informativas están basadas en información a priori sustancial, tal como estudios previos considerados relevantes para el conjunto de datos actual $\mathbf{y}$. Sin embargo, las a prioris informativas raramente están disponibles en aplicaciones reales BJ, particularmente aquellas relacionados a políticas públicas. Las a priori difusas (o no informativas) se diseñan para reflejar la falta de información acerca de $\boldsymbol{\lambda}$. La elección de una a priori difusa no es única y algunas a priori difusas impropias pueden llevar a posteriores impropias. Es por consiguiente esencial asegurarse que la elección de la a priori difusa, $f(\boldsymbol{\lambda})$, lleva a una $f(\boldsymbol{\mu}|\mathbf{y})$ a posteriori apropiada. También, es deseable seleccionar una a priori difusa que conduzca a inferencias bien calibradas en el sentido de validez en el marco frecuentista. En la práctica, el sesgo frecuentista, $E(\widehat{\phi}^{BJ} - \phi)$, del estimador BJ de $\widehat{\phi}^{BJ}$ y el sesgo relativo frecuentista de la varianza a posteriori como un estimador del ECM($\widehat{\phi}^{BJ}$) debe ser pequeño. Además, la cobertura frecuentista de un intervalo BJ sobre ϕ debe estar cercano al nivel nominal (Dawid, 1985; Browne y Draper, 2001).

Aplicando el teorema de bayes, tenemos

$$f(\boldsymbol{\mu}, \boldsymbol{\lambda}|\mathbf{y}) = \frac{f(\mathbf{y}, \boldsymbol{\mu}|\boldsymbol{\lambda}) f(\boldsymbol{\lambda})}{f_1(\mathbf{y})}, \tag{2.208}$$

donde $f_1(\mathbf{y})$ es la densidad marginal de $\mathbf{y}$:

$$f_1(\mathbf{y}) = \int f(\mathbf{y}, \boldsymbol{\mu}|\boldsymbol{\lambda}) f(\boldsymbol{\lambda}) d\boldsymbol{\mu} d\boldsymbol{\lambda}. \tag{2.209}$$

La densidad a posteriori deseada $f(\boldsymbol{\mu}|\mathbf{y})$ es obtenida de (2.208) como

$$f(\boldsymbol{\mu}|\mathbf{y}) = \int f(\boldsymbol{\mu}, \boldsymbol{\lambda}|\mathbf{y}) d\boldsymbol{\lambda} \tag{2.210}$$

$$= \int f(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\lambda}) f(\boldsymbol{\lambda}|\mathbf{y}) d\boldsymbol{\lambda}. \tag{2.211}$$

La forma (2.211) muestra que $f(\boldsymbol{\mu}|\mathbf{y})$ es una mixtura de densidades condicionales $f(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\lambda})$; nótese que $f(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\lambda})$ se usa para inferencias BE. Debido a la

forma de la mixtura (2.211), BJ también es llamado BE bayes o totalmente bayes.

Es claro de (2.208) y (2.210) que la evaluación de $f(\boldsymbol{\mu}|\mathbf{y})$ y cantidades posteriores asociadas tales como $E[h(\boldsymbol{\mu})|\mathbf{y}]$ involucra integración multidimensional. No obstante, es a menudo posible realizar integración analítica con respecto a alguna de las componentes de $\boldsymbol{\mu}$ y $\boldsymbol{\lambda}$. Si el problema reducido involucra integración de una o dos dimensiones, entonces puede usarse la integración numérica directa para calcular las cantidades a posteriori deseadas. Para problemas complejos, sin embargo, se hace necesario evaluar integrales de dimensión superior. Recientemente métodos desarrollados de Monte Carlo mediante Cadenas de Markov (MCMC), parecen superar las dificultades computacionales para una gran dimensión. Estos métodos generan muestras de la distribución a posteriori, y entonces usan las muestras simuladas para aproximar las cantidades a posteriori deseadas. Los paquetes de Software BUGS y CODA implementan MCMC y diagnósticos de convergencia. En la siguiente sección se aplicarán los métodos MCMC; en particular, el algorítmo de muestreo de Gibbs y el algorítmo de Metropolis-Hastings (M-H).

Modelo básico de nivel de área

En esta subsección aplicamos la aproximación BJ para el modelo básico de nivel de área (2.134), asumiendo una distribución a priori sobre los parámetros del modelo $(\boldsymbol{\beta}, \sigma_v^2)$. Primero consideremos el caso de σ_v^2 conocida y asumamos una a priori "monótona" sobre $\boldsymbol{\beta}$ dada por $f(\boldsymbol{\beta}) \propto 1$, y reescribimos (2.134) como un modelo BJ:

$$\begin{aligned} (i) \quad & \widehat{\theta}_d|\theta_d, \boldsymbol{\beta}, \sigma_v^2 \overset{\text{ind}}{\sim} N(\theta_d, \psi_d), \qquad d = 1, \cdots, D \\ (ii) \quad & \theta_d|\boldsymbol{\beta}, \sigma_v^2 \overset{\text{ind}}{\sim} N(\mathbf{z}_d^T\boldsymbol{\beta}, b_d^2\sigma_v^2), \qquad d = 1, \cdots, D \\ & (iii) \quad f(\boldsymbol{\beta}) \propto 1. \end{aligned} \tag{2.212}$$

Entonces extendemos el resultado al caso de σ_v^2 desconocida reemplazando (iii) en (2.212) por

$$\text{(iii)}' \quad f(\boldsymbol{\beta}, \sigma_v^2) = f(\boldsymbol{\beta})f(\sigma_v^2) \propto f(\sigma_v^2), \tag{2.213}$$

donde $f(\sigma_v^2)$ es una a priori sobre σ_v^2.

σ_v^2 conocida: Se ha demostrado por cálculos que la distribución a posteriori de θ_d dado $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \cdots, \widehat{\theta}_D)$ y σ_v^2, bajo el modelo BJ (2.212), es normal con media igual al estimador BLUP $\tilde{\theta}_d^H$ y varianza igual a $M_{1d}(\sigma_v^2)$ dado por (2.138). Es decir, el estimador BJ de θ_d es

$$\tilde{\theta}_d^{BJ}(\sigma_v^2) = E(\theta_d|\widehat{\boldsymbol{\theta}}, \sigma_v^2) = \tilde{\theta}_d^H, \tag{2.214}$$

y la varianza a posteriori de θ_d es

$$V(\theta_d|\widehat{\boldsymbol{\theta}}, \sigma_v^2) = M_{1d}(\sigma_v^2) = \text{ECM}(\tilde{\theta}_d^H). \tag{2.215}$$

Ahora, cuando σ_v^2 se asume conocida y $f(\boldsymbol{\beta}) \propto 1$, las aproximaciones BJ y BLUP bajo normalidad conducen a la misma estimación puntual.

σ_v^2 desconocida: Integración numérica. En la práctica, σ_v^2 es desconocida y es necesario tener en cuenta la incertidumbre en torno a σ_v^2 asumiendo una a priori, $f(\sigma_v^2)$, sobre σ_v^2. El modelo BJ está dado por (i) e (ii) de (2.212) y (iii) de (2.213). Obtenemos el estimador BJ de θ_d como

$$\widehat{\theta}_d^{BJ} = E(\theta_d|\widehat{\boldsymbol{\theta}}) = E_{\sigma_v^2}[\tilde{\theta}_d^{BJ}(\sigma_v^2)], \tag{2.216}$$

donde $E_{\sigma_v^2}$ denota la esperanza con respecto a la distribución a posteriori de σ_v^2, $f(\sigma_v^2|\widehat{\boldsymbol{\theta}})$. La varianza a posteriori de θ_d está dada por

$$V(\theta_d|\widehat{\boldsymbol{\theta}}) = E_{\sigma_v^2}[M_{1d}(\sigma_v^2)] + V_{\sigma_v^2}[\tilde{\theta}_d^{BJ}(\sigma_v^2)], \tag{2.217}$$

donde $V_{\sigma_v^2}$ denota la varianza con respecto a $f(\sigma_v^2|\widehat{\boldsymbol{\theta}})$. Se sigue de (2.216) y de (2.217) que la evaluación de $\widehat{\theta}_d^{BJ}$ y $V(\theta_d|\widehat{\boldsymbol{\theta}})$ involucra solamente integración unidimensional.

La función de distribución a posteriori $f(\sigma_v^2|\widehat{\boldsymbol{\theta}})$ puede obtenerse de la función de verosimilitud restringida $L_R(\sigma_v^2)$ como

$$f(\sigma_v^2|\widehat{\boldsymbol{\theta}}) \propto L_R(\sigma_v^2) f(\sigma_v^2), \tag{2.218}$$

donde

$$\log[L_R(\sigma_v^2)] = \text{const} - \frac{1}{2}\sum_{d=1}^{D}\log(\sigma_v^2 b_d^2 + \psi_d) - \frac{1}{2}\log\left|\sum_{d=1}^{D}(\sigma_v^2 b_d^2 + \psi_d)^{-1}\mathbf{z}_d\mathbf{z}_d^T\right|$$

$$-\frac{1}{2}\sum_{d=1}^{D}[\widehat{\theta}_d - \mathbf{z}_d^T\tilde{\boldsymbol{\beta}}(\sigma_v^2)]^2/(\sigma_v^2 b_d^2 + \psi_d), \tag{2.219}$$

con $\mathbf{z}_d$ es el vector de $p\times 1$ covariables, y $\tilde{\boldsymbol{\beta}}(\sigma_v^2) = \tilde{\boldsymbol{\beta}}$ es el estimador de mínimos cuadrados ponderado de $\boldsymbol{\beta}$; ver Harville (1977).

La media a posteriori de σ_v^2 bajo la a priori monótona $f(\sigma_v^2) \propto 1$ puede ser expresada como

$$E(\sigma_v^2|\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}_{vHB}^2 = \int_0^\infty \sigma_v^2 L_R(\sigma_v^2) d\sigma_v^2 \;/\; \int_0^\infty L_R(\sigma_v^2) d\sigma_v^2. \tag{2.220}$$

Este estimador es siempre único y positivo, al contrario de la moda a posteriori o el estimador MVR de σ_v^2.

El sesgo de $\widehat{\sigma}_{vHB}^2$ puede ser sustancial cuando σ_v^2 es pequeña (Browne y Draper, 2001). Sin embargo, $\widehat{\theta}_d^{BJ}$ puede no verse afectado por el sesgo de $\widehat{\sigma}_{vHB}^2$ porque el estimador combinado $a_d\widehat{\theta}_d + (1-a_d)\mathbf{z}_d^T\widehat{\boldsymbol{\beta}}$ es aproximadamente insesgado para θ_d para cualquier elección del peso $a_d (0 \leq a_d \leq 1)$.

La varianza a posteriori (2.217) se usa como una medida de incertidumbre asociada con $\widehat{\theta}_d^{BJ}$. Es deseable seleccionar una a priori impropia que conduzca a inferencias bien calibradas. En particular, la varianza a posteriori debe ser insesgada de segundo orden para el ECM, es decir, $E[V(\theta_d|\widehat{\boldsymbol{\theta}})] - \text{ECM}(\widehat{\theta}_d^{BJ}) = o(D^{-1})$. Datta, Rao y Smith (2002), mostraron que la a priori correspondiente para el caso especial de $b_d = 1$ está dada por

$$f_d(\sigma_v^2) \propto (\sigma_v^2 + \psi_d)^2 \sum_{l=1}^{D} (\sigma_v^2 + \psi_l)^{-2}. \tag{2.221}$$

La demostración de (2.221) puede verse en Datta et al. (2002).

Para el caso balanceado, $\psi_d = \psi$, la a priori correspondiente (2.221) se reduce a la a priori monótona $f(\sigma_v^2) \propto 1$ y la inferencia BJ resultante tiene justificación dual. Sin embargo, el uso de la a priori monótona cuando la varianza muestral varía significativamente en las áreas podría llevar a varianzas posteriores que no conducen al ECM.

σ_v^2 desconocida: Muestreo de Gibbs. Ahora aplicamos el muestreador de Gibbs al modelo básico de nivel de área, dado por (i) y (ii) de (2.212), asumiendo la a priori (2.213) sobre $\boldsymbol{\beta}$ y σ_v^2 con $\sigma_v^{-2} \sim G(a,b),\ a>0,\ b>0$, es decir, una distribución gamma con parámetro de forma a y parámetro de escala b. Nótese que σ_v^2 se distribuye como una gamma inversa $IG(a,b)$, con $f(\sigma_v^2) \propto \exp(-b/\sigma_v^2)(1/\sigma_v^2)^{a+1}$. Las constantes positivas a y b se fijan en valores muy pequeños (BUGS usa $a = b =$0.001 como el valor fijo por default). Es fácil verificar que las condicionales Gibbs están dadas por

$$\text{(i)} \quad [\boldsymbol{\beta}|\boldsymbol{\theta}, \sigma_v^2, \widehat{\boldsymbol{\theta}}] \sim N_p\left[\boldsymbol{\beta}^*, \sigma_v^2\left(\sum_d \tilde{\mathbf{z}}_d\tilde{\mathbf{z}}_d^T\right)^{-1}\right] \tag{2.222}$$

$$\text{(ii)} \quad [\theta_d|\boldsymbol{\beta}, \sigma_v^2, \widehat{\boldsymbol{\theta}}] \sim N[\widehat{\theta}_d^B(\boldsymbol{\beta}, \sigma_v^2), \gamma_d\psi_d], \quad d = 1, \cdots, D \tag{2.223}$$

$$\text{(iii)} \quad [\sigma_v^{-2}|\boldsymbol{\beta}, \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}] \sim G\left[\frac{D}{2} + a, \frac{1}{2}\sum_d (\tilde{\theta}_d - \tilde{\mathbf{z}}_d^T\boldsymbol{\beta})^2 + b\right], \tag{2.224}$$

donde $\tilde{\theta}_d = \theta_d/b_d,\ \tilde{\mathbf{z}}_d = \mathbf{z}_d/b_d,\ \boldsymbol{\beta}^* = (\sum_d \tilde{\mathbf{z}}_d\tilde{\mathbf{z}}_d^T)^{-1}(\sum_d \tilde{\mathbf{z}}_d\tilde{\theta}_d),\ N_p(\cdot)$ denota una normal p-variada, y $\widehat{\theta}_d^B(\boldsymbol{\beta},\sigma_v^2) = \gamma_d\widehat{\theta}_d + (1-\gamma_d)\mathbf{z}_d^T\boldsymbol{\beta}$ es el estimador de bayes de θ_d. Todas las condicionales Gibbs tienen forma explícita y entonces las muestras MCMC pueden generarse directamente de las condicionales (i)-(iii).

Denotemos las muestras MCMC como $\{(\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma_v^{2(k)}),\ k = r+1, \cdots, r+R\}$. Usando (2.223), podemos obtener estimadores de Rao-Blackwell de la media a posteriori y de la varianza a posteriori de θ_d:

$$\widehat{\theta}_d^{BJ} = \frac{1}{R}\sum_{k=r+1}^{r+R} \widehat{\theta}_d^B(\boldsymbol{\beta}^{(k)}, \sigma_v^{2(k)}) = \widehat{\theta}_d^B(\cdot,\cdot) \tag{2.225}$$

y

$$\widehat{V}(\theta_d|\widehat{\boldsymbol{\theta}}) = \frac{1}{R}\sum_{k=r+1}^{r+R} g_{1d}(\sigma_v^{2(k)}) +$$

$$\frac{1}{R-1}\sum_{k=r+1}^{r+R}\left[\widehat{\theta}_d^B(\boldsymbol{\beta}^{(k)}, \sigma_v^{2(k)}) - \widehat{\theta}_d^B(\cdot,\cdot)\right]^2. \tag{2.226}$$

Pueden obtenerse estimadores más eficientes aprovechando los resultados de forma explícita de la subsección 2.5.2 para σ_v^2 conocida. Tenemos

$$\widehat{\theta}_d^{BJ} = \frac{1}{R}\sum_{k=r+1}^{r+R} \tilde{\theta}_d^H(\sigma_v^{2(k)}) = \tilde{\theta}_d^H(\cdot) \tag{2.227}$$

y

$$\widehat{V}(\theta_d|\widehat{\boldsymbol{\theta}}) = \frac{1}{R}\sum_{k=r+1}^{r+R}\left[g_{1d}(\sigma_v^{2(k)}) + g_{2d}(\sigma_v^{2(k)})\right]$$

$$+\frac{1}{R-1}\sum_{k=r+1}^{r+R}\left[\tilde{\theta}_d^H(\sigma_v^{2(k)}) - \tilde{\theta}_d^H(\cdot)\right]^2. \tag{2.228}$$

Basado en el estimador Rao-Blackwell $\widehat{\theta}_d^{BJ}$, un estimador del total $Y_d = g^{-1}(\theta_d)$ está dado por $g^{-1}(\widehat{\theta}_d^{BJ})$, pero no es igual a la media a posteriori deseada $E(Y_d|\widehat{\boldsymbol{\theta}})$. Sin embargo, las muestras MCMC marginales $\{Y_d^{(k)} = g^{-1}(\theta_d^{(k)})\}$ pueden usarse directamente para estimar la media a posteriori de Y_d como

$$\widehat{Y}_d^{BJ} = \frac{1}{R}\sum_{k=r+1}^{r+R} Y_d^{(k)} = Y_d^{(\cdot)}. \tag{2.229}$$

Similarmente, la varianza a posteriori de Y_d se estima como

$$\widehat{V}(Y_d|\widehat{\boldsymbol{\theta}}) = \frac{1}{R-1}\sum_{k=r+1}^{r+R}\left(Y_d^{(k)} - Y_d^{(\cdot)}\right)^2. \tag{2.230}$$

Si se generan L ejecuciones independientes, en lugar de una sola ejecución larga, entonces la media a posteriori se estima como

$$\widehat{Y}_d^{BJ} = \frac{1}{Lr}\sum_{l=1}^{L}\sum_{k=r+1}^{2r} Y_d^{(lk)} = \frac{1}{L}\sum_{l=1}^{L} Y_d^{(l\cdot)} = Y_d^{(\cdot\cdot)} \tag{2.231}$$

donde $Y_d^{(lk)}$ es el k-ésimo valor obtenido en la l-ésima ejecución de longitud $2r$ con las primeras r iteraciones suprimidas. La varianza a posteriori se estima de $\widehat{V} = \frac{r-1}{r}W + \frac{1}{r}B$:

$$\widehat{V}(Y_d|\widehat{\boldsymbol{\beta}}) = \frac{r-1}{r}W_d + \frac{1}{r}B_d, \tag{2.232}$$

donde $B_d = r\sum_{l=1}^{L}(Y_d^{(l\cdot)} - Y_d^{(\cdot\cdot)})^2/(L-1)$ es la varianza entre ejecuciones y $W_d = \sum_{l=1}^{L}\sum_{k=r+1}^{2r}(Y_d^{(lk)} - Y_d^{(l\cdot)})^2/[L(r-1)]$ es la varianza dentro de ejecuciones.

Modelo básico de nivel unidad

En esta subsección se explica la aproximación BJ para el modelo básico de nivel unidad (2.144) con igual varianza de los errores (es decir, $k_{dj} = 1$), asumiendo una distribución a priori sobre los parámetros del modelo $(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)$. Consideremos primero el caso de σ_v^2 y σ_e^2 conocidas y asumiendo una a priori "monótona" sobre $\boldsymbol{\beta} : f(\boldsymbol{\beta}) \propto 1$. Reescribimos (2.144) como un modelo BJ:

(i) $y_{dj}|\boldsymbol{\beta}, v_d, \sigma_e^2 \overset{\text{ind}}{\sim} N(\mathbf{x}_{dj}^T\boldsymbol{\beta} + v_d, \sigma_e^2), \quad j = 1, \cdots, n_d; \quad d = 1, \cdots, D$

(ii) $v_d|\sigma_v^2 \overset{\text{iid}}{\sim} N(0, \sigma_v^2), \quad d = 1, \cdots, D$

(iii)

$$f(\boldsymbol{\beta}) \propto 1. \tag{2.233}$$

Extendemos entonces el resultado al caso de σ_v^2 y σ_e^2 desconocidas reemplazando (iii) en (2.233) por

$$\text{(iii)}' \quad f(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2) = f(\boldsymbol{\beta})f(\sigma_v^2)f(\sigma_e^2) \propto f(\sigma_v^2)f(\sigma_e^2), \tag{2.234}$$

donde $f(\sigma_v^2)$ y $f(\sigma_e^2)$ son las a prioris sobre σ_v^2 y σ_e^2. Por simplicidad, tomamos $\mu_d = \overline{\mathbf{X}}_d^T\boldsymbol{\beta} + v_d$ como la media de área pequeña d, asumiendo que el tamaño de la población, N_d, es grande.

σ_v^2 y σ_e^2 conocidas: Cuando σ_v^2 y σ_e^2 se asumen conocidas, las aproximaciones BJ y BLUP bajo normalidad conducen a estimaciones puntuales y medidas de variabilidad idénticas, asumiendo una a priori monótona sobre $\boldsymbol{\beta}$. Este resultado, de hecho, es válido para el modelo mixto lineal general con varianza de los parámetros conocida. El estimador BJ de μ_d está dado por

$$\tilde{\mu}_d^{BJ}(\sigma_v^2, \sigma_e^2) = E(\mu_d|\mathbf{y}, \sigma_v^2, \sigma_e^2) = \tilde{\mu}_d^H, \tag{2.235}$$

donde $\mathbf{y}$ es el vector de observaciones muestrales y $\tilde{\mu}_d^H$ es el estimador BLUP dado por (2.148).

σ_v^2 y σ_e^2 desconocidas: Integración numérica. En la práctica, σ_v^2 y σ_e^2 son desconocidas y es necesario tener en cuenta la incertidumbre sobre σ_v^2 y σ_e^2 asumiendo una a priori sobre σ_v^2 y σ_e^2. El modelo BJ está dado por (i) y (ii) de (2.233) y (iii)' dado por (2.234). Obtenemos el estimador BJ de μ_d y la varianza posterior de μ_d como

$$\widehat{\mu}_d^{BJ} = E(\mu_d|\mathbf{y}) = E_{\sigma_v^2,\sigma_e^2}[\tilde{\mu}_d^{BJ}(\sigma_v^2, \sigma_e^2)] \tag{2.236}$$

y

$$V(\mu_d|\mathbf{y}) = E_{\sigma_v^2,\sigma_e^2}[M_{1d}(\sigma_v^2, \sigma_e^2)] + V_{\sigma_v^2,\sigma_e^2}[\tilde{\mu}_d^{BJ}(\sigma_v^2, \sigma_e^2)], \tag{2.237}$$

donde $E_{\sigma_v^2,\sigma_e^2}$ y $V_{\sigma_v^2,\sigma_e^2}$, respectivamente denotan la esperanza y varianza con respecto a la distribución a posteriori $f(\sigma_v^2, \sigma_e^2|\mathbf{y})$.

Como en la subsección 2.5.2, la densidad a posteriori $f(\sigma_v^2, \sigma_e^2|\mathbf{y})$ puede obtenerse de la función de verosimilitud restringida $L_R(\sigma_v^2, \sigma_e^2)$ como

$$f(\sigma_v^2, \sigma_e^2|\mathbf{y}) \propto L_R(\sigma_v^2, \sigma_e^2) f(\sigma_v^2) f(\sigma_e^2). \tag{2.238}$$

Bajo la a priori monótona $f(\sigma_v^2) \propto 1$ y $f(\sigma_e^2) \propto 1$, la posterior $f(\sigma_v^2, \sigma_e^2|\mathbf{y})$ es propia (sujeta a una restricción de tamaño de muestra moderado) y proporcional a $L_R(\sigma_v^2, \sigma_e^2)$. La evaluación de la media a posteriori (2.236) y la varianza a posteriori (2.237), usando $f(\sigma_v^2, \sigma_e^2|\mathbf{y}) \propto L_R(\sigma_v^2, \sigma_e^2)$, involucra integrales de dos dimensiones.

Si asumimos una gamma difusa a priori sobre σ_v^{-2}, $G(a_e, b_e)$ con $a_e \geq 0$ y $b_e \geq 0$, entonces es posible integrar sin σ_e^2 con respecto a $f(\sigma_e^2|\tau_v, \mathbf{y})$, donde $\tau_v = \sigma_v^2/\sigma_e^2$. La evaluación de (2.236) y (2.237) se ha reducido ahora a una integración unidimensional con respecto a la a posteriori de τ_v, $f(\tau_v|\mathbf{y})$.

σ_v^2 y σ_e^2 desconocidas: Muestreo de Gibbs. Explicamos ahora el muestreo de Gibbs aplicado al modelo básico de nivel unidad dado por (i) y (ii) de (2.233), asumiendo la a priori (2.234) sobre $\boldsymbol{\beta}, \sigma_v^2$ y σ_e^2 con $\sigma_v^{-2} \sim G(a_v, b_v)$, $a_v \geq$

$0,\ b_v > 0$ y $\sigma_e^{-2} \sim G(a_e, b_e),\ a_e \geq 0,\ b_e > 0$. Es fácil verificar que las condicionales Gibbs están dadas por

$$\text{(i)} \quad [\boldsymbol{\beta}|\mathbf{v}, \sigma_v^2, \sigma_e^2, \mathbf{y}] \sim N_p\left[\left(\sum_d \sum_j \mathbf{x}_{dj}\mathbf{x}_{dj}^T\right)^{-1} \times \right.$$

$$\left. \sum_d \sum_j \mathbf{x}_{dj}(y_{dj} - v_d), \sigma_e^2 \left(\sum_d \sum_j \mathbf{x}_{dj}\mathbf{x}_{dj}^T\right)^{-1}\right] \tag{2.239}$$

$$\text{(ii)} \quad \text{Para} \quad d = 1, \cdots, D,$$

$$[v_d|\mathbf{v}, \sigma_v^2, \sigma_e^2, \mathbf{y}] \sim \left[\gamma_d(\overline{y}_d - \overline{\mathbf{x}}_d^T\boldsymbol{\beta}), g_{1d}(\sigma_v^2, \sigma_e^2) = \gamma_d\sigma_e^2/n_d\right] \tag{2.240}$$

$$\text{(iii)} \quad [\sigma_e^{-2}|\boldsymbol{\beta}, \mathbf{v}, \sigma_v^2, \mathbf{y}] \sim$$

$$G\left[\frac{n}{2} + a_e, \frac{1}{2}\sum_d \sum_j (y_{dj} - \mathbf{x}_{dj}^T\boldsymbol{\beta} - v_d)^2 + b_e\right] \tag{2.241}$$

$$\text{(iv)} \quad [\sigma_e^{-2}|\boldsymbol{\beta}, \mathbf{v}, \sigma_e^2, \mathbf{y}] \sim G\left(\frac{D}{2} + a_v, \frac{1}{2}\sum_d v_d^2 + b_v\right), \tag{2.242}$$

donde $n = \sum_d n_d,\ \mathbf{v} = (v_1, \cdots, v_D)^T,\ \overline{y}_d = \sum_j y_{dj}/n_d,\ \overline{\mathbf{x}}_d = \sum_j \mathbf{x}_{dj}/n_d$, y $\gamma_d = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/n_d)$. Observe que todas las condicionales de Gibbs tienen forma explícita y entonces las muestras MCMC pueden ser generadas directamente de las condicionales (i) a (iv). Puede usarse BUGS para generar muestras de las anteriores condicionales usando cualquier a priori gamma inversa (con $a_e = b_e$ =0.001 y $a_v = b_v$ =0.001) o con valores especificados de a_e, b_e, a_v y b_v.

Denotando las muestras MCMC de una sola ejecución larga por $\{\boldsymbol{\beta}^{(k)}, \mathbf{v}^{(k)}, \sigma_v^{2(k)}, \sigma_e^{2(k)}, k = r+1, \cdots, r+R\}$. La muestra marginal MCMC $\{\boldsymbol{\beta}^{(k)}, \mathbf{v}^{(k)}\}$ puede usarse directamente para estimar la media a posteriori de μ_d como

$$\widehat{\mu}_d^{BJ} = \frac{1}{R}\sum_{k=r+1}^{r+R} \mu_d^{(k)} = \mu_d^{(\cdot)}, \tag{2.243}$$

donde $\mu_d^{(k)} = \overline{\mathbf{X}}_d^T\boldsymbol{\beta}^{(k)} + v_d^{(k)}$. Similarmente, la varianza a posteriori de μ_d se estima como

$$V(\mu_d|\mathbf{y}) = \frac{1}{R-1}\sum_{k=r+1}^{r+R}\left(\mu_d^{(k)} - \mu_d^{(\cdot)}\right)^2. \tag{2.244}$$

2.6. Otros métodos de estimación

Aunque en general, los métodos de estimación de áreas pequeñas descritos hasta ahora producen buenas estimaciones, en la última década se han desarrollado otros métodos para abordar problemas específicos o problemas en los cuales se pueda prescindir de supuestos tan restrictivos como la distribución de los datos. En este grupo podemos clasificar los métodos basados en modelos m-quantile para la estimación en áreas pequeñas propuestos por Chambers, R. et al. (2006) y los modelos no paramétricos, uno de los cuales fue propuesto por Salvati, N. et al. (2010).

2.6.1. Estimación con regresión m-quantile

De acuerdo con lo descrito hasta ahora, sabemos que los métodos de estimación basados en modelos pueden clasificarse en dos categorías, los métodos basados en modelos de efectos fijos, es decir, modelos que explican la variación entre áreas en la variable de interés usando únicamente la información auxiliar y, métodos basados en modelos de efectos mixtos (aleatorios) que incluyen efectos aleatorios de área específica para considerar la variación entre áreas, además de la explicada por la información auxiliar.

Los modelos de efectos mixtos han sido usados ampliamente en la estimación de áreas pequeñas (Fay y Herriot, 1979, Battese, et al. 1988, Rao, 2003). Sin embargo, tales modelos dependen de supuestos paramétricos y distribucionales, así como también de la especificación de la parte aleatoria del modelo. Además, la inferencia robusta bajo estos modelos no es comprensible.

En estos modelos se asume que el vector de p variables auxiliares x_{ij} es conocido para cada unidad poblacional i en el área pequeña j y la información para la variable de interés y está disponible para las unidades en la muestra. El objetivo es usar estos datos para estimar varias cantidades de área específica, incluyendo, la media m_j de y del área j. En el caso general un modelo lineal de efectos fijos es de la forma

$$y_{ij} = x_{ij}^T\beta + z_{ij}^T\gamma_j + \varepsilon_{ij} \qquad (i = 1 \cdots, n, j = 1, \cdots, d),$$

donde γ_j denota un vector de efectos aleatorios y z_{ij} denota un vector de variables auxiliares cuyos valores son conocidos para todas las unidades en la población. La estimación de los parámetros del modelo puede realizarse usando máxima verosimilitud o máxima verosimilitud restringida, usualmente bajo el supuesto de normalidad. Las medias de dominio se estiman por

$$\widehat{m}_j = N_j^{-1}\left\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}(x_{ij}^T\widehat{\beta} + z_{ij}^T\widehat{\gamma}_j)\right\},$$

donde s_j denota las n_j unidades muestreadas en el área j y r_j denota las restantes $N_j - n_j$ unidades en el área. Este estimador es más comúnmente conocido como mejor predictor empírico lineal insesgado (EBLUP) de m_j (Henderson, 1953). Pueden destacarse dos casos especiales de este estimador. El primero es cuando z_{ij} consiste de un conjunto de indicadores para las áreas pequeñas, en cuyo caso los γ_j son efectos escalares de área pequeña; entonces nos referimos al estimador EBLUP como un estimador de "interceptos aleatorios". El segundo es la situación mas general donde los z_{ij} incluyen indicadores de área pequeña así como otras covariables, en cuyo caso nos referimos al estimador EBLUP como estimador de "pendientes aleatorias". El papel de los efectos aleatorios en el modelo es para caracterizar diferencias en la distribución condicional de y dado x entre las áreas pequeñas. Sin embargo, el modelo de efectos aleatorios no es la única manera en que podemos realizar tal caracterización.

Una aproximación alternativa a la estimación en áreas pequeñas está basada en la modelación de los coeficientes quantile-like de la distribución condicional de la variable de interés, dadas las covariables. Se evitan los efectos aleatorios, en cambio, la variación inter dominio es caracterizada por la variación en valores específicos de área de estos coeficientes quantile-like. Es fácil ajustar estos nuevos modelos y tienen un buen número de ventajas prácticas, incluyendo fácil especificación no paramétrica y simple incorporación de pesos de muestreo.

Regresión quantile.

La teoría clásica de modelos estadísticos lineales es fundamentalmente una teoría de esperanzas condicionales; es decir, un modelo de regresión resume el comportamiento de la media de y a cada punto en un conjunto de x's (Mosteller y Tukey, 1977). Desafortunadamente, este resumen proporciona una imagen bastante incompleta, de la misma manera como la media da una imagen muy incompleta de una distribución. Es normalmente mucho mejor ajustar una familia de modelos de regresión, cada uno resumiendo el comportamiento de un punto porcentual diferente, o quantile, de y para cada punto en este conjunto de x's. Tal ejercicio de modelado se llama usualmente como una regresión quantile.

En el caso lineal, la regresión quantile conduce a una familia de hiperplanos de regresión indexados por coeficientes $q \in (0, 1)$ (Koenker y Bassett, 1978). Para cada valor de q, los planos correspondientes muestran como $Q_q(x)$, el q-ésimo quantile de la distribución condicional de y dado x, varía con x. Un modelo lineal para el q-ésimo quantile condicional de y dado un vector de

covariables x es $Q_q(x) = x^T\beta_q$. El vector β_q se estima minimizando

$$\sum_{i=1}^{n} \left[|y_i - x_i^T b|\{(1-q)I(y_i - x_i^T b \leq 0) + qI(y_i - x_i^T b > 0)\}\right]$$

con respecto a b. La solución para este problema de minimización requiere métodos de programación lineal (Koenker y D'Orey, 1987) y funciones para ajustar planos de regresión quantile, disponibles en software estadístico estándar, tal como R.

Regresión m-quantile

La regresión quantile puede ser vista como una generalización de regresión de la mediana. Similarmente, expectile regression (Newey y Powell, 1987) es una generalización "quantile-like" de la regresión de la media. En regresión m-quantile (Breckling y Chambers, 1988) estos conceptos son integrados en una estructura común definida por una generalización "quantile-like" de regresión robusta basada en funciones de influencia.

El m-quantile de orden q para la densidad condicional de y dado x se define como la solución $Q_q(x;\psi)$ de la ecuación de estimación $\int \psi_q(y-Q)f(y|x)dy = 0$, donde ψ denota la función de influencia asociada con el m-quantile. Un modelo de regresión lineal m-quantile es uno en el cual el m-quantile condicional de orden q permanece en el plano $Q_q(x) = x^T\beta_q$. Para q y ψ especificados, los estimadores de los parámetros de regresión m-quantile pueden obtenerse resolviendo las ecuaciones de estimación

$$\sum_{i=1}^{n} \psi_q(r_{iq\psi})x_i = 0, \tag{2.245}$$

donde $r_{iq\psi} = y_i - x_i^T\beta_\psi(q)$, $\psi_q(r_{iq\psi}) = 2\psi(s^{-1}r_{iq\psi})\{qI(r_{iq\psi} > 0) + (1-q)I(r_{iq\psi} \leq 0)\}$ y s es un estimador de escala robusto apropiado, tal que la desviación absoluta media del estimador $s = \text{med}|r_{iq\psi}|/0{,}6745$. Es fácil ver que cualquier función de influencia válida puede usarse como base para la regresión m-quantile. Vamos a considerar la propuesta 2 de Hubel de la función de influencia, $\psi(u) = uI(-c \leq u \leq c) + c\ \text{sgn}(u)$, donde c está acotada en un entorno de cero.

Se usa regresión m-quantile en lugar de modelos de regresión quantile "estándar" como la base de este método por razones esencialmente prácticas. Los algoritmos para ajustar modelos de regresión quantile no garantizan necesariamente convergencia y solución única. En contraste, el algoritmo de mínimos cuadrados reponderados usado para ajustar una regresión m-quantile

converge a una única solución (Kokic et. al., 1997) cuando se usa una función de influencia monótona continua.

Un problema común para todos los planos de regresión ajustados "tipo quantile" es que pueden cruzar por diferentes valores de q. Esto implica que la condicional estimada m-quantile definida por estos planos no es entonces completamente monótona en q para todo x. Este es un problema de muestra finita y es usualmente debido a una combinación de falta de especificación del modelo, colinealidad y datos con valores influyentes. Fijemos $q_k < q_j < 0{,}5$. En el caso del escalar x, la monotonía en x requiere que $\beta_0(q_k) + \beta_1(q_k)x \leq \beta_0(q_j) + \beta_1(q_j)x$. Para conseguir esto sobre una cuadrícula de q valores y sobre el rango (x_{min}, x_{max}), uno puede tocar por abajo ligeramente con la linea ajustada correspondiente a los valores de $q < 0{,}5$ y tocar por arriba ligeramente con las líneas ajustadas correspondientes a los valores $q > 0{,}5$. Esto se logra cambiando el punto de inicio o el punto final de la línea ajustada definida por el valor más pequeño (más grande) de q para que sea más pequeño (más grande) que el punto correspondiente de la línea ajustada obtenida por el valor más grande (más pequeño) de q. Diversos estudios sobre monotonía pueden consultarse por ejemplo en un informe técnico presentado por R. Chambers en http://eprints.soton.ac.uk/14075/, o también en Koenker (1984), He (1997), Aragon et al. (2006), Mukarjee y Stern, (1994) y Robertson et al., (1988). Una aplicación del uso de modelos no paramétricos m-quantile puede verse en Pratesi, M. et al. (2008).

Estimación para características de área pequeña

Supongamos que el m-quantile condicional sigue un modelo lineal, con $\beta_\psi(q)$ una función suficientemente uniforme de q tal que las siguientes aproximaciones de primer orden se conservan:

$$m_j = N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \beta_\psi(q_i)\right\}$$

$$\doteq N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \beta_\psi(\theta_j)\right\} + N_j^{-1}\sum_{i\in r_j} x_i^T\left\{\frac{\partial \beta_\psi(\theta_j)}{\partial \theta_j}\right\}(q_i - \theta_j).$$

Aquí θ_j es el valor promedio de los coeficientes m-quantile de las unidades en el área j. Usualmente el primer término en el lado derecho de la anterior expresión tiende a dominar, sugiriendo un predictor de m_j de la forma

$$\widehat{m}_j = N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \widehat{\beta}_\psi(\widehat{\theta}_j)\right\}. \tag{2.246}$$

Son posibles definiciones alternativas de θ_j y de los estimadores $\widehat{\theta}_j$ de estas cantidades, como la mediana del área j del coeficiente m-quantile de las unidades, es decir, podemos referirnos a θ_j como el coeficiente m-quantile del área j.

Una aproximación ligeramente diferente habría sido el ajustar coeficientes de regresión m-quantile $\widehat{\beta}_\psi(q_i)$ correspondientes a los coeficientes m-quantile, q_i, de todos los individuos en la misma área y entonces usar estos coeficientes promedio en (2.246).

Independientemente de cómo se definan los coeficientes m-quantile θ_j para el área j, (2.246) es equivalente usando $x_i^T \beta_\psi(\theta_j)$ para predecir los valores no observados y_i para una unidad no muestreada i en el área j. Esto sugiere que valores predichos para otras características de área pequeña pueden también calcularse usando estas predicciones de nivel unidad. Por ejemplo, un estimador de la función de distribución de la población finita definida por los valores y en el área j implicados por esta aproximación es

$$\widehat{F}_j(t) = N_j^{-1} \left[\sum_{i \in s_j} I(y_i \leq t) + \sum_{i \in r_j} I\{x_i^T \widehat{\beta}_\psi(\widehat{\theta}_j) \leq t\} \right]. \tag{2.247}$$

Estimadores de los quantiles de la distribución de y dentro del área j son fácilmente derivados de (2.247). En particular, la mediana estimada de los valores y en el área j es el valor mediano del conjunto $\{y_i; i \in s_j\} \cup \{x_i^T \widehat{\beta}_\psi(\widehat{\theta}_j); i \in r_j\}$, con otros quantiles estimados definidos similarmente. El error cuadrático medio de las estimaciones se obtiene iterativamente usando el algoritmo de mínimos cuadrados ponderados.

2.6.2. Estimación en áreas pequeñas usando estimadores no paramétricos directos

Los métodos de estimación en áreas pequeñas, generalmente incluyen el uso de estimadores indirectos. Describiremos a continuación una aproximación desarrollada por Salvati, et al. (2010), en la cual, además de usar métodos no paramétricos de estimación, hace uso de un estimador directo, con objeto de ofrecer una alternativa a la estimación indirecta que es muy sensible a la falta de especificación del modelo, colinealidad y valores influyentes. Chandra y Chambers (2005, 2009) han propuesto recientemente una estimación directa basada en el modelo (MBDE) como una composición entre estos dos extremos. El MBDE para la media de área pequeña es un promedio ponderado de los valores de la muestra del área, pero con pesos derivados de un predictor lineal de la correspondiente media poblacional bajo un modelo lineal con efectos alea-

torios de área. Los pesos usados en el MBDE están basados en una extensión de Royal (1976) para el caso de un modelo de efectos mixtos.

Cuando la forma funcional de la relación entre la variable respuesta y las covariables es desconocida o tiene una forma funcional complicada, una aproximación basada en el uso de un modelo no lineal puede ofrecer ventajas significativas comparada con una basada en un modelo lineal. En esta subsección nos enfocaremos en un modelo de regresión no paramétrico basado en una aproximación p-spline para la verdadera función de regresión (ver Eilers y Marx, 1996; Ruppert et al., 2003) y se extenderá al caso donde la relación entre la variable de interés y un subconjunto de covariables del modelo a nivel de población es no lineal. En particular, se usa un p-spline para modelar esta no linealidad, aun cuando al mismo tiempo se incluyan efectos aleatorios de área pequeña. El resultado no paramétrico MBDE, en adelante NPMBDE, es entonces una suma ponderada de los valores muestrales del área pequeña de interés, con pesos derivados del EBLUP de la media poblacional definida por esto modelos de regresión p-spline con efectos aleatorios de área.

Estimación de áreas pequeñas basada en regresión spline penalizada

La suavización no paramétrica es una manera popular de modelar una relación de regresión no lineal, y la suavización de modelos basados en p-splines son particularmente atractivos porque representan una extensión relativamente simple de los modelos de regresión lineal (Eilers y Marx, 1996).

El EBLUP no paramétrico: Sea y_j el valor de la variable de interés y y x_j el valor de la variable auxiliar x asociada con la unidad j. Para el caso univariado, el modelo de regresión fundamental se escribe como $y_j = m(x_j)+\varepsilon_j$, donde los ε_j son variables aleatorias independientes con media cero. La función $m(x)$ es desconocida y se asume que puede ser aproximado suficientemente bien por

$$m(x, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)_+^p, \tag{2.248}$$

donde p es el grado del spline, $(t)_+^p = t^p$ si $t > 0$ y 0 en otro caso, κ_k para $k = 1, \cdots, K$ es un conjunto de contantes fijas llamadas nodos, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$ es el vector de coeficientes de la porción paramétrica del modelo y $\boldsymbol{\gamma} = (\gamma_0, \cdots, \gamma_p)^T$ es el vector de coeficientes spline. Si el número de nodos K es suficientemente grande, la clase de funciones en la expresión (2.248) puede aproximar la mayoría de las funciones suaves. Ruppert et al. (2003) sugiere el uso de un nudo para cualesquiera cuatro observaciones, hasta un máximo de 40 nodos para una aplicación univariante. Note que la función de aproximación

$m(x, \boldsymbol{\beta}, \boldsymbol{\gamma})$ en la ecuación (2.248) usa funciones base polinomiales truncadas por simplicidad.

Usando un número grande de nodos en la expresión (2.248) puede conducir a un ajuste inestable. Para superar este problema, un límite superior es normalmente impuesto sobre el tamaño del vector de coeficientes spline $\boldsymbol{\gamma}$. Estimado $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ por minimización de los cuadrados de la desviaciones del modelo (2.248) de los actuales valores de datos sujeto a estas restricciones es equivalente a minimizar la siguiente función de pérdida penalizada

$$\sum_j (y_j - m(x_j, \boldsymbol{\beta}, \boldsymbol{\gamma}))^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{\gamma}. \tag{2.249}$$

Aquí λ es un multiplicador de Lagrange que controla el nivel de suavización de los resultados ajustados y puede definirse, por ejemplo, por validación cruzada. Alternativamente, Wand (2003) y Ruppert et al. (2003) notaron la equivalencia entre la expresión de minimización (2.249) y la maximización de la verosimilitud de la variable respuesta y los coeficientes spline bajo un modelo de efectos aleatorios, para que la solución a la expresión (2.249) defina el BLUP de $m(x_j, \boldsymbol{\beta}, \boldsymbol{\gamma})$. En particular, sea $\mathbf{y} = (y_1, \cdots, y_N)^T$, donde N es el tamaño de la población,

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^p \end{bmatrix}$$

y

$$\mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+^p & \cdots & (x_1 - \kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_N - \kappa_1)_+^p & \cdots & (x_N - \kappa_K)_+^p \end{bmatrix}.$$

La aproximación spline de la ecuación (2.248) puede escribirse entonces como el modelo lineal mixto

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e} \tag{2.250}$$

donde γ y $\mathbf{e}$ se sume ahora que son vectores aleatorios independientes Gaussianos de dimensión K y N, respectivamente. En particular, se asume que

$$\gamma \sim N_K(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_K) \qquad y \qquad \mathbf{e} \sim N_N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N) \tag{2.251}$$

donde $\mathbf{I}_t$ denota la matriz identidad de dimensión t. Opsomer et al. (2008) usaron p-splines en el contexto de estimación de áreas pequeñas agregando efectos aleatorios de área pequeña al modelo (2.250) para capturar las disimilaridades entre áreas pequeñas que no son explicadas por las covariables incluidas en el

modelo. Sea A el número de áreas pequeñas. Entonces el modelo (2.250) puede ser extendido a

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} + \mathbf{e} \tag{2.252}$$

donde $\mathbf{D} = (\mathbf{d}_1, \cdots, \mathbf{d}_N)^T$ es una matriz de covariables conocidas de dimensión $N \times A$ caracterizando diferencias entre las áreas y $\mathbf{u}$ es el vector de A efectos aleatorios de área. En el caso más simple, $\mathbf{D}$ está dado por una matriz cuya i-ésima columna, para $i = 1, \cdots, A$, es una variable indicadora que toma los valores 1 si una unidad está en el área i y es 0 en otro caso. Si se asume que los efectos de área $\mathbf{u}$ están distribuidos independientemente de los efectos spline γ y los efectos individuales $\mathbf{e}$, con $\mathbf{u} \sim N_A(\mathbf{0}, \sigma_u^2\mathbf{I}_A)$, así que la matriz de covarianza del vector $\mathbf{y}$ está dada por

$$Var(\mathbf{y}) = \mathbf{V} = \sigma_\gamma^2\mathbf{Z}\mathbf{Z}^T + \sigma_u^2\mathbf{D}\mathbf{D}^T + \sigma_e^2\mathbf{I}_N.$$

Los parámetros σ_γ^2, σ_u^2 y σ_e^2 son llamados usualmente los componentes de varianza de (2.252).

Asumimos que el método de muestreo usado es no informativo para los valores de $\mathbf{y}$ de la población dados los correspondientes valores de las variables auxiliares y se conocen las afiliaciones de área de las unidades de la población. Se sigue que podemos particionar $\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{D}$ y $\mathbf{e}$ en componente definidas por la unidades n de la población muestreadas y $N-n$ no muestreadas, denotadas por los subíndices s y r, respectivamente. Podemos por consiguiente escribir (2.252) como sigue:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{bmatrix} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} \mathbf{D}_s \\ \mathbf{D}_r \end{bmatrix} \mathbf{u} + \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \end{bmatrix}, \tag{2.253}$$

con matriz de varianza dada por

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}. \tag{2.254}$$

Así, $\mathbf{X}_s$ en (2.253) representa la matriz definida por por los valores de la muestra n del vector de variables auxiliares, mientras $\mathbf{V}_{rr}$ en (2.254) es la matriz de covarianzas de la variable respuesta entre las unidades $N-n$ no muestreadas. Usamos un subíndice de i para denotar la restricción al área pequeña i. El tamaño de la población global es $N = \sum_{i=1}^{A} N_i$, y el tamaño de la muestra total es $n = \sum_{i=1}^{A} n_i$. Similarmente, $s_i(r_i)$ el conjunto de la unidades de la población no muestreadas del área i, y $U_i = s_i \cup r_i$ denota el conjunto de unidades de la población que constituyen el área pequeña i.

Cuando los componentes de varianza son conocidos, la teoría establecida sugiere (ver McCulloch y Searle, 2001) el siguiente estimador de mínimos cuadrados generalizados de $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_s\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s\mathbf{V}_{ss}^{-1}\mathbf{y}_s, \tag{2.255}$$

y el siguiente mejor predictor lineal insesgado ($BLUP_s$) para $\boldsymbol{\gamma}$ y $\mathbf{u}$

$$\widehat{\boldsymbol{\gamma}} = \sigma_\gamma^2 \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}}) \quad y \quad \widehat{\mathbf{u}} = \sigma_u^2 \mathbf{D}_s^T \mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}}). \tag{2.256}$$

En la práctica los componentes de varianza son desconocidos y deben estimarse de los datos muestrales usando métodos tales como máxima verosimilitud o máxima verosimilitud restringida. En lo que sigue usaremos ($\widehat{\sigma}_\gamma^2$, $\widehat{\sigma}_u^2$, $\widehat{\sigma}_e^2$) para denotar tales estimaciones, permitiéndonos definir el estimador plug-in $\mathbf{V}_{ss} = \widehat{\sigma}_\gamma^2 \mathbf{Z}_s \mathbf{Z}_s^T + \widehat{\sigma}_u^2 \mathbf{D}_s \mathbf{D}_s^T + \widehat{\sigma}_e^2 \mathbf{I}_n$, donde $\mathbf{I}_n$ es la matriz identidad de orden n. Esto conduce al mejor estimador empírico lineal insesgado $\widehat{\boldsymbol{\beta}}^{EBLUE} = (\mathbf{X}_s \widehat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s \widehat{\mathbf{V}}_{ss}^{-1} \mathbf{y}_s$ y al empírico $\widehat{\boldsymbol{\gamma}}^{EBLUP} = \widehat{\sigma}_\gamma^2 \mathbf{Z}_s^T \widehat{\mathbf{V}}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s \widehat{\boldsymbol{\beta}}^{EBLUE})$ y $\widehat{\mathbf{u}}^{EBLUP} = \widehat{\sigma}_u^2 \mathbf{D}_s^T \widehat{\mathbf{V}}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s \widehat{\boldsymbol{\beta}}^{EBLUE})$.

Bajo (2.252), el EBLUP para para la media $\overline{y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_j$ de y en el área pequeña es

$$\widehat{\overline{y}}_i^{NPBLUP} = N_i^{-1} \left[\sum_{j \in s_i} y_j + \sum_{j \in r_i} \widehat{y}_j^{NPBLUP} \right], \tag{2.257}$$

donde $\widehat{y}_j^{NPBLUP} = \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}^{EBLUE} + \mathbf{z}_j^T \widehat{\boldsymbol{\gamma}}^{EBLUP} + \mathbf{d}_j^T \widehat{\mathbf{u}}^{EBLUE}$, y $\mathbf{x}_j^T$, $\mathbf{z}_j^T$ y $\mathbf{d}_j^T$ denotan las filas de $\mathbf{X}$, $\mathbf{Z}$ y $\mathbf{D}$, respectivamente, que corresponden a la unidad j en el área i. El EBLUP (2.257) es referido como el EBLUP no paramétrico o NPEBLUP y es propuesto por Opsomer et al. (2008), quien estudió sus propiedades teóricas y proporcionó un estimador analítico y un estimador basado en bootstrap no paramétrico mas intensivo computacionalmente para su ECM. En particular, estos autores proporcionaron una aproximación de segundo orden para el error cuadrático medio del NPEBLUP.

Capítulo 3

Estimación para variables cualitativas

3.1. Introducción

La estimación de la proporción poblacional de individuos que optan por una opción A, frente a otra opción B o la proporción de individuos de una población que presentan una característica A, frente a otros que no la presentan, son ejemplos comunes de una gran variedad de situaciones prácticas que surgen en diversos campos como ensayos clínicos, experimentos farmacéuticos, estadística médica, estudios de opinión, investigación de mercados, etc. En situaciones como las descritas anteriormente, estamos interesados en estimar la proporción poblacional P de individuos que poseen cierto atributo. Para hacerlo, tomamos una muestra s de tamaño $n < N$, bajo un diseño de muestreo prefijado y si se satisfacen ciertas condiciones, usamos la proporción muestral p como estimador de la proporción poblacional P, es decir

$$\widehat{P} = p = n^{-1} \sum_{s} A_i,$$

donde es evidente que $\widehat{P}$ es la frecuencia relativa de individuos que poseen el atributo A o, en términos de probabilidad podemos afirmar que, si $\widehat{P}$ es la probabilidad que un individuo posea la caracterśtica A, el número esperado de los que poseen la característica es simplemente $N\widehat{P}$.

Aunque la variable de interés tenga más de dos opciones de respuesta, siempre es posible organizar los datos en dos grupos, los que poseen la característica de interés y los que no la poseen. Bajo este supuesto, la i-ésima respuesta $A_i \sim Be(p)$, de tal forma que para la muestra s, si se conservan las características de la población, $\sum_s A_i \sim B(n, p)$.

A partir de la distribución de probabilidad de la respuesta, se han propuesto distintos métodos para estimar proporciones, incluyendo la estimación directa, sintética y compuesta, bajo el enfoque basado en el diseño con probabilidades de inclusión prefijadas; considerando probabilidades condicionales en tablas de contingencia para respuesta binaria o politómica, modelos de regresión logística, regresión de Poisson, modelos de regresión para muestras pareadas, estimación con datos correlados y con valores faltantes, etc. (ver por ejemplo, Fleiss et al. (2003), Agresti (1996) y Fienberg, S.E (2007)), aunque hasta ahora, no hay una discusión completa sobre los métodos de estimación para variables no continuas (por ejemplo, variables categóricas o de frecuencias) cuyo parámetro central son las proporciones de áreas pequeñas, como se ha hecho para variables continuas (Rao, 2003). En años recientes, bajo el enfoque de estimación basada en modelos, se han propuesto otros métodos como "empirical best prediction" (EBP) y "hierarchical bayes" (HB) para hacer inferencia sobre proporciones, como se ha descrito en las subsecciones (2.5.1) y (2.5.2). Los métodos de estimación que hemos considerado hasta ahora, para variables cuantitativas, pueden adaptarse a la estimación de proporciones en dominios.

Supongamos que U_d denota un dominio (o subpoblación) de interés y que deseamos estimar la proporción de individuos que poseen una característica de interés, es decir, y_j es binaria. Escribimos Y_j como A_j, donde $A_j = 1$ si la unidad poblacional posee la característica A y $A_j = 0$, en caso contrario. Definimos

$$A_{dj} = \begin{cases} A_j, & \text{si } j \in U_d; \\ 0, & \text{en otro caso,} \end{cases} \tag{3.1}$$

$$a_{dj} = \begin{cases} 1, & \text{si } j \in U_d; \\ 0, & \text{en otro caso,} \end{cases} \tag{3.2}$$

con

$$T(A_d) = \sum_{j \in U} A_{dj} = \sum_{j \in U_d} A_j = T_{A_d}$$

y

$$T(a_d) = \sum_{j \in U} a_{dj} = \sum_{j \in U_d} 1 = N_d.$$

Notemos que A_{dj} puede escribirse también como $a_{dj}A_j$. Si los dominios de interés, digamos, $U_1, \cdots, U_D$ forman una partición de U (o de un dominio grande), es deseable desde el punto de vista del usuario garantizar que las estimaciones de totales de dominio sumen la estimación del total poblacional.

En ausencia de información auxiliar poblacional, usamos el estimador expandido del total

$$\widehat{T}_{A_d} = \sum_{j \in s} w_j A_{dj} = \sum_{j \in s_d} A_j, \tag{3.3}$$

donde s_d denota los elementos de la muestra que pertenecen al dominio U_d.

La proporción de dominio $P_{A_d} = T(A_d)/A(a_d)$ se estima por

$$\widehat{P}A_d = \frac{\widehat{T}(A_d)}{\widehat{T}(a_d)} = \frac{\widehat{T}_{A_d}}{\widehat{N}_d}, \tag{3.4}$$

donde N_d puede ser conocido (situación poco frecuente en la práctica) o puede ser una variable aleatoria, dando origen a un estimador de razón de totales, cuya estimación se hará en el capítulo siguiente.

Nuestra contribución a la estimación de áreas pequeñas se desarrolla bajo el enfoque basado en el diseño, por esta razón presentaremos en primer lugar la descripción de los métodos basados en el modelo para la estimación de proporciones en áreas pequeñas, y dejaremos para el final la descripción de los métodos basados en el diseño.

3.2. Estimación basada en modelos

3.2.1. Modelos básicos de nivel de área

Los estimadores para variables cuantitativas específicos de dominio, también pueden aplicarse a proporciones de dominio, siempre que tenga significado la estimación de una media para variables cualitativas o la asociación entre variables para el caso de los modelos de enlace, toda vez que la estimación en los modelos básicos de nivel de área (2.50), el modelo multivariante de Fay-Herriot, el modelo básico de nivel unidad (2.92) y los modelos mixtos (2.109), hacen uso de un modelo de regresión en la fase de estimación, para incorporar la información auxiliar disponible y mejorar las estimaciones.

3.2.2. Modelos de efectos mixtos y estimadores EBLUP

Los modelos básicos de nivel de área, pueden considerarse como casos especiales de un modelos mixto lineal general de efectos fijos y aleatorios, por lo que para la estimación de proporciones en modelos de efectos mixtos y estimadores EBLUP, no hace falta nueva teoría y puede hacerse a partir de los resultados de Rao (2003), haciendo ajustes a expresiones que involucran medias y redefiniendo expresiones para las varianzas y covarianzas. Los modelos lineales mixtos tienen una larga historia, pero han recibido especial interés en las últimas décadas. Esto es en parte debido a la pesada carga computacional de los métodos de estimación usando tales modelos. Recientes desarrollos en hardware computacional, software y métodos de estimación han contribuido

a aumentar la atención sobre el uso de los modelos mixtos para el análisis de datos.

Los modelos lineales mixtos son muy buenos para predecir combinaciones lineales de efectos fijos y aleatorios, la cual es una de sus más atractivas propiedades. En una serie de trabajos, Henderson (1948, 1949, 1959, 1963, 1973, 1975) desarrolló los estimadores BLUP, "mejor predictor lineal insesgado", para los modelos mixtos.

Los métodos BLUP se han convertido en un procedimiento poderoso y ampliamente usado para ajustar modelos para tendencias genéticas en poblaciones animales basados en rasgos medidos en escalas continua y categórica. Sin embargo, los métodos BLUP descritos por Henderson (1949-1975) asumen conocidas la varianza asociada a los efectos aleatorios en el modelo mixto (componentes de varianza). En la práctica, por supuesto, los componentes de varianza son desconocidos y deben estimarse de los datos. Existen varios métodos para estimar componentes de varianza. Harville (1977) revisa estos métodos, incluyendo máxima verosimilitud y máxima verosimilitud residual y otros tres métodos sugeridos por Henderson. El predictor obtenido del estimador BLUP cuando los componentes de varianza son desconocidos y son reemplazados por estimadores asociados recibe el nombre de estimador EBLUP "mejor predictor empírico lineal insesgado" y se describe en Harville (1990), Robinson (1990), Harville (1991) y Rao (2003). En las últimas décadas se han desarrollado varias aproximaciones importantes para estimar/predecir el valor de una combinación lineal de efectos fijos y aleatorios en variable respuesta de datos discretos. En prácticamente todas estas aproximaciones los efectos aleatorios son normalmente distribuidos. Schall (1991), Breslow y Clayton (1994), McGilchrist y Aisbett (1991) y McGilchrist (1994), han extendido el EBLUP a modelos lineales generalizados. Tres técnicas de expansión de verosimilitud, incluyendo la aproximación de Salomon y Cox (1992) se comparan en Breslow y Lin (1995) y Lin y Breslow (1996). Zeger y Karim (1991) introdujeron una aproximación del muestreo de Gibbs para modelos lineales mixtos generalizados. Las aproximaciones computacionales Monte Carlo EM (MCEM) y Monte Carlo Newton-Raphson (MCNR) se describen en McCulloch (1994) y McCulloch (1997), respectivamente y brevemente en el capítulo anterior, dedicado a la estimación bajo el enfoque bayesiano.

3.2.3. Estimación bayesiana de proporciones

Además de EBLUP, los métodos de estimación e inferencia bayes empírico (EB) y bayes Jerárquico (HB), también se han aplicado a la estimación en áreas pequeñas para proporciones. Bajo la aproximación EB, la estimación e inferencia de la aproximación de bayes se usan en cuanto la distribución

posterior se estima de los datos. Bajo la aproximación HB, los parámetros desconocidos del modelo (incluyendo componentes de varianza) son tratados como aleatorios, con valores extraídos de distribuciones a priori. Entonces, la distribución posterior de las características de área pequeña se obtienen por integración sobre estas a priori, con inferencias basadas en esta distribución posterior. Ghosh y Rao (1994) y Rao (2003) revisaron la aplicación de estos métodos a la estimación de áreas pequeñas. Maiti (1998) usó a prioris no informativas para el hiperparámetro cuando aplicó el método HB. You y Rao (2000) usaron los métodos HB para estimar medias de áreas pequeñas bajo modelos de efectos aleatorios.

Ahora, si los datos son exclusivamente discretos o categóricos, Ghosh et al. (1998), desarrollaron una aproximación general para estimación de áreas pequeñas basada en modelos lineales generalizados. Malec et al. (1999) extendió los modelos descritos por Malec et al. (1997) incluyendo una componente de sobremuestreo en la verosimilitud. Farrel et al. (1997) extendió el modelo mixto logístico de MacGibbon y Tomberlin (1989) a diseños multifase. Farrell (2000), propone una metodología EB para estimar proporciones en áreas pequeñas. La idea consiste en incorporar efectos aleatorios que reflejen la estructura del diseño de la muestra en el modelo de regresión logística. Moura y Migon (2002), desarrollaron una aproximación de modelo jerárquico logístico para la predicción de proporciones de áreas pequeñas, teniendo en cuenta los efectos de heterogeneidad estructurada y espacial, es decir, introducen una componente para considerar la estructura espacialmente correlada de los datos de la respuesta binaria. MacGibbon y Tomberlin (2002), en esta nueva contribución, estiman proporciones mediante EB incorporando efectos aleatorios y efectos aleatorios anidados en modelos que reflejan la estructura compleja de un diseño de muestreo de muchas fases, como fue propuesto originalmente por Dempster y Tomberlin (1980). La metodología propuesta usa en ambos, el modelo de regresión logística múltiple y la técnicas EB, un efecto aleatorio, lo cual produce directamente estimaciones de la incertidumbre asociada con las proporciones estimadas para las áreas pequeñas. Larsen (2003), introduce un método de estimación que más bien es apropiado para estimación en eventos raros. Compara estimaciones basadas en dos modelos, uno que usa conglomerados geográficos simples y uno que usa información más detallada de covariables que relacionan los subconjuntos entre sí. Se emplean técnicas de estimación EB, se seleccionan covariables para el segundo modelo paso a paso hasta agregar otra variable que no produzca una disminución en un criterio objetivo previamente establecido. González-Manteiga et al. (2007) utilizan un modelo mixto logístico para la estimación de proporciones de dominio, aunque el objetivo central del documento es la estimación del error cuadrático medio de estas estimaciones. Xie, et al. (2007), proponen una extensión robusta del

modelo de Fay-Herriot, toda vez que el método de F-H puede comportarse bien globalmente en la estimación de áreas pequeñas, pero es vulnerable a outliers. En el nuevo modelo se asume que los efectos aleatorios de área siguen una distribución t considerando como parámetro desconocido los grados de libertad. Trevisani et al. (2007), hacen una comparación de modelos alternativos HB, cuando las estimaciones de áreas pequeñas consisten en datos de conteo o frecuencias. Los modelos alternativos que probaron fueron el modelo logNormal-Normal, Normal-logNormal, Normal-Poisson-logNormal, Poisson-logNormal y Gamma-Poisson-logNormal. Liu (2009) propuso un modelo para reducir el riesgo de mala especificación del modelo debido a desviaciones significativas del supuesto de normalidad, que usualmente se asume en los efectos aleatorios. La distribución de los efectos aleatorios se selecciona adaptativamente de la distribución de probabilidad de la clase exponencial power y es la riqueza de la clase exponencial power quien asegura la robustez de la aproximación de HB contra la desviación de la normalidad. Liu (2009), en otro contribución, estudia una aproximación de una mejor predicción empírica usando el mismo modelo de su trabajo previo ($Bernoulli - Logit - EP$) para la estimación en áreas pequeñas.

3.2.4. Modelos mixtos lineales generalizados

Describiremos en mayor detalle los modelos mixtos lineales generalizados porque pueden ser especialmente ajustados para datos binarios y de frecuencias.

Modelos de regresión logística

Supongamos que A_{dj} es binaria, es decir, $A_{dj} = 0$ ó 1, y los parámetros de interés son las proporciones de áreas pequeñas $P_{A_d} = \sum_j A_{dj}/N_d$. MacGibbon y Tomberlin (1989), usaron un modelo de regresión logística con efectos aleatorios de área para estimar P_{A_d}. Para la unidad j dentro del área d, A_{dj} toma los valores 1 y 0 con probabilidades P_{dj} y $1 - P_{dj}$, respectivamente, por lo tanto $E[A_{dj}] = P_{dj}$ y $Var[A_{dj}] = P_{dj}(1 - P_{dj})$. Asumamos que las A_{dj} son variables aleatorias independientes Bernoulli (P_{dj}), y los P_{dj} obedecen al siguiente modelo de regresión logística con efectos aleatorios de área v_d:

$$\text{logit}(P_{dj}) = \log \frac{P_{dj}}{1 - P_{dj}} = \mathbf{B}_{dj}^T \boldsymbol{\beta} + v_d, \tag{3.5}$$

o si se escribe en términos de P_{dj} se obtiene

$$P_{dj} = \frac{\exp\{\boldsymbol{\eta}_{dj}\}}{1 + \exp\{\boldsymbol{\eta}_{dj}\}}, \quad \boldsymbol{\eta}_{dj} = \mathbf{B}_{dj}\boldsymbol{\beta} + u_d \tag{3.6}$$

donde $v_d \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ y las $\mathbf{B}_{dj}$ son covariables a nivel de unidad.

El estimador de P_{Ad} bajo el modelo es de la forma

$$\widehat{P}_{Ad} = \frac{1}{N_d}\left(\sum_{j\in s_d} A_{dj} + \sum_{j\in s_d^c} \widehat{P}_{dj}\right),$$

donde s_d^c está formada por los elementos de N_d que o están en s_d, $\widehat{P}_{dj}$ se obtiene de (3.6) estimando $\boldsymbol{\beta}$ y v_d utilizando los métodos BE o BJ.

Malec, Sedransk, Moriarity y LeClere (1997), consideraron un modelo de regresión logística con coeficientes de regresión aleatorios. Suponen que las unidades están agrupadas en j clases en cada área pequeña d, y dados los P_{dj}, los valores $A_{djl}\,(l = 1, \cdots, N_{dj})$ en la (d, j)-ésima celda son variables independientes Bernoulli con probabilidad común P_{dj}. Además, asumen que

$$\theta_{dj} = \text{logit}(P_{dj}) = \mathbf{B}_j^T \boldsymbol{\beta}_d \tag{3.7}$$

y

$$\boldsymbol{\beta}_d = \mathbf{Z}_d \boldsymbol{\alpha} + \mathbf{v}_d \tag{3.8}$$

con $\mathbf{v}_d \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_v)$, donde $\mathbf{Z}_d$ es una matriz de covariables de dimensión $p \times q$ a nivel de área y $\mathbf{B}_j$ es un vector de covariables en cada clase.

Modelo mixto logístico

González-Manteiga, et al. (2007), utilizaron un modelo mixto logístico para la estimación de proporciones en dominio y la estimación de su varianza por métodos de replicación. Describimos en seguida el modelo para la estimación de proporciones.

Se considera una población finita de N unidades $U = \{1, \cdots, N\}$, particionada en D subconjuntos U_d con tamaños $N_d, d = 1, \cdots, D$, donde $N = \sum_{d=1}^{D} N_d$. Para la unidad j dentro del área d, A_{dj} toma los valores 1 y 0 con probabilidad P_{dj} y $1 - P_{dj}$, respectivamente, por lo tanto $E[A_{dj}] = P_{dj}$ y $Var[A_{dj}] = P_{dj}(1 - P_{dj})$.

Tanto la media como la varianza dependen de j, por lo tanto cualquier factor que afecte la esperanza también afectará la varianza. Sea $\mathbf{B}_{dj}$ el vector de dimensión $1 \times p$ que contiene los valores de p atributos auxiliares, $\mathbf{B}_{dp} = (B_{d1}, B_{d2}, \cdots, B_{dp})$. Para el área pequeña d, sea u_d un efecto aleatorio normalmente distribuido con media cero y varianza constante φ. Se asume que $u_1, \cdots, u_D$ son independientes y que, dado u_d, las observaciones A_{dj} son también independientes con distribución

$$A_{dj}|u_d \sim Bin(n_d, P_{dj}), \quad j = 1, \cdots, N_d, \quad d = 1, \cdots, D. \tag{3.9}$$

Denotemos por $\mu_{dj} = n_d P_{dj}$ y $\sigma_{dj} = n_d P_{dj}(1 - P_{dj})$, respectivamente, la media y varianza de A_{dj} dado μ_{dj}. La distribución condicional de A_{dj} pertenece a la familia exponencial, y el parámetro natural es

$$\log\left\{\frac{P_{dj}}{1-P_{dj}}\right\}, \quad j = 1, \cdots, N_d, \quad d = 1, \cdots, D.$$

Si se designa por $\boldsymbol{\eta}_{dj} = \mathbf{B}_{dj}\,\boldsymbol{\beta} + u_d$ el predictor lineal, donde $\boldsymbol{\beta}$ es el vector de p efectos de los p atributos auxiliares y se asume que se conserva el modelo para la población, entonces

$$\log\left\{\frac{P_{dj}}{1-P_{dj}}\right\} = \mathbf{B}_{dj}\,\boldsymbol{\beta} + u_d, \quad j = 1, \cdots, N_d, \quad d = 1, \cdots, D. \tag{3.10}$$

Expresando el modelo en términos de $P_{A_{dj}}$ se obtiene

$$P_{dj} = \frac{\exp\{\boldsymbol{\eta}_{dj}\}}{1+\exp\{\boldsymbol{\eta}_{dj}\}}, \quad \boldsymbol{\eta}_{dj} = \mathbf{B}_{dj}\boldsymbol{\beta} + u_d, \quad j = 1, \cdots, N_d, \quad d = 1, \cdots, D, \tag{3.11}$$

y su estimador será

$$\widehat{P}_{dj} = \frac{\exp\{\widehat{\boldsymbol{\eta}}_{dj}\}}{1+\exp\{\widehat{\boldsymbol{\eta}}_{dj}\}}, \quad \widehat{\boldsymbol{\eta}}_{dj} = \mathbf{B}_{dj}\widehat{\boldsymbol{\beta}}, \quad j = 1, \cdots, N_d, \quad d = 1, \cdots, D, \tag{3.12}$$

Para formular el modelo dado por (3.12) en forma matricial, se asume que $\mathbf{A}$ es el vector de dimensión $N \times 1$ con elementos A_{dj}, $\mathbf{B}$ la matriz de $N \times p$ con filas $\mathbf{B}_{dj}$ y $\mathbf{u} = (u_1, \cdots, u_D)^t$ el vector de $D \times 1$ efectos aleatorios de área. Se define la matriz $\mathbf{Z} = \text{diag}\{\mathbf{1}_{N_d}, d = 1, \cdots, D\}$ de dimensión $N \times D$, donde $\mathbf{1}_a$ el vector columna de unos de tamaño a. Además, $\boldsymbol{\eta}$ es el vector de predictores lineales $\boldsymbol{\eta}_{dj}$, el cual en notación matricial es $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$. Sea $\boldsymbol{\mu} = E[\mathbf{A}|\mathbf{u}]$ el vector de medias condicionales con elementos μ_{dj}, y $\boldsymbol{\Sigma} = Var[\mathbf{A}|\mathbf{u}]$ la matriz de covarianzas condicional, la cual es diagonal con elementos σ_{dj}. El interés es predecir parámetros lineales para áreas pequeñas, es decir, un vector de parámetros del tipo

$$\boldsymbol{\delta} = \mathbf{M}\mathbf{A}, \tag{3.13}$$

donde $\mathbf{M} = \text{diag}\{a_d^t, d = 1, \cdots, D\}$ es una matriz de $D \times N$ y $a_d^t = (a_{d1}, \cdots, a_{dN_d})$ es un vector con elementos conocidos y uniformemente acotados. Un caso particular de $\boldsymbol{\delta}$ es el vector de proporciones de áreas pequeñas $\mathbf{P}_{A_d} = (P_{A_1}, \cdots, P_{A_D})^t$, donde

$$P_{A_d} = N_d^{-1} \sum_{j=1}^{N_d} A_{dj}, \quad d = 1, \cdots, D,$$

para el cual $a_d^t = N_d^{-1}\mathbf{1}_{N_d}^t, \quad d = 1, \cdots, D.$

Seleccionemos una muestra aleatoria $s \subset U$ de tamaño $n < N$ de acuerdo a un diseño de muestreo. Designemos por $r = U - s$ el conjunto de unidades no incluidas en la muestra. Sea $s_d = s \cap U_d$ el conjunto de unidades muestrales extraídas del área d, con tamaño $n_d,\ d = 1, \cdots, D$, donde $n = \sum_{d=1}^{D} n_d$. Además, sea $r_d = r \cap U_d$ el conjunto de unidades de U_d que no caen en la muestra, con tamaño $N_d - n_d,\ d = 1, \cdots, D$. Reordenando las observaciones de acuerdo a su pertenencia a s o r, se definen las particiones

$$\mathbf{M} = [\ \mathbf{M}_s \quad \mathbf{M}_r\ ],\ \mathbf{A} = \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_r \end{bmatrix},\ \mathbf{B} = \begin{bmatrix} \mathbf{B}_s \\ \mathbf{B}_r \end{bmatrix},\ \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{bmatrix},\ \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_s \\ \boldsymbol{\eta}_r \end{bmatrix},$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_r \end{bmatrix},\ \boldsymbol{\Sigma} = \text{diag}[\ \boldsymbol{\Sigma}_s \quad \boldsymbol{\Sigma}_r\ ].$$

Entonces, el parámetro lineal que deseamos estimar puede escribirse como

$$\boldsymbol{\delta} = P_{A_d} = \mathbf{M}_s\mathbf{A}_s + \mathbf{M}_r\mathbf{A}_r. \tag{3.14}$$

El único elemento desconocido involucrado en (3.14) es $\mathbf{A}_r$, el cual puede predecirse ajustando el modelo; es decir, obteniendo un predictor lineal $\widehat{\eta}_{dj} = \mathbf{B}_{dj}\widehat{\boldsymbol{\beta}} + \widehat{u}_d$ y usando como predicción de cada elemento A_{dj} de $\mathbf{A}_r$, la predicción de su valor esperado

$$\widehat{\mu}_{dj} = \widehat{P}_{dj} = \frac{\exp\{\widehat{\eta}_{dj}\}}{1 + \exp\{\widehat{\eta}_{dj}\}}.$$

Denotemos por $\widehat{\mu}_r$ el vector de valores predichos $\widehat{\mu}_{dj}$ correspondientes a las observaciones no muestreadas. Entonces, el predictor de $\boldsymbol{\delta}$ está dado por

$$\widehat{\boldsymbol{\delta}} = \widehat{P}_{A_d} = \mathbf{M}_s\mathbf{A}_s + \mathbf{M}_r\widehat{\mu}_r. \tag{3.15}$$

Cuando φ es conocida, la maximización de la densidad conjunta de $\mathbf{A}_s$ y $\mathbf{u}$ conduce a un sistema de ecuaciones que proporciona un estimador de cuasi-verosimilitud penalizada (PQL) de $\boldsymbol{\beta}$ y un predictor PQL de $\mathbf{u}$ (Breslow y Clayton, 1993). Denotemos por $\widehat{\boldsymbol{\beta}}$ el estimador PQL de $\boldsymbol{\beta}$, y por $\widehat{\mathbf{u}} = (\widehat{u}_1, \cdots, \widehat{u}_D)^t$ el predictor PQL de $\mathbf{u} = (u_1, \cdots, u_D)^t$. Pueden calcularse vía un algoritmo scorin de Fisher directo o via el mismo algoritmo, pero en la forma de mínimos cuadrados ponderados iterados (IWLS). De la misma forma, denotemos por $\widehat{\eta}_{dj} = \mathbf{B}_{dj}\widehat{\boldsymbol{\beta}} + \widehat{u}_d$ el predictor de η_{dj}, $\widehat{P}_{dj}$ el predictor de P_{dj} obtenido reemplazando $\widehat{\eta}_{dj}$ en (3.12), y finalmente $\widehat{\sigma}_{dj} = \widehat{P}_{A_{dj}}(1 - \widehat{P}_{A_{dj}})$ el predictor de σ_{dj}.

En el caso de φ desconocida, el algoritmo propuesto por Schall (1991) proporciona estimadores de $\boldsymbol{\beta}$ y φ, y un predictor de $\mathbf{u}$. Para describir sus

resultados brevemente, se reescribe el modelo (3.10) en la forma de un modelo lineal generalizado, es decir,

$$g_{dj}(\mu_{dj}) = \eta_{dj} \quad \text{con} \quad g_{dj}(\mu_{dj}) = \log\left\{\frac{\mu_{dj}}{1-\mu_{dj}}\right\}, \tag{3.16}$$

para $j = 1, \cdots, N_d$ y $d = 1, \cdots, D$. Considere la expansión de Taylor de primer orden de $g_{dj}(A_{dj})$, en torno a μ_{dj},

$$g_{dj}(A_{dj}) \cong \eta_{dj} + (A_{dj} - \mu_{dj})g'_{dj}(\mu_{dj}) \triangleq \xi_{dj}.$$

Los momentos condicionales de las variables transformadas ξ_{dj} están dados por

$$E[\xi_{dj}|\mathbf{u}] = \eta_{dj}, \quad Var[\xi_{dj}|\mathbf{u}] = g'_{dj}(\mu_{dj})^2\sigma_{dj},$$

y

$$Cov[\xi_{dj}, \xi_{d'j'}|\mathbf{u}] = 0 \quad \text{para} \quad d \neq d' \quad \text{o} \quad j \neq j'.$$

Además, la media condicional de ξ_{dj} es $\mathbf{B}_{dj}^t\boldsymbol{\beta}$, y los efectos aleatorios u_d son independientes, normalmente distribuidos con media cero y varianza constante igual a φ. Ahora, podemos construir el modelo lineal

$$\xi_{dj} = \mathbf{B}_{dj}^t\boldsymbol{\beta} + u_d + e_{dj} \quad j = 1, \cdots, N_d, \quad d = 1, \cdots, D, \tag{3.17}$$

siendo e_{dj} variables aleatorias independientes, independientes de $\mathbf{u}$, con media cero y varianza $v_{dj} = g'(\mu_{dj})^2\sigma_{dj}$. El modelo (3.17) es una aproximación lineal del modelo (3.16). Denotemos por ξ_s el vector de observaciones transformadas ξ_{dj}, con $j \in s_d$, $d = 1, \cdots, D$. Entonces se cumple que

$$Var[\xi_s] = \varphi\mathbf{Z}_s\mathbf{Z}_s^t + \boldsymbol{\Sigma}_{es} \triangleq \mathbf{V}_s, \tag{3.18}$$

donde $\boldsymbol{\Sigma}_{es}$ es una matriz diagonal cuyos elementos son las varianzas v_{dj} de los residuos e_{dj}.

Observemos que si $\mathbf{V}_s$ es conocida, entonces el mejor estimador lineal insesgado (BLUE) de $\boldsymbol{\beta}$, y el mejor predictor lineal insesgado (BLUP) de $\mathbf{u}$ en (3.17) estarían dados por las fórmulas

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_s^t\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^t\mathbf{V}_s^{-1}\boldsymbol{\xi}_s, \quad \widehat{\mathbf{u}} = \varphi\mathbf{Z}_s^t\mathbf{V}_s^{-1}(\boldsymbol{\xi}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}}). \tag{3.19}$$

Desgraciadamente, $\mathbf{V}_s$ es desconocida, dado que depende de la varianza del error condicional v_{dj}, la cual depende del predictor lineal $\eta_{dj} = \mathbf{B}_{dj}\boldsymbol{\beta} + u_d$, es decir $\mathbf{V}_s = \mathbf{V}_s(\boldsymbol{\beta}, \mathbf{u})$ para un valor dado de φ. Reemplazando $\mathbf{V}_s$ por su predicción $\widehat{\mathbf{V}}_s = \mathbf{V}_s(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}})$ en (3.19), las ecuaciones obtenidas son las ecuaciones IWLS para obtener el estimador PQL de $\boldsymbol{\beta}$ y el predictor PQL de $\mathbf{u}$ del modelo

(3.16) vía el algoritmo scoring de Fisher. Estas ecuaciones se obtienen sin el supuesto de normalidad de los errores e_{dj}. Además, $\mathbf{V}_s$ también depende de φ y así este parámetro debe estimarse.

Siguiendo a Schall (1991), es fácil ver que la ecuación de máxima verosimilitud (MV) para φ, obtenida maximizando la densidad marginal de $\boldsymbol{\xi}_s$ está dada por

$$\varphi = (D-\nu)^{-1}\sum_{d=1}^{D} u_d^2 \quad \text{para} \quad \nu = \sum_{d=1}^{D}(1+\varphi\sigma_{di\cdot})^{-1}, \tag{3.20}$$

donde $\sigma_{dj\cdot} = \sum_{j\in s_d}\sigma_{dj}$.

Modelos para mortalidad y tasas de enfermedad

La mortalidad y las tasas de enfermedad de áreas pequeñas en una región o un país se usan con frecuencia para construir mapas de enfermedad, como por ejemplo, mapas del cáncer. Tales mapas se usan para visualizar la variabilidad geográfica de una enfermedad e identificar áreas de proporciones altas, garantizando así la intervención. Un modelo simple de área pequeña se obtiene asumiendo que las frecuencias observadas de área pequeña, $y_d = \sum_{U_d} A_j$, son variables independientes Poisson con media condicional $E(y_d|\lambda_d) = n_d\lambda_d$ y que $\lambda_d \overset{\text{iid}}{\sim} \text{gamma}(\alpha,\nu)$. Aquí λ_d y n_d son la tasa verdadera y el número de expuestos en el área d, y (α,ν) son los parámetros de escala y forma de la distribución gamma. Bajo este modelo, se obtienen estimadores suavizados de λ_d usando los métodos de BE o BJ (Clayton y Kaldor, 1987; Datta, Ghosh y Waller, 2000). También se han propuesto modelos espaciales autoregresivos condicionales (CAR) de la forma (2.90) sobre el log de la razón $\theta_d = \log(\lambda_d)$ (Klayton y Kaldor, 1987). El modelo sobre λ_d puede extenderse para incorporar covariables de nivel de área $\mathbf{z}_d$, por ejemplo, $\theta_d = \mathbf{z}_d^T\boldsymbol{\beta} + v_d$ con $v_d \overset{\text{iid}}{\sim} N(0,\sigma_v^2)$. Nandram, Sedransk y Pickle (1999), estudiaron los modelos de regresión sobre el log de las tasas de edad específica $\theta_{dj} = \log\lambda_{dj}$ incluyendo pendientes aleatorias, donde j denota edad. Pueden modelarse también tasas de mortalidad conjunta de las frecuencias observadas en dos sitios diferentes (y_{1d}, y_{2d}), asumiendo que las (y_{1d}, y_{2d}) son condicional e independientemente ditribuídas sobre $(\lambda_{1d},\lambda_{2d})$ y $\theta_d = (\log\lambda_{1d}, \log\lambda_{2d})^T \overset{\text{iid}}{\sim} N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Además, y_{1d} y y_{2d} se asumen como variables de Poisson condicionalmente independientes con $E(y_{1d}|\lambda_{1d}) = n_{1d}\lambda_{1d}$ y $E(y_{2d}|\lambda_{2d}) = n_{2d}\lambda_{2d}$. Como un ejemplo de este modelo bivariado, y_{1d} y y_{2d} denotan el número de muertes debidas al cáncer en los sitios 1 y 2 y (n_{1d}, n_{2d}) la población en riesgo en los sitios 1 y 2. DeSouza (1992), mostró que el modelo bivariado conduce a estimadores mejorados de las tasas $(\lambda_{1d},\lambda_{2d})$ comparado a estimadores basados en modelos univariantes separados.

Modelos de la familia exponencial

Ghosh, Natarajan, Stroud y Carlin (1998), propusieron modelos lineales generalizados con efectos aleatorios de área. Los estadísticos muestrales de las observaciones en el dominio $A_{dj}(j = 1, \cdots, n_d; d = 1, \cdots, D)$ condicionales sobre los θ_{dj}, se asumen independientemente distribuidos con función de densidad de probabilidad que pertenece a la familia exponencial con parámetros canónicos θ_{dj}, es decir,

$$f(A_{dj}|\theta_{dj}) = \exp\left[\frac{1}{\phi_{dj}}(\theta_{dj}A_{dj} - a(\theta_{dj})) + b(A_{dj}, \phi_{dj})\right] \tag{3.21}$$

para $\phi_{dj}(> 0)$ y funciones conocidas $a(\cdot)$ y $b(\cdot)$. La familia exponencial (3.21) cubre distribuciones conocidas incluyendo las distribuciones normal, binomial y Poisson. Por ejemplo, $\theta_{dj} = \text{logit}(p_{dj})$ y $\phi_{dj} = 1$ si A_{dj} es Binomial (n_{dj}, p_{dj}), y $\theta_{dj} = \log(\lambda_{dj})$ y $\phi_{dj} = 1$ si A_{dj} es Poisson (λ_{dj}). Los θ_{dj} son modelados como

$$\theta_{dj} = \mathbf{B}_{dj}^T\boldsymbol{\beta} + v_d + u_{dj}, \tag{3.22}$$

donde v_d y u_{dj} son mutuamente independientes con $v_d \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ y $u_{dj} \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$.

El objetivo aquí es hacer inferencias sobre los parámetros θ_{dj} del área pequeña. Por ejemplo, $\theta_{dj} = \text{logit}(p_{dj})$ y p_{dj} denota la proporción asociada con la variable binaria en la j-ésima categoría de edad y sexo en la región d.

Ghosh, Natarajan, Waller y Kim (1999), extendieron el modelo de enlace (3.22) para manejar datos espaciales, y aplicaron el modelo del mapa de la enfermedad.

Modelos semiparamétricos

También se han propuesto modelos semiparamétricos basados solamente en la especificación de los dos primeros momentos de las respuestas A_{dj}, condicionadas sobre la media de área pequeñas μ_d, y de los $\mu_d's$. En la ausencia de covariables, Ghosh y Lahiri (1987), asumen el siguiente modelo:

- Para cada d, condicionada sobre los θ_d, los A_{dj} son iid con media θ_d y varianza $\mu_2(\theta_{dj})$, denotada $A_{dj}|\theta_d \overset{\text{iid}}{\sim}(\theta_d, \mu_2(\theta_d)), j = 1, \cdots, N_d; d = 1, \cdots, D$;
- $\theta_d \overset{\text{iid}}{\sim}(\mu, \sigma_v^2)$;
- $0 < \sigma_e^2 = E\mu_2(\theta_i) < \infty$.

Raghunathan (1993), incorporó información de covariables de nivel de área $\mathbf{z}_d$ como sigue:

- condicionada sobre θ_d, $A_{dj} \overset{\text{iid}}{\sim}(\theta_d, b_1(\phi, \theta_d, a_{dj}))$ donde $b(\cdot)$ es una función positiva conocida de un parámetro de "dispersión" ϕ, medias de área pequeña θ_d y constantes conocidas a_{dj};

- $\theta_d \overset{\text{ind}}{\sim}(\tau_d = h(\mathbf{z}_d^T\boldsymbol{\beta}), b_2(\psi, \tau_d, a_d))$ donde $h(\cdot)$ es una función conocida y $b_2(\cdot)$ es una función positiva conocida de un parámetro de "dispersión" ψ, la media τ_d y una constante conocida a_d.

El modelo "longitudinal" (2.108) con covariables de nivel unidad puede ser generalizado fijando

$$E(A_{dj}|\mathbf{v}_d) = \mu_{dj}, \quad V(A_{dj}|\mathbf{v}_d) = \phi b(\mu_{dj}) \tag{3.23}$$

y

$$h(\mu_{dj}) = \mathbf{B}_{dj1}^T\boldsymbol{\beta} + \mathbf{B}_{dj2}^T\mathbf{v}_d, \quad \mathbf{v}_d \overset{\text{iid}}{\sim}(\mathbf{0}, \boldsymbol{\Sigma}_v); \tag{3.24}$$

es decir, $\mathbf{v}_d$ son independientes e idénticamente distribuidas con media $\mathbf{0}$ y matriz de covarianzas $\boldsymbol{\Sigma}_v$ (Breslow y Claiton, 1993).

3.2.5. Modelos especialmente diseñados para datos binarios

En esta subsección describiremos modelos de nivel unidad, que han sido pensados especialmente para respuestas binarias, A_{dj}, es decir, $A_{dj} = 1$ o 0. En este caso, los modelos lineales mixtos no son apropiados y se han propuesto modelos alternativos. Si todas la covariables $\mathbf{B}_{dj}$ asociadas con A_{dj} son de área específica, es decir $\mathbf{B}_{dj} = \mathbf{B}_d$, entonces podemos transformar las proporciones de la muestra de área, $\widehat{P}_{A_d} = \sum_j A_{dj}/n_d = y_d/n_d$, usando una transformación arco seno y el modelo se reduce a un modelo de nivel área. No obstante, los estimadores transformados resultantes, $\widehat{\theta}_d$, pueden no satisfacer el modelo de muestreo con errores medios de muestreo cero si n_d es pequeña.

Modelar la unidad de nivel evita las dificultades anteriores, y pueden obtenerse estimadores BE de proporciones directamente del caso general de covariables de unidad específica.

Modelos sin covariables

En este caso se asume un modelo de dos fases sobre las observaciones muestrales A_{dj}, $j = 1, \cdots, n_d$; $d = 1, \cdots, D$. En la primera fase, asumimos que $A_{dj}|p_d \overset{\text{iid}}{\sim} \text{Bernoulli}(p_d)$, $d = 1, \cdots, D$. En la segunda fase se asume un modelo

que liga los p_d; en particular, $p_d \overset{\text{iid}}{\sim} \text{beta}(\alpha, \beta)$; $\alpha > 0,\ \beta > 0$ donde $\text{beta}(\alpha, \beta)$ denota la distribución beta con parámetros α y β:

$$f(p_d|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_d^{\alpha-1}(1 - p_d)^{\beta-1}; \quad \alpha > 0,\ \beta > 0 \tag{3.25}$$

donde $\Gamma(\cdot)$ es la función gamma. Reducimos el vector de frecuencias observadas $\mathbf{y}_d = (y_{d1}, \cdots, y_{dn_d})^T$, al total de frecuencias observadas en la muestra de dominio $y_d = \sum_j A_{dj}$, notando que y_d es un estadístico minimal suficiente para el modelo de la primera fase.

Nótese que $y_d|p_d \overset{\text{ind}}{\sim} \text{Binomial}(n_d, p_d)$, es decir,

$$f(y_d|p_d) = \left(\begin{array}{c} n_d \\ y_d \end{array} \right) p_d^{y_d}(1 - p_d)^{n_d - y_d}. \tag{3.26}$$

Se sigue de (3.25) y (3.26) que $p_d|y_d, \alpha, \beta \overset{\text{ind}}{\sim} \text{beta}(y_d + \alpha, n_d - y_d + \beta)$. Por lo tanto, el estimador de bayes de p_d y la varianza a posteriori de p_d están dadas por

$$\widehat{p}_d^{\,B}(\alpha, \beta) = E(p_d|y_d, \alpha, \beta) = (y_d + \alpha)/(n_d + \alpha + \beta) \tag{3.27}$$

y

$$V(p_d|y_d, \alpha, \beta) = \frac{(y_d + \alpha)(n_d - y_d + \beta)}{(n_d + \alpha + \beta + 1)(n_d + \alpha + \beta)^2}. \tag{3.28}$$

Nótese que la distribución de enlace, $f(p_d|\alpha, \beta)$, es una "a priori conjugada" en el sentido que la distribución a posteriori, $f(p_d|y_d, \alpha, \beta)$, tendrá la misma forma que la distribución a priori.

Obtenemos estimadores de los parámetros del modelo de la distribución marginal: $y_d|\alpha, \beta \overset{\text{ind}}{\sim}$ Beta-binomial dada por

$$f(y_d|\alpha, \beta) = \left(\begin{array}{c} n_d \\ y_d \end{array} \right) \frac{\Gamma(\alpha + y_d)\Gamma(\beta + n_d - y_d)}{\Gamma(\alpha + \beta + n_d)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}. \tag{3.29}$$

Los estimadores de máxima verosimilitud (MV), $\widehat{\alpha}_{\text{MV}}$ y $\widehat{\beta}_{\text{MV}}$ pueden obtenerse maximizando la log-verosimilitud:

$$l(\alpha, \beta) = \text{const} + \sum_{i=1}^{D} \left[\sum_{h=0}^{y_d - 1} \log(\alpha + h) \right.$$

$$\left. + \sum_{h=0}^{n_d - y_d - 1} \log(\beta + h) - \sum_{h=0}^{n_d - 1} \log(\alpha + \beta + h) \right], \tag{3.30}$$

donde $\sum_{h=0}^{y_d-1} \log(\alpha + h)$ se toma como cero si $y_d = 0$ y $\sum_{h=0}^{n_d - y_d - 1} \log(\beta + h)$ se toma como cero si $y_d = n_d$. Una representación conveniente es en términos de

la media $E(A_{dj}) = \mu = \alpha/(\alpha+\beta)$ y $\tau = 1/(\alpha+\beta)$ la cual está relacionada con la correlación intraclase $\rho = \text{Corr}(A_{dj} - A_{dk}) = 1/(\alpha+\beta+1)$ para $j \neq k$. Usando μ y τ, (3.30) toma la forma

$$l(\mu,\tau) = \text{const} + \sum_{i=1}^{D}\left[\sum_{h=0}^{y_d-1}\log(\mu+h\tau)\right.$$

$$\left. + \sum_{h=0}^{n_d-y_d-1}\log(1-\mu+h\tau) - \sum_{h=0}^{n_d-1}\log(1+h\tau)\right]. \tag{3.31}$$

Las expresiones explícitas para $\widehat{\alpha}_{\text{MV}}$ y $\widehat{\beta}_{\text{MV}}$ (ó $\widehat{\mu}_{\text{MV}}$ y $\widehat{\tau}_{\text{MV}}$) no existen, pero pueden obtenerse estimaciones MV por el método de Newton-Raphson o algún otro método iterativo.

Podemos usar también estimadores simples por el método de momentos de α y β. Igualamos la proporción muestral ponderada $p = \sum_d (n_d/n_T)\widehat{p}_d$ y la varianza muestral ponderada $s_p^2 = \sum_d (n_d/n_T)(\widehat{p}_d - \widehat{p})^2$ a sus valores esperados y resolvemos las ecuaciones de momentos resultantes para α y β, donde $n_T = \sum_d n_d$. Esto lleva a estimadores de momentos, $\widehat{\alpha}$ y $\widehat{\beta}$, dados por

$$\widehat{\alpha}/(\widehat{\alpha}+\widehat{\beta}) = \widehat{p} \tag{3.32}$$

y

$$\frac{1}{\widehat{\alpha}+\widehat{\beta}+1} = \frac{n_T s_p^2 - \widehat{p}(1-\widehat{p})(D-1)}{\widehat{p}(1-\widehat{p})[n_T - \sum_d n_d^2/n_T - (D-1)]}; \tag{3.33}$$

ver Kleinman (1973). Nótese que los estimadores de momentos no son únicos, a diferencia de los estimadores MV.

Sustituimos los estimadores de momentos $\widehat{\alpha}$ y $\widehat{\beta}$ en (3.27) para obtener un estimador BE de p_d como

$$\widehat{p}_d^{BE} = \widehat{p}_d^{B}(\widehat{\alpha},\widehat{\beta}) = \widehat{\gamma}_d \widehat{p}_d + (1-\widehat{\gamma}_d)\widehat{p}, \tag{3.34}$$

donde $\widehat{\gamma}_d = n_d/(n_d+\widehat{\alpha}+\widehat{\beta})$. Observe que $\widehat{p}_d^{BE}$ es un estimador combinado del estimador directo $\widehat{p}_d$ y el estimador sintético $\widehat{p}$ y se da más peso a $\widehat{p}_d$ conforme el tamaño de muestra de área, n_d, se incrementa. Por lo tanto es similar al estimador de Fay-Herriot para el modelo de nivel básico de área, pero el peso $\widehat{\gamma}_d$ evita el supuesto de varianza de muestreo conocida de $\widehat{p}_d$. El estimador $\widehat{p}_d^{BE}$ es aproximadamente insesgado para p_d en el sentido que su sesgo, $E(\widehat{p}_d^{BE} - p_d)$, es de orden D^{-1}, para D grande. Una aproximación BE sencilla usa $\widehat{p}_d^{BE}$ como el estimador de p_d, y su variabilidad se mide por la varianza a posteriori estimada $V(p_d|y_d,\widehat{\alpha},\widehat{\beta}) = g_{1i}(\widehat{\alpha},\widehat{\beta},y_d)$. Sin embargo, la varianza a posteriori estimada,

$g_{1i}(\widehat{\alpha}, \widehat{\beta}, y_d)$, puede conducir a una severa subestimación del ECM($\widehat{p}_d^{BE}$) porque ignora la variabilidad asociada con $\widehat{\alpha}$ y $\widehat{\beta}$.

Se han propuesto también modelos alternativos de dos fases. El modelo de primera fase no se cambia, pero el modelo de segunda fase se cambia a cualquier (i) $\text{logit}(p_d) = \log[p_d/(1-p_d)] \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ o (ii) $\Phi^{-1}(p_d) \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, donde $\Phi(\cdot)$ es la función de distribución acumulada (FDA) de una variable $N(0,1)$. Los modelos (i) y (ii) se llaman modelos logit-normal y probit-normal, respectivamente. La implementación de BE es más complicada para los modelos alternativos porque no existen expresiones explícitas para el estimador de bayes y la varianza a posteriori de p_d.

Para el modelo logit-normal, el estimador de bayes de p_d puede expresarse como una razón de integrales unidimensionales. Escribiendo p_d como $p_d = h_1(\mu + \sigma z_d)$, donde $h_1(a) = e^a/(1+e^a)$ y $z_d \sim N(0,1)$, obtenemos $\widehat{p}_d^B(\mu, \sigma) = E(p_d|y_d, \mu, \sigma)$ de la distribución condicional de z_d dada y_d. Tenemos

$$\widehat{p}_d^B(\mu, \sigma) = A(y_d, \mu, \sigma)/B(y_d, \mu, \sigma), \tag{3.35}$$

donde

$$A(y_d, \mu, \sigma) = E[h_1(\mu + \sigma z)\exp\{h_2(y_d, \mu + \sigma z)\}] \tag{3.36}$$

y

$$B(y_d, \mu, \sigma) = E[\exp\{h_2(y_d, \mu + \sigma z)\}], \tag{3.37}$$

donde $h_2(y_d, a) = a y_d - n_d \log(1 + a^a)$ y la esperanza es sobre $z \sim N(0,1)$; ver McCulloch y Searle (2001, p. 67). Podemos evaluar (3.36) y (3.37) simulando muestras de una $N(0,1)$. Alternativamente puede usarse, como esbozamos antes, integración numérica.

La log verosimilitud, $l(\mu, \sigma)$, para el modelo logit-normal puede escribirse como

$$l(\mu, \sigma) = \text{const} + \sum_{i=1}^{D} \log[B(y_d, \mu, \sigma)], \tag{3.38}$$

donde $B(y_d, \mu, \sigma)$ está dada por (3.37). Las derivadas de $l(\mu, \sigma)$, necesitarán métodos tipo Newton-Raphson para calcular estimaciones MV, y pueden ser aproximadas de una manera similar.

Usando estimadores MV de $\widehat{\mu}$ y $\widehat{\sigma}$, obtenemos un estimador BE de p_d como $\widehat{p}_d^{BE} = \widehat{p}_d^B(\widehat{\mu}, \widehat{\sigma})$. La varianza a posteriori, $V(p_d|y_d, \mu, \sigma)$, también puede expresarse en términos de la esperanza sobre $z \sim N(0,1)$, notando que $V(p_d|y_d, \mu, \sigma) = E(p_d^2|y_d, \mu, \sigma) - [\widehat{p}_d^B(\mu, \sigma)]^2$. Denotando $v(p_d|y_d, \mu, \sigma) = g_{1i}(\mu, \sigma, y_d)$, puede aplicarse el método jackknife para obtener un estimador aproximadamente insesgado del ECM($\widehat{p}_d^{BE}$).

Jiang y Lahiri (2001) usaron $\widehat{p}_d^{BE}$ basado en los estimadores de momentos del primer paso, y obtuvieron un estimador por desarrollo de Taylor del

ECM($\widehat{p}_d^{BE}$), similar a los estimadores del ECM de segundo orden del modelo lineal mixto. Este estimador es aproximadamente insesgado, como en el caso del estimador jackknife del ECM.

Modelos con covariables

El modelo logit-normal de la subsección anterior se extiende sin dificultad al caso de covariables. En la primera fase, asumimos que $A_{dj}|p_{dj} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{dj})$ para $j = 1, \cdots, N_d;\ d = 1, \cdots, D$. Los p_{dj} están ligados en la segunda etapa asumiendo un modelo de regresión logística con efectos de área aleatorios: $\text{logit}(p_{dj}) = \mathbf{B}_{dj}^T\boldsymbol{\beta} + v_d$ donde $v_d \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ y $\mathbf{B}_{dj}$ es el vector de covariables fijas. El modelo de dos variables se llama modelo logístico lineal mixto. Pertenece a la clase de los modelos lineales generalizados mixtos.

Los parámetros de área pequeña son las proporciones de área pequeña $P_d = \sum_j A_{dj}/N_d$. Como en el caso del modelo básico de nivel unidad, asumimos que los modelos se conservan para la muestra $\{(A_{dj}, \mathbf{B}_{dj}), j \in s_d; d = 1, \cdots, D\}$, donde s_d es la muestra de tamaño n_d para el área d.

Expresamos P_d como $P_d = f_d p_d + (1 - f_d)p_d^*$, donde $f_d = n_d/N_d$, p_d es la proporción muestral y $p_d^* = \sum_{l \in \overline{s}_d} A_{dl}/(N_d - n_d)$ es la proporción de las unidades no muestreadas $\overline{s}_d$ en el área d. Ahora, notando que $E(A_{dl}|p_{dl}, \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v) = p_{dl}$ para $l \in \overline{s}_d$, el estimador de bayes de p_d^* está dada por $\widehat{p}_{d(c)}^B = E(p_{d(c)}|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v)$, donde $\widehat{p}_{d(c)} = \sum_{l \in \overline{s}_d} p_{dl}/(N_d - n_d)$ y $\mathbf{y}_d$ es el vector de valores A_j de la muestra del área d. Por lo tanto, el estimador de bayes de P_d puede expresarse como

$$\widehat{P}_d^B = \widehat{P}_d^B(\boldsymbol{\beta}, \boldsymbol{\sigma}_v) = f_d p_d + (1 - f_d)p_{d(c)}^B. \tag{3.39}$$

Si la fracción de muestreo es despreciable, podemos expresar $\widehat{P}_d^B$ como

$$\widehat{P}_d^B \approx \frac{1}{N_d} E\left(\sum_{l=1}^{N_d} p_{dl}|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v\right). \tag{3.40}$$

La varianza a posteriori de P_d se reduce a

$$V(P_d|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v) = (1 - f_d)^2 E(\overline{y}_d^* - \widehat{p}_{d(c)}^B)^2$$

$$= N_d^{-2}\left[E\left\{\sum_{l \in \overline{s}_d} p_{dl}(1 - p_{dl})|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v\right\} + V\left\{\sum_{l \in \overline{s}_d} p_{dl}|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v\right\}\right]; \tag{3.41}$$

ver Malec et al. (1997). Nótese que (3.41) incluye esperanzas de la forma $E(\sum_l p_{dl}^2|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v)$ y $E[(\sum_l p_{dl})^2|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v]$ así como $E(\sum_l p_{dl}|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v) = (N_d - n_d)\widehat{p}_{dc}^B$. No existe ninguna expresión explícita para estas esperanzas. Sin embargo, podemos expresar las esperanzas como razones de integrales de una sola dimensión, similares a (3.35).

La estimación por máxima verosimilitud de los parámetros del modelo, $\boldsymbol{\beta}$ y σ_v, para el modelo logístico lineal mixto y otros modelos lineales generalizados mixtos, ha recibido considerable atención en años recientes. Los métodos propuestos incluyen cuadratura numérica, algoritmo EM, Monte Carlo mediante cadenas de Markov (MCMC) y aproximación estocástica. Para detalles de los algoritmos, consultar McCulloch y Searle (2001). También se han propuesto métodos más simples, cuasi-verosimilitud penalizada (PQL), basada en maximizar la distribución conjunta de $\mathbf{y} = (\mathrm{y}_1^T, \cdots, \mathrm{y}_D^T)^T$ y $\mathbf{v} = (v_1, \cdots, v_D)^T$ con respecto a $\boldsymbol{\beta}$ y $\mathbf{v}$.

Farrel, MacGibbon y Tomberlin (1997a), usaron una medida bootstrap de precisión similar al bootstrap paramétrico de Laird y Louis (1987) (también llamado bootstrap tipo III). Los resultados de simulación, basados en $D = 20$ y $n_d = 50$, indicaron un buen funcionamiento del método BE propuesto. Farrell, MacGibbon y Tomberlin (1997b), relajaron el supuesto de normalidad sobre las v_d, usando el siguiente modelo de enlace: $\text{logit}(p_{dj}) = \mathbf{B}_{dj}^T\boldsymbol{\beta} + v_d$ y $v_d \overset{\text{iid}}{\sim}$ distribución no especificada. Usaron un método MV no paramétrico, propuesto por Laird (1978), para obtener un estimador BE de P_d. Usaron también el bootstrap tipo II de Laird y Louis (1987), para obtener una medida bootstrap de precisión.

3.3. Estimación basada en el diseño.

3.3.1. Estimación de proporciones en dominios

Bajo este apartado consideraremos las adaptaciones de los métodos directo, sintético y combinado, señalados en Rao (2003), a la estimación de proporciones en dominios, los propuestos por Särndal et al. (1992) y los modelos *LGREG* propuestos por Lehtonen y Veijanen (1998). La estimación directa de proporciones de dominio, se hace sin dificultad a partir del estimador directo de la media poblacional, es decir

$$\widehat{P} = N_d^{-1} \sum_{s_d} y_j$$

o en su forma expandida

$$\widehat{P} = N_d^{-1} \sum_{s_d} w_j y_j$$

donde el atributo de interés $y_j = A_j$, w_j son las probabilidades de inclusión de primer orden y N_d se asume conocido, toda vez que si no lo es, el estimador de la proporción no sería lineal y habría que estimarla usando un estimador de razón de totales de H-T. La estimación se hace bajo el supuesto de que se conservan las características de la población en la muestra de dominio y puede extenderse a diseños muestrales más complejos, como muestreo estratificado y estratificado multietápico.

El estimador de regresión generalizado dado por (2.11), que hace uso de información auxiliar en forma de totales poblacionales, puede adaptarse también a la estimación de proporciones, aunque la justificación teórica del uso de estimadores de regresión para estimar proporciones es conceptualmente difícil, toda vez que el método se ha diseñado para variables continuas, pero siempre que tenga significado la estimación de la media o la asociación entre variables cualitativas, pueden aplicarse. Särndal et al. (1992) hace un desarrollo de los estimadores de regresión para dominios, pero para variables continuas. En los trabajos recientes de Rueda, M. et al. (2010), si bien no está pensado para la estimación en dominios, se adapta el modelo GREG a la estimación de proporciones, proponiendo un estimador de regresión para el cual se establecen propiedades teóricas y condiciones de optimalidad. Si el modelo GREG se escribe en la forma expandida (2.13), en el cual se cambian los pesos iniciales por nuevos pesos y se optimiza bajo ciertas restricciones de calibración, Rueda, M. et al. (2009) obtienen un estimador de calibración para proporciones. Si se dispone de una sola variables auxiliar, (2.11) se reduce al estimador de razón (2.18), que también ha sido aplicado a la estimación de proporciones por Rueda M. et al. (2008), además, justificaron también la estimación de proporciones de un atributo relacionado a múltiples atributos auxiliares.

Si tomamos información auxiliar de otro dominio u otro tiempo, podemos hacer la estimación de proporciones mediante la estimación directa modificada. Es posible también la estimación sintética de proporciones, usando las expresiones para la media sintética (2.29) si no disponemos de información auxiliar y el estimador de regresión sintético (2.30). A partir de (2.30) se pueden derivar, el estimador de razón sintético, si se cuenta con una sola variable auxiliar y un estimador post-estratificado sintético dado por (2.31). Los estimadores combinados también pueden aplicarse a la estimación de proporciones, que es una forma natural de balancear el sesgo potencial de la estimación sintética, frente a la inestabilidad de un estimador directo, tomando un estimador combinado de los pesos del par de estimadores. Si usamos los pesos comunes $\phi_i = \phi$ en una estimación compuesta, tenemos los estimadores de James-Stein, que pueden usarse en la estimación de proporciones.

3.3.2. Estimación en áreas pequeñas por regresión logística

Un modelo que puede clasificarse en un enfoque bajo el diseño, fue desarrollado por Lehtonen y Veijanen (1998), para la estimación de frecuencias de clase de una variable respuesta discreta. Propusieron un estimador de regresión generalizado (GREG) para el total $\widehat{T}_{GREG}$, que combina los valores predichos $\widehat{A}_j = \widehat{P}r(A_j = 1|\pi_j)$ basado en un modelo apropiado y el estimador de H-T de los residuos $e_j = A_j - \widehat{A}_j$ de las unidades muestreadas,

$$\widehat{T}_{GREG} = \sum_{j=1}^{N} \widehat{A}_j + \sum_{j \in s} e_j/\pi_j, \tag{3.42}$$

si el modelo de probabilidad apropiado es el de regresión logística,

$$\widehat{A}_j = \frac{exp(\widehat{\eta}_j)}{1 + exp(\widehat{\eta}_j)} \tag{3.43}$$

con $\widehat{\eta}_j = \mathbf{B}_j^T \widehat{\beta}$ y $\widehat{\beta}$ es el estimador de las pendientes comunes, obtenidas por mínimos cuadrados ordinarios, entonces el modelo se denomina LGREG.

Duchesne (2003) usó el estimador LGREG en la construcción de un estimador para proporciones, dividiendo por el tamaño poblacional conocido N, es decir

$$\widehat{P}_{LGREG} = \frac{1}{N} \left\{ \sum_{j=1}^{N} \widehat{A}_j + \sum_{j \in s} e_j/\pi_j \right\}. \tag{3.44}$$

Modelo de regresión logística a nivel de dominio

Ahora, bajo la formulación del modelo con intercepto específico de dominio tenemos en (3.43) que $\mathbf{B} = (I_{1j}, \cdots, I_{Dj}, B_{1j}, \cdots, B_{pj})$ conocido para todo $j \in U$, donde $I_{dj} = 1$ si $j \in U_d, I_{dj} = 0$ en otro caso. $\boldsymbol{\beta} = (\beta_{01}, \cdots, \beta_{0D}, \beta_1, \cdots, \beta_p)'$, donde β_{0D} son interceptos específicos de dominio y β_j son pendientes comunes y $d = 1 \cdots, D$.

Los valores ajustados para $\widehat{A}_j$ en el modelo de regresión logística (3.43), se calculan para todos los $j \in U$ y se sustituyen en (3.42). Entonces el modelo para el estimador GREG logístico de totales de dominio, es

$$\widehat{T}_{d_{LGREG}} = \sum_{j \in U_d} \widehat{A}_j + \sum_{j \in s_d} \left(\frac{A_j - \widehat{A}_j}{\pi_j} \right) \tag{3.45}$$

La estimación de la proporción de individuos que poseen el atributo A a nivel de dominio, se obtiene dividiendo por N_d el total de dominio dado por (3.45), es decir

$$\widehat{P}_{d_{LGREG}} = \frac{1}{N_d} \left\{ \sum_{j \in U_d} \widehat{A}_j + \sum_{j \in s_d} e_j / \pi_j \right\}, \tag{3.46}$$

donde si N_d es desconocido, debe estimarse.

Una expresión para la varianza estimada del estimador de la proporción estará dada por

$$\widehat{AV}(\widehat{P}_{d_{LGREG}}) = \frac{1}{\widehat{N}_d^2} \sum \sum_{j,k \in s_d} \frac{\Delta_{jk}}{\pi_{jk}} \frac{e_j}{\pi_j} \frac{e_k}{\pi_k}. \tag{3.47}$$

Capítulo 4

Aportaciones a la estimación de proporciones

Como se desprende de los capítulos precedentes, se han propuesto distintos métodos para la estimación en áreas pequeñas. Algunos de estos métodos de estimación, bajo el diseño o bajo el modelo, hacen uso de información auxiliar en la fase de estimación, tomando información de otros dominios adyacentes o similares, o bien de registros administrativos, censos o estudios previos. En las últimas décadas se ha privilegiado el uso de modelos que relacionan la variable de interés con la información auxiliar disponible, mediante modelos jerárquicos de bayes. Estos métodos demandan el uso de aproximaciones computacionales como métodos de Monte Carlo mediante cadenas de Markov, muestreo de Gibbs, integración numérica, Monte Carlo Newton-Raphson, métodos iterativos para la estimación de la varianza, etc., además de información completa. Como hemos afirmado antes, en la práctica no es fácil disponer de información completa, por lo que los métodos basados en modelos podrían no ser factibles. Es bajo este escenario que proponemos nuestros estimadores.

Como es usual con variables categóricas o de frecuencias, el objetivo es estimar la proporción poblacional de individuos que poseen el atributo A, dentro del dominio d, es decir, $P_{A_d} = N_d^{-1} \sum_{j=1}^{U_d} A_{dj}$, si disponemos de un vector de información auxiliar del atributo B, en forma de totales o proporciones, a nivel poblacional o a nivel de dominio, que se supone relacionado con el atributo de interés A.

Sea U_d nuestro dominio de interés, deseamos estimar la proporción P_{A_d} del atributo A a nivel de dominio, a partir de la muestra aleatoria s bajo un diseño de muestreo prefijado. Entonces, la proporción de individuos que poseen el atributo A en el dominio estará dada por

$$P_{A_d} = N_d^{-1} \sum_{j \in U} X_{dj} A_{dj} = N_d^{-1} \sum_{j \in U_d} X_j A_j = N_d^{-1} \sum_{j \in s_d} A_j \tag{4.1}$$

y la proporción de individuos que poseen el atributo B será

$$P_{B_d} = N_d^{-1} \sum_{j \in U} X_{dj} B_{dj} = N_d^{-1} \sum_{j \in U_d} X_j B_j = N_d^{-1} \sum_{j \in s_d} B_j \tag{4.2}$$

donde N_d es desconocido en la práctica.

De la expresión (3.2) se deduce que N_d, el tamaño del dominio, es igual a $T(X_{dj})$, entonces el estimador de Horvitz-Thompson de este total será

$$\widehat{N}_d = \sum_{j \in s} \frac{X_{dj}}{\pi_j} = \sum_{j \in s_d} \frac{1}{\pi_j}, \tag{4.3}$$

su varianza

$$V[\widehat{N}_d] = \sum_{k,l \in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) X_k X_l$$

y un estimador de la varianza

$$\widehat{V}[\widehat{N}_d] = \sum_{k,l \in s} \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) X_k X_l \tag{4.4}$$

Para el caso de un diseño $MAS(N, n)$ el estimador dado por (4.3) será

$$\widehat{N}_d = n_d \frac{N}{n} = N \frac{n_d}{n}$$

y el estimador de la varianza en (4.4) estará dado por

$$\widehat{V}[\widehat{N}_d] = N^2 \frac{1-f}{n-1} \frac{n_d}{n} \left(1 - \frac{n_d}{n} \right).$$

Por lo anterior, reescribiendo (4.1) para P_{A_d} tendremos

$$P_{A_d} = \frac{T(A_{dj})}{T(X_{dj})} = \frac{T_1}{T_2} = \frac{\sum_{j \in U_d} A_{dj}}{\sum_{j \in U_d} X_{dj}} = \frac{\sum_{j \in U_d} A_j}{N_d} \tag{4.5}$$

que corresponde a un parámetro no lineal en la forma de una razón simple de totales. Aplicando la estimación de Horvitz-Thompson, su estimador será

$$\widehat{P}_{A_d} = \frac{\sum_s A_{dj}/\pi_j}{\sum_s X_{dj}/\pi_j} = \frac{\sum_{s_d} A_j/\pi_j}{\sum_{s_d} 1/\pi_j} = \frac{\sum_{s_d} A_j/\pi_j}{\widehat{N}_d}. \tag{4.6}$$

Una expresión para la varianza se obtiene de forma conjunta de la siguiente manera:

sabemos que una estimación de esta razón será

$$\widehat{P}_{A_d} = \frac{\widehat{T}_1}{\widehat{T}_2}$$

por lo que la varianza aproximada será obtenida considerando la variable auxiliar E_j definida como

$$E_j = \frac{1}{T_2}(A_j - P_{A_d}X_j) \text{ y } e_j = \frac{E_j}{\pi_j}.$$

Se tiene que

$$\widehat{R}_0 = P_{A_d} + \frac{1}{T_2}\sum_{j\in s}\frac{(A_j - P_{A_d}X_j)}{\pi_j} = P_{A_d} + \sum_{j\in s} e_j$$

entonces,

$$AV[\widehat{P}_{A_d}] \approx V[\widehat{R}_0] = V\left[\sum_{j\in s} e_j\right]$$

por lo que la expresión para la varianza será

$$AV[\widehat{P}_{A_d}] = \frac{1}{(T_2)^2}\sum\sum_{kl\in U}\Delta_{kl}\frac{(A_k - P_{A_d}X_k)}{\pi_k}\frac{(A_l - P_{A_d}X_l)}{\pi_l} \tag{4.7}$$

y su estimador

$$\widehat{AV}[\widehat{P}_{A_d}] = \frac{1}{(\widehat{T}_2)^2}\sum\sum_{kl\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{(A_k - \widehat{P}_{A_d})}{\pi_k}\frac{(A_l - \widehat{P}_{A_d})}{\pi_l} \tag{4.8}$$

donde $\widehat{T}_2 = \widehat{N}_d$.

Bajo un diseño $MAS(N,n)$, $\widehat{N}_d = N\frac{n_d}{n}$, por lo que la ecuación (4.6) se escribe como

$$\widehat{P}_{A_d} = \frac{\sum_{s_d} A_j/\pi_j}{\widehat{N}_d} = \frac{\sum_{s_d} A_j}{n_d} = p_{A_d}. \tag{4.9}$$

La varianza aproximada dada por (4.7) para $\widehat{P}_{A_d}$ será

$$AV[\widehat{P}_{A_d}] = \frac{1}{N_d^2}N^2\frac{1-f}{n}\frac{1}{N-1}\sum_{j\in U}(A_j - P_{A_d}X_j)^2$$

pero $S^2_{A_d} = \frac{1}{N_d-1}\sum_{j\in U_d}(A_j - P_{A_d}X_j)^2$, donde $P_{A_d}X_j = P_{A_d}$, por estar en el dominio. Entonces, de la expresión anterior se obtiene

$$AV[\widehat{P}_{A_d}] = \frac{N^2}{N_d^2}\frac{1-f}{n}\frac{N_d-1}{N-1}S^2_{A_d} = \frac{N}{N_d}\frac{1-f}{n}P_{A_d}Q_{A_d}$$

ya que $S^2_{A_d} = \frac{N_d}{(N_d-1)} P_{A_d} Q_{A_d}$. Entonces

$$AV[\widehat{P}_{A_d}] = \frac{N}{N_d} \frac{1-f}{n} P_{A_d} Q_{A_d}, \tag{4.10}$$

y el estimador de la varianza dado por (4.8)

$$\widehat{AV}[\widehat{P}_{A_d}] = \frac{N^2}{(\widehat{N}_d)^2} \frac{1-f}{n} \frac{1}{n-1} \sum_{j \in s} (A_j - \widehat{P}_{A_d})^2,$$

$$\widehat{AV}[\widehat{P}_{A_d}] = \frac{N^2}{(N\frac{n_d}{n})^2} \frac{1-f}{n} \frac{1}{n-1} \sum_{j \in s} (A_j - p_{A_d})^2,$$

$$\widehat{AV}[\widehat{P}_{A_d}] = \frac{n^2}{n_d^2} \frac{1-f}{n} \frac{1}{n-1} \sum_{j \in s} (A_j - p_{A_d})^2,$$

$$\widehat{AV}[\widehat{P}_{A_d}] = \frac{1}{n_d} \frac{1-f}{n_d} \frac{n}{n-1} \sum_{j \in s_d} (A_j - p_{A_d})^2,$$

ahora, si designamos por

$$s^2_{A_d} = \frac{1}{n_d - 1} \sum_{j \in s_d} (A_j - p_{A_d})^2$$

la cuasivarianza muestral calculada sobre los elementos que en la muestra pertenecen a la subpoblación, entonces la varianza estimada queda como

$$\widehat{AV}[\widehat{P}_{A_d}] = \frac{n_d - 1}{n_d} \frac{1-f}{n_d} \frac{n}{n-1} s^2_{A_d} = \frac{1-f}{n_d} \frac{n}{n-1} p_{A_d} q_{A_d} \tag{4.11}$$

donde $(n_d - 1)s^2_{A_d} = n_d p_{A_d} q_{A_d}$, p_{A_d} es la proporción de individuos que poseen el atributo A en la submuestra s_d y $q_{A_d} = 1 - p_{A_d}$.

Ahora, si la muestra es suficientemente grande de forma que $n/(n-1) \approx 1$, entonces (4.11) se puede aproximar por

$$\widehat{AV}[\widehat{P}_{A_d}] = \frac{1-f}{n_d} p_{A_d} q_{A_d}. \tag{4.12}$$

4.1. Estimadores de razón propuestos

Los estimadores tipo razón, que incluyen información auxiliar relacionada a la variable de interés, pretenden mejorar la presición de un estimador simple. Estos estimadores han demostrado, en ciertas condiciones, ser mejores que las estimaciones directas sin información auxiliar. En esta sección proponemos algunos estimadores de razón para la proporción de un atributo de interés de una subpoblación o dominio, en los cuales se cuenta con información auxiliar en forma de totales a nivel poblacional o a nivel de dominio. Consideraremos inicialmente estimadores simples y posteriormente los correspondientes estimadores combinados de estos estimadores. Se hace a continuación la presentación individual de cada uno de ellos.

4.1.1. Información auxiliar en dominios: se conoce P_{B_d}

Asumamos conocido $P_B^{U_d} = P_{B_d}$, que es la proporción del atributo B a nivel de dominio. Bajo este supuesto el estimador de razón para P_{A_d} estará dado por

$$\widehat{P}_{rA_d}^{(1)} = \widehat{R}_d P_{B_d} = \frac{\widehat{P}_{A_d}}{\widehat{P}_{B_d}} P_{B_d} \tag{4.13}$$

donde $R_d = \frac{P_{A_d}}{P_{B_d}}$, $\widehat{P}_{A_d} = \widehat{N}_d^{-1} \sum_{j \in s_d} A_j$, $\widehat{P}_{B_d} = \widehat{N}_d^{-1} \sum_{j \in s_d} B_j$ y $P_{A_d} = N_d^{-1} \sum_{j \in U_d} B_j$.

Mediante la técnica de linealización de Taylor podemos formular una expresión para la varianza del estimador. La aproximación lineal para $\widehat{R}_d$ es

$$\widehat{R}_d = R_d + \frac{1}{P_{B_d}} (\widehat{P}_{A_d} - R_d \widehat{P}_{B_d})$$

de donde se deduce que

$$V(\widehat{R}_d) = V\left[R_d + \frac{1}{P_{B_d}} (\widehat{P}_{A_d} - R_d \widehat{P}_{B_d}) \right] = \frac{1}{P_{B_d}^2} V(\widehat{P}_{A_d} - R_d \widehat{P}_{B_d})$$

$$= \frac{1}{P_{B_d}^2} \left[V(\widehat{P}_{A_d}) + R_d^2 V(\widehat{P}_{B_d}) - 2R_d cov(\widehat{P}_{A_d}, \widehat{P}_{B_d}) \right]$$

Por lo tanto, la varianza aproximada del estimador de $\widehat{P}_{rA_d}^{(1)}$ será

$$AV(\widehat{P}_{rA_d}^{(1)}) = V[\widehat{R}_d P_{B_d}] = V\left[\left(R_d + \frac{1}{P_{B_d}} (\widehat{P}_{A_d} - R_d \widehat{P}_{B_d}) \right) P_{B_d} \right]$$

$$= V(\widehat{P}_{A_d} - R_d \widehat{P}_{B_d})$$

$$AV(\widehat{P}_{rA_d}^{(1)}) = V(\widehat{P}_{A_d}) + R_d^2 V(\widehat{P}_{B_d}) - 2R_d cov(\widehat{P}_{A_d}, \widehat{P}_{B_d}). \tag{4.14}$$

El estimador $\widehat{P}_{rA_d}^{(1)}$ es sesgado y su sesgo aproximado puede obtenerse usando un resultado dado por Sampath, S., (2001, pág. 98), a partir del cual una expresión de este sesgo viene dada por

$$B(\widehat{P}_{rA_d}^{(1)}) = P_{A_d}\left[\frac{V(\widehat{P}_{B_d})}{P_{B_d}^2} - \frac{cov(\widehat{P}_{A_d}, \widehat{P}_{B_d})}{P_{A_d}P_{B_d}}\right]. \tag{4.15}$$

Entonces, bajo un diseño $MAS(N, n)$, el estimador de su varianza estará dado por

$$\widehat{AV}(\widehat{P}_{rA_d}^{(1)}) = \frac{1-f}{n_d}\frac{n}{n-1}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}_d^2\widehat{P}_{B_d}\widehat{Q}_{B_d} - 2\widehat{R}_d\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\right] \tag{4.16}$$

y una aproximación para el sesgo estimado

$$\widehat{B}(\widehat{P}_{rA_d}^{(1)}) = \frac{1-f}{n_d}\frac{n}{n-1}\left[\widehat{R}_d\widehat{Q}_{B_d} - \widehat{\phi}_d\frac{\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}}{\widehat{P}_{B_d}}\right], \tag{4.17}$$

donde $\widehat{P}_{A_d} = p_{A_d}$, la proporción de individuos de la muestra que están en el dominio y que poseen el atributo A, $\widehat{Q}_{A_d} = 1 - p_{A_d}$, $\widehat{P}_{B_d} = p_{B_d}$, $\widehat{Q}_{B_d} = 1 - p_{B_d}$ y $\widehat{\phi}_d = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$, es el estimador del coeficiente V de Cramer a partir de la tabla de doble entrada formada por los datos de los atributos A, B y sus complementos en la submuestra del dominio s_d (ver apéndice A).

4.1.2. Información auxiliar poblacional: se conoce P_B

La proporción de individuos P_{A_d} que poseen el atributo A, a nivel de dominio, asumiendo conocida la proporción poblacional de individuos que poseen el atributo B, se puede estimar usando un estimador de razón de la forma

$$\widehat{P}_{rA_d}^{(2)} = \widehat{R}_d P_B, \tag{4.18}$$

donde $\widehat{R}_d = \widehat{P}_{A_d}/\widehat{P}_{B_d}$ es el estimador de la razón a nivel de dominio, $R_d = P_{A_d}/P_{B_d}$, $\widehat{P}_{A_d} = \widehat{N}_d^{-1}\sum_{j\in s_d} A_j$, $\widehat{P}_{B_d} = \widehat{N}_d^{-1}\sum_{j\in s_d} B_j$ y $P_B = N^{-1}\sum_{j=1}^{N} B_j$.

El estimador de razón $\widehat{R}_d = \frac{\widehat{P}_{A_d}}{\widehat{P}_{B_d}}$, tiene como aproximación lineal de Taylor

$$\widehat{R}_d = R_d + \frac{1}{P_{B_d}}(\widehat{P}_{A_d} - R_d\widehat{P}_{B_d}), \tag{4.19}$$

y su varianza aproximada, deducida de la aproximación de Taylor (4.19), es como sigue

$$AV(\widehat{R}_d) = V\left[R_d + \frac{1}{P_{B_d}}(\widehat{P}_{A_d} - R_d\widehat{P}_{B_d})\right] = \frac{1}{P_{B_d}^2}V[\widehat{P}_{A_d} - R_d\widehat{P}_{B_d}]$$

$$= \frac{1}{P_{B_d}^2}\left[V(\widehat{P}_{A_d}) + R_d^2V(\widehat{P}_{B_d}) - 2R_dcov(\widehat{P}_{A_d}, \widehat{P}_{B_d})\right].$$

Entonces, la varianza aproximada de $\widehat{P}_{rA_d}^{(1)}$ será,

$$AV(\widehat{P}_{rA_d}^{(2)}) = V\left[\left(R_d + \frac{1}{P_{B_d}}(\widehat{P}_{A_d} - R_d\widehat{P}_{B_d})\right)P_B\right] = \left(\frac{P_B}{P_{B_d}}\right)^2 V(\widehat{P}_{A_d} - R_d\widehat{P}_{B_d})$$

$$AV(\widehat{P}_{rA_d}^{(2)}) = \left(\frac{P_B}{P_{B_d}}\right)^2\left[V(\widehat{P}_{A_d}) + R_d^2V(\widehat{P}_{B_d}) - 2R_dcov(\widehat{P}_{A_d}, \widehat{P}_{B_d})\right], \quad (4.20)$$

cuyo estimador será

$$\widehat{AV}(\widehat{P}_{rA_d}^{(2)}) = \left(\frac{P_B}{P_{B_d}}\right)^2\left[\widehat{V}(\widehat{P}_{A_d}) + \widehat{R}_d^2\widehat{V}(\widehat{P}_{B_d}) - 2\widehat{R}_d\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_d})\right]. \quad (4.21)$$

Bajo un diseño $MAS(N, n)$, la varianza dada por (4.20) será

$$AV(\widehat{P}_{rA_d}^{(2)}) =$$

$$\left(\frac{P_B}{P_{B_d}}\right)^2\frac{N}{N_d}\frac{1-f}{n}\left[P_{A_d}Q_{A_d} + R_d^2P_{B_d}Q_{B_d} + 2R_d\phi_d\sqrt{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}\right], \quad (4.22)$$

donde $Q_{A_d} = 1 - P_{A_d}$, $Q_{B_d} = 1 - P_{B_d}$ y $\phi_d = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}}$, es el coeficiente V de Cramer basado en una tabla de doble entrada, formada por los datos de los atributos A y B en el dominio, con sus respectivos complementos y, $cov(\widehat{P}_{A_d}, \widehat{P}_{B_d}) = \phi_d\sqrt{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}$.

El estimador de la varianza será

$$\widehat{AV}(\widehat{P}_{rA_d}^{(2)}) =$$

$$\left(\frac{P_B}{P_{B_d}}\right)^2\frac{1-f}{n_d}\frac{n}{n-1}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}_d^2\widehat{P}_{B_d}\widehat{Q}_{B_d} - 2\widehat{R}_d\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\right] \quad (4.23)$$

si se cumple que $n/(n-1) \approx 1$, podemos usar como aproximación

$$\widehat{AV}(\widehat{P}_{rA_d}^{(2)}) = \left(\frac{P_B}{P_{B_d}}\right)^2\frac{1-f}{n_d}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}_d^2\widehat{P}_{B_d}\widehat{Q}_{B_d} - 2\widehat{R}_d\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\right]. \quad (4.24)$$

Su sesgo aproximado

$$B(\widehat{P}^{(2)}_{rA_d}) = P_{A_d}\left(\frac{P_B}{P_{B_d}}\right)^2\left[\frac{V(\widehat{P}_{B_d})}{P^2_{B_d}} - \frac{cov(\widehat{P}_{A_d},\widehat{P}_{B_d})}{P_{A_d}P_{B_d}}\right]$$

$$B(\widehat{P}^{(2)}_{rA_d}) = \left(\frac{P_B}{P_{B_d}}\right)^2\left[R_d V(\widehat{P}_{B_d}) - \frac{cov(\widehat{P}_{A_d},\widehat{P}_{B_d})}{P_{B_d}}\right]$$

bajo un diseño $MAS(N,n)$, será

$$B(\widehat{P}^{(2)}_{rA_d}) = \frac{N}{N_d}\frac{1-f}{n}\left(\frac{P_B}{P_{B_d}}\right)^2\left[P_{A_d}Q_{B_d} - \frac{\phi_d\sqrt{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}}{P_{B_d}}\right]$$

y su estimador

$$\widehat{B}(\widehat{P}^{(2)}_{rA_d}) = \frac{1-f}{n_d}\frac{n}{n-1}\left(\frac{P_B}{P_{B_d}}\right)^2\left[\widehat{P}_{A_d}\widehat{Q}_{B_d} - \frac{\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}}{\widehat{P}_{B_d}}\right], \quad (4.25)$$

donde $\widehat{P}_{A_d} = p_{A_d}$, la proporción de individuos de la muestra que están en el dominio y que poseen el atributo A, $\widehat{Q}_{A_d} = 1 - p_{A_d}$, $\widehat{P}_{B_d} = p_{B_d}$, $\widehat{Q}_{B_d} = 1 - p_{B_d}$ y $\widehat{\phi}_d = \frac{n_{11}n_{22}-n_{12}n_{21}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$, es el estimador del coeficiente V de Cramer a partir de la tabla de doble entrada formada por los datos de los atributos A, B y sus complementos en la submuestra del dominio s_d (ver apéndice A).

4.1.3. Estimador de razón sintético

En este caso suponemos que el estimador de razón $\widehat{R} = \frac{\widehat{P}_A}{\widehat{P}_B}$, válido para la población, puede usarse para estimar la proporción P_{A_d} correspondiente al dominio d, si se asume conocida información auxiliar en forma de totales de dominio, es decir, se conoce P_{B_d}. El estimador propuesto es

$$\widehat{P}^{(3)}_{rA_d} = \widehat{R}P_{B_d} = \frac{\widehat{P}_A}{\widehat{P}_B}P_{B_d} \quad (4.26)$$

donde $R = \frac{P_A}{P_B}$, $\widehat{P}_A = n^{-1}\sum_{j=1}^n A_j$, $\widehat{P}_B = n^{-1}\sum_{j=1}^n B_j$ y $P_{B_d} = N_d^{-1}\sum_{j\in U_d} B_j$.

La aproximación lineal de Taylor para $\widehat{R}$ es

$$\widehat{R} = R + \frac{1}{P_B}(\widehat{P}_A - R\widehat{P}_B)$$

de la cual se deduce que

$$VA(\widehat{R}) = \frac{1}{P_B^2} V(\widehat{P}_A - R\widehat{P}_B) = \frac{1}{P_B^2}\left[V(\widehat{P}_A) + R^2 V(\widehat{P}_B) - 2Rcov(\widehat{P}_A, \widehat{P}_B)\right].$$

Ahora, una aproximación para la varianza de $\widehat{P}^{(3)}_{rA_d}$ será

$$VA(\widehat{P}^{(3)}_{rA_d}) = \left(\frac{P_{B_d}}{P_B}\right)^2 \left[V(\widehat{P}_A) + R^2 V(\widehat{P}_B) - 2Rcov(\widehat{P}_A, \widehat{P}_B)\right] \tag{4.27}$$

y su estimador

$$\widehat{VA}(\widehat{P}^{(3)}_{rA_d}) = \left(\frac{P_{B_d}}{P_B}\right)^2 \left[\widehat{V}(\widehat{P}_A) + \widehat{R}^2 V(\widehat{P}_B) - 2\widehat{R}\widehat{cov}(\widehat{P}_A, \widehat{P}_B)\right] \tag{4.28}$$

Una aproximación para el sesgo

$$B(\widehat{P}^{(3)}_{rA_d}) = P_A \left(\frac{P_{B_d}}{P_B}\right)^2 \left[\frac{V(\widehat{P}_B)}{P_B^2} - \frac{cov(\widehat{P}_A, \widehat{P}_B)}{P_A P_B}\right] \tag{4.29}$$

y su estimador

$$\widehat{B}(\widehat{P}^{(3)}_{rA_d}) = \widehat{P}_A \left(\frac{P_{B_d}}{P_B}\right)^2 \left[\frac{\widehat{V}(\widehat{P}_B)}{\widehat{P}_B^2} - \frac{\widehat{cov}(\widehat{P}_A, \widehat{P}_B)}{\widehat{P}_A \widehat{P}_B}\right] \tag{4.30}$$

Bajo un diseño muestral $MAS(N, n)$ podemos escribir (4.27) como

$$VA(\widehat{P}^{(3)}_{rA_d}) = \frac{N-n}{(N-1)n}\left(\frac{P_{B_d}}{P_B}\right)^2 \left[P_A Q_A + R^2 P_B Q_B - 2R\phi\sqrt{P_A Q_A P_B Q_B}\,\right] \tag{4.31}$$

y su estimador

$$\widehat{VA}(\widehat{P}^{(3)}_{rA_d}) = \frac{1-f}{n-1}\left(\frac{P_{B_d}}{P_B}\right)^2 \left[\widehat{P}_A \widehat{Q}_A + \widehat{R}^2 \widehat{P}_B \widehat{Q}_B - 2\widehat{R}\widehat{\phi}\sqrt{\widehat{P}_A \widehat{Q}_A \widehat{P}_B \widehat{Q}_B}\,\right]. \tag{4.32}$$

El sesgo dado por (4.29) será

$$B(\widehat{P}^{(3)}_{rA_d}) = \frac{N-n}{(N-1)n}\left(\frac{P_{B_d}}{P_B}\right)^2 \left[RQ_B - \phi\frac{\sqrt{P_A Q_A P_B Q_B}}{P_B}\right] \tag{4.33}$$

y el sesgo estimado

$$\widehat{B}(\widehat{P}^{(3)}_{rA_d}) = \frac{1-f}{n-1}\left(\frac{P_{B_d}}{P_B}\right)^2 \left[\widehat{R}\widehat{Q}_B - \widehat{\phi}\frac{\sqrt{\widehat{P}_A \widehat{Q}_A \widehat{P}_B \widehat{Q}_B}}{\widehat{P}_B}\right], \tag{4.34}$$

donde $\widehat{P}_A = p_A$, la proporción muestral de individuos que poseen el atributo A en la muestra s, $\widehat{Q}_A = 1 - p_A$, $\widehat{P}_B = p_B$ y $\widehat{Q}_B = 1 - p_B$. Sabemos que $\widehat{\phi}$ es el estimador del coeficiente V de Cramer, aunque en este caso habrá que notar que se calcula de forma similar a la indicada antes, pero con los datos de la muestra aleatoria de tamaño n.

4.1.4. Extensión de los estimadores de razón a un diseño general de muestreo

Consideremos el caso en el cual una muestra aleatoria s se selecciona de acuerdo a un diseño general de muestreo, con probabilidades de inclusión de primero y segundo orden, π_j y π_{jk}, respectivamente.

Sabemos por (4.6) que en ausencia de información auxiliar, el estimador de la proporción de individuos que poseen el atributo A en el dominio d, es $\widehat{P}_{A_d} = \widehat{N}_d^{-1} \sum_{s_d} A_j/\pi_j$. Usando información auxiliar hemos definido el estimador de razón dado por (4.18) o (4.13). Usando las propiedades de los estimadores de H-T, podemos reescribir la expresión para la varianza como

$$AV(\widehat{P}_{rA_d}) = V\left(\sum_{j\in U_d} \frac{A_j - R_d B_j}{\pi_j}\right),$$

$$AV(\widehat{P}_{rA_d}) = \frac{1}{N_d^2} \sum_{j,k\in U_d} (\pi_j\pi_k - \pi_{jk}) \left(\frac{E_j}{\pi_j}\right)\left(\frac{E_k}{\pi_k}\right)$$

y su estimador

$$\widehat{AV}(\widehat{P}_{rA_d}) = \frac{1}{\widehat{N}_d^2} \sum_{j,k\in s_d} \left(\frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(\frac{e_j}{\pi_j}\right)\left(\frac{e_k}{\pi_k}\right), \tag{4.35}$$

con $E_j = A_j - R_d B_j$ y $e_j = A_j - \widehat{R}_d B_j$.

4.1.5. Estimadores combinados de razón

Para balancear el sesgo potencial en los estimadores de razón que hemos desarrollado, proponemos cinco estimadores combinados de razón expresados como una combinación lineal, con pesos ponderados, de los estimadores $\widehat{P}_{A_d}$, $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(2)}_{rA_d}$ y $\widehat{P}^{(3)}_{rA_d}$. Los estimadores que se pueden derivar son

$$\widehat{P}^{(c_1)}_{rA_d} = \alpha\widehat{P}_{A_d} + (1-\alpha)\widehat{P}^{(1)}_{rA_d}, \tag{4.36}$$

$$\widehat{P}^{(c_2)}_{rA_d} = \alpha\widehat{P}_{A_d} + (1-\alpha)\widehat{P}^{(2)}_{rA_d}, \tag{4.37}$$

$$\widehat{P}^{(c_3)}_{rA_d} = \alpha\widehat{P}_{A_d} + (1-\alpha)\widehat{P}^{(3)}_{rA_d}, \tag{4.38}$$

$$\widehat{p}^{(c_4)}_{rA_d} = \alpha\widehat{P}^{(1)}_{rA_d} + (1-\alpha)\widehat{P}^{(3)}_{rA_d}, \tag{4.39}$$

$$\widehat{P}^{(c_5)}_{rA_d} = \alpha\widehat{P}^{(2)}_{rA_d} + (1-\alpha)\widehat{P}^{(3)}_{rA_d}, \tag{4.40}$$

donde $0 \leq \alpha \leq 1$.

Habrá que estimar el valor para α en el sentido de mínima varianza dentro de la clase de estimadores de $\widehat{P}^{(c_k)}_{rA_d}$, $k = 1, \cdots, 5$.

El estimador es de la forma

$$\widehat{P}^{(c_k)}_{rA_d} = \alpha\widehat{t_1} + (1-\alpha)\widehat{t_2}$$

donde $\widehat{t_1}$ representa el primer estimador y $\widehat{t_2}$ el segundo. Entonces, debemos hallar el valor de α que minimiza la varianza de $\widehat{P}^{(c_k)}_{rA_d}$. Esta varianza estará dada por

$$AV(\widehat{P}^{(c_k)}_{rA_d}) = V\left[\alpha\widehat{t_1} + (1-\alpha)\widehat{t_2}\right]$$

$$= \alpha^2 V(\widehat{t_1}) + (1-\alpha)^2 V(\widehat{t_2}) + 2\alpha(1-\alpha)cov(\widehat{t_1}, \widehat{t_2}).$$

Designemos por $V_1 = V(\widehat{t_1})$, $V_2 = V(\widehat{t_2})$ y $C = cov(\widehat{t_1}, \widehat{t_2})$. La varianza de $\widehat{P}^{(c_k)}_{rA_d}$ puede expresarse como

$$AV(\widehat{P}^{(c_k)}_{rA_d}) = \alpha^2 V_1 + (1-\alpha)^2 V_2 + 2\alpha(1-\alpha)C. \tag{4.41}$$

Minimizando esta expresión de la varianza obtenemos

$$\frac{\partial AV(\widehat{P}^{(c_k)}_{rA_d})}{\partial\alpha} = 2\alpha V_1 - 2(1-\alpha)V_2 + 2C - 4\alpha C,$$

igualando a cero y resolviendo para α se obtiene

$$2\alpha V_1 - 2(1-\alpha)V_2 + 2C - 4\alpha C = \alpha V_1 - V_2 + \alpha V_2 + C - 2\alpha C = 0$$

$$\alpha(V_1 + V_2 - 2C) = V_2 - C,$$

$$\alpha_{opt} = \frac{V_2 - C}{V_1 + V_2 - 2C} \tag{4.42}$$

y es un mínimo toda vez que

$$\frac{\partial^2 AV(\widehat{P}_{rA_d}^{(c_d)})}{\partial^2 \alpha} = 2V_1 + 2V_2 - 4C = 2(V_1 + V_2 - 2C) = 2V(\widehat{t}_1 - \widehat{t}_2) > 0,$$

y concluimos que α minimiza la expresión para la $AV(\widehat{P}_{rA_d}^{(c_k)})$, por lo que el estimador puede escribirse como

$$\widehat{P}_{rA_d}^{(c_k)} = \alpha_{opt}\widehat{t}_1 + (1 - \alpha_{opt})\widehat{t}_2.$$

El la práctica $\widehat{P}_{rA_d}^{(c_k)}$ no se conoce, dado que α_{opt} depende de las varianzas poblacionales, las cuales son generalmente desconocidas. En este caso, podemos usar el estimador

$$\widehat{P}_{rA_d}^{(c_k)} = \widehat{\alpha}_{opt}\widehat{t}_1 + (1 - \widehat{\alpha}_{opt})\widehat{t}_2 \tag{4.43}$$

donde

$$\widehat{\alpha}_{opt} = \frac{\widehat{V}(\widehat{t}_2) - \widehat{cov}(\widehat{t}_1, \widehat{t}_2)}{\widehat{V}(\widehat{t}_1) + \widehat{V}(\widehat{t}_2) - 2\widehat{cov}(\widehat{t}_1, \widehat{t}_2)}. \tag{4.44}$$

Operando algebraicamente en la expresión para la varianza de $\widehat{P}_{rA_d}^{(c_k)}$ en (4.41), esta puede ser expresada como

$$AV(\widehat{P}_{rA_d}^{(c_k)}) = \frac{V_1 V_2 - C^2}{V_1 + V_2 - 2C} \tag{4.45}$$

donde V_1 y V_2 son las expresiones para la varianza de $\widehat{t}_1$ y $\widehat{t}_2$, respectivamente y $C = cov(\widehat{t}_1, \widehat{t}_2)$, generalmente desconocida, que debe estimarse. A partir de esta última expresión, un estimador de la varianza del estimador óptimo puede escribirse como

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_k)}) = \frac{\widehat{V}(\widehat{t}_1)\widehat{V}(\widehat{t}_2) - \widehat{cov}^2(\widehat{t}_1, \widehat{t}_2)}{\widehat{V}(\widehat{t}_1) + \widehat{V}(\widehat{t}_2) - 2\widehat{cov}(\widehat{t}_1, \widehat{t}_2)} \tag{4.46}$$

Si asumimos que la covarianza C en (4.44), (4.45) y (4.46) es pequeña, entonces la aproximación óptima para los pesos α_{opt} depende solamente de la razón de varianzas, por lo que (4.44) puede escribirse como

$$\alpha_{opt} = \frac{1}{1 + F} \tag{4.47}$$

donde $F = \frac{V(\widehat{t}_1)}{V(\widehat{t}_2)}$.

Además, la varianza de $\widehat{P}_{rA_d}^{(c_k)}$ con pesos óptimos α_{opt} se reduce a

$$AV(\widehat{P}_{rA_d}^{(c_k)}) = \frac{V_1 V_2}{V_1 + V_2}$$

y su estimador

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_k)}) = \frac{\widehat{V}(\widehat{t}_1)\widehat{V}(\widehat{t}_2)}{\widehat{V}(\widehat{t}_1) + \widehat{V}(\widehat{t}_2)} \tag{4.48}$$

Adicionalmente podemos afirmar que, el ECM con pesos óptimos α_{opt} se reduce a

$$ECM(\widehat{P}_{rA_d}^{(c_k)}) = \alpha_{opt} ECM(\widehat{t}_1) = (1 - \alpha_{opt}) ECM(\widehat{t}_2).$$

Se sigue de esta última expresión que la reducción en ECM conseguida por el estimador óptimo con respecto al más pequeño de los $ECMs$ que componen al estimador está dada por α_{opt} si $0 \leq \alpha_{opt} \leq 1/2$, e igual a $1 - \alpha_{opt}$, si $1/2 \leq \alpha_{opt} \leq 1$. Así la máxima reducción del 50 % se consigue cuando $\alpha_{opt} = 1/2$ o $F = 1$.

Si ocurre que la covarianza C es grande, entonces la aproximación óptima para los pesos α_{opt} depende, además, de la covarianza, que deberá estimarse. Especialmente la covarianza de estos estimadores combinados es difícil de estimar, pero podríamos emplear una aproximación usada por Eustat (2005), como se muestra a continuación.

Sabemos que

$$ECM(\widehat{P}_{rA_d}^{(c_k)}) = ECM\left[\alpha(\widehat{t}_1) + (1-\alpha)(\widehat{t}_2)\right]$$

$$= \alpha^2 ECM(\widehat{t}_1) + (1-\alpha)^2 ECM(\widehat{t}_2) + 2\alpha(1-\alpha)E[(\widehat{t}_1 - P_{A_{d1}})(\widehat{t}_2 - P_{A_{d2}})]$$

Una posible aproximación de

$$E[(\widehat{t}_1 - P_{A_{d1}})(\widehat{t}_2 - P_{A_{d2}})] = E[\widehat{t}_1 \times \widehat{t}_2] - P_{A_{d2}} E[\widehat{t}_1] - P_{A_{d1}} E[\widehat{t}_2] + P_{A_{d1}} P_{A_{d2}}$$

$$= E[\widehat{t}_1 \times \widehat{t}_2] - P_{A_{d2}} P_{A_{d1}} - P_{A_{d1}}(E[\widehat{t}_2] - P_{A_{d2}})$$

$$= E[\widehat{t}_1 \times \widehat{t}_2] - P_{A_{d2}} P_{A_{d1}} - P_{A_{d1}}[sesgo(\widehat{t}_2)]$$

$$\approx E[(\widehat{t}_1)^2] - P_{A_{d1}}^2 - P_{A_{d1}}[sesgo(\widehat{t}_2)]$$

$$= ECM[\widehat{t}_1] - P_{A_{d1}}[sesgo(\widehat{t}_2)]$$

En este caso $E[\widehat{t}_1 \times \widehat{t}_2]$ puede aproximarse con $E[(\widehat{t}_1)^2]$ debido a que la covarianza es distinta de cero solamente para los términos comunes de ambos

estimadores, es decir, para los términos que intervienen en el cálculo del estimador $\widehat{t}_1$. En este caso asumimos que $P_{A_{d1}} = P_{A_{d2}} = P_{A_d}$, la proporción de individuos en el dominio que poseen el atributo A y que $VA(\widehat{t}_1) > VA(\widehat{t}_2)$, si ocurre lo contrario, habrá que hacer los correspondientes ajustes en las expresiones para los estimadores de $\widehat{P}_{rA_d}^{(c_k)}$ y su varianza.

Por lo anterior, un estimador del ECM, estará dado por

$$\widehat{ECM}(\widehat{P}_{rA_d}^{(c_k)}) = \widehat{\alpha}_{opt}^2 \widehat{ECM}(\widehat{t}_1) + (1-\widehat{\alpha}_{opt})^2 \widehat{ECM}(\widehat{t}_2) \\ +2\widehat{\alpha}_{opt}(1-\widehat{\alpha}_{opt}) \left\{ \widehat{ECM}[\widehat{t}_1] - \widehat{P}_{A_{d1}}[\widehat{sesgo}(\widehat{t}_2)] \right\}$$

O como hemos expresado en (4.48), si asumimos que $ECM(\widehat{P}_{rA_d}^{(c_k)}) \doteq V(\widehat{P}_{rA_d}^{(c_k)})$, Särndal, et al. (1992, pág. 174), entonces

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_k)}) = \frac{\widehat{V}(\widehat{t}_1)\widehat{V}(\widehat{t}_2) - \left\{ \widehat{V}[\widehat{t}_1] - \widehat{P}_{A_d1}[\widehat{B}(\widehat{t}_2)] \right\}^2}{\widehat{V}(\widehat{t}_1) + \widehat{V}(\widehat{t}_2) - 2\left\{ \widehat{V}[\widehat{t}_2] - \widehat{P}_{A_{d1}}[\widehat{B}(\widehat{t}_2)] \right\}}. \tag{4.49}$$

Para un diseño de muestreo $MAS(N,n)$ se sustituyen en (4.48) las varianzas y sesgo estimados dados por (4.23) y (4.25), (4.32) y (4.34), si la covarianza C es pequeña, o en (4.49) si la covarianza es grande, para obtener las estimaciones de las varianzas de los estimadores combinados (4.36), (4.37), (4.38), (4.39) y (4.40).

Entonces, los estimadores combinados y los estimadores de sus varianzas serán, considerando el valor óptimo para α, los siguientes:

Para $\widehat{P}_{rA_d}^{(c_1)}$,

$$\widehat{P}_{rA_d}^{(c_1)} = \widehat{\alpha}_{op}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(1)}, \tag{4.50}$$

donde

$$\widehat{\alpha}_{opt} = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(1)}) - \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{rA_d}^{(1)})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(1)}) - 2\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{rA_d}^{(1)})}. \tag{4.51}$$

Su varianza estimada, en caso de que la covarianza sea pequeña, estará dada por

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_1)}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}_{rA_d}^{(1)})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(1)})} \tag{4.52}$$

y si la covarianza es grande

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_1)}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}_{rA_d}^{(1)}) - \left\{ \widehat{V}[\widehat{P}_{A_d}] - \widehat{P}_{A_d}[\widehat{B}(\widehat{P}_{rA_d}^{(1)})] \right\}^2}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(1)}) - 2\left\{ \widehat{V}[\widehat{P}_{A_d}] - \widehat{P}_{A_d}[\widehat{B}(\widehat{P}_{rA_d}^{(1)})] \right\}}. \tag{4.53}$$

Para $\widehat{P}_{rA_d}^{(c_2)}$,

$$\widehat{P}_{rA_d}^{(c_2)} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(2)}, \tag{4.54}$$

donde

$$\widehat{\alpha}_{opt} = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(2)}) - \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{rA_d}^{(2)})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(2)}) - 2\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{rA_d}^{(2)})}. \tag{4.55}$$

Su varianza estimada, en caso de que la covarianza sea pequeña, estará dada por

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_2)}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}_{rA_d}^{(2)})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(2)})} \tag{4.56}$$

y si la covarianza es grande

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_2)}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}_{rA_d}^{(2)}) - \left\{\widehat{V}[\widehat{P}_{A_d}] - \widehat{P}_{A_d}[\widehat{B}(\widehat{P}_{rA_d}^{(2)})]\right\}^2}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(2)}) - 2\left\{\widehat{V}[\widehat{P}_{A_d}] - \widehat{P}_{A_d}[\widehat{B}(\widehat{P}_{rA_d}^{(2)})]\right\}}. \tag{4.57}$$

Para $\widehat{P}_{rA_d}^{(c_3)}$,

$$\widehat{P}_{rA_d}^{(c_3)} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(3)}, \tag{4.58}$$

donde

$$\widehat{\alpha}_{opt} = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(3)}) - \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{rA_d}^{(3)})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)}) - 2\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{rA_d}^{(3)})}. \tag{4.59}$$

Su varianza estimada, en caso de que la covarianza sea pequeña, estará dada por

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_3)}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}_{rA_d}^{(3)})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)})} \tag{4.60}$$

y si la covarianza es grande

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_3)}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}_{rA_d}^{(3)}) - \left\{\widehat{V}[\widehat{P}_{A_d}] - \widehat{P}_{A_d}[\widehat{B}(\widehat{P}_{rA_d}^{(3)})]\right\}^2}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)}) - 2\left\{\widehat{V}[\widehat{P}_{A_d}] - \widehat{P}_{A_d}[\widehat{B}(\widehat{P}_{rA_d}^{(3)})]\right\}}. \tag{4.61}$$

Para $\widehat{P}_{rA_d}^{(c_4)}$,

$$\widehat{P}_{rA_d}^{(c_4)} = \widehat{\alpha}_{opt}\widehat{P}_{rA_d}^{(1)} + (1-\widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(3)}, \tag{4.62}$$

donde

$$\widehat{\alpha}_{opt} = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(3)}) - \widehat{cov}(\widehat{P}_{rA_d}^{(1)}, \widehat{P}_{rA_d}^{(3)})}{\widehat{V}(\widehat{P}_{rA_d}^{(1)}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)}) - 2\widehat{cov}(\widehat{P}_{rA_d}^{(1)}, \widehat{P}_{rA_d}^{(3)})}. \tag{4.63}$$

Su varianza estimada, en caso de que la covarianza sea pequeña, estará dada por

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_4)}) = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(1)})\widehat{V}(\widehat{P}_{rA_d}^{(3)})}{\widehat{V}(\widehat{P}_{rA_d}^{(1)}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)})} \tag{4.64}$$

y si la covarianza es grande

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_4)}) = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(1)})\widehat{V}(\widehat{P}_{rA_d}^{(3)}) - \left\{\widehat{V}[\widehat{P}_{rA_d}^{(1)}] - \widehat{P}_{rA_d}^{(1)}[\widehat{B}(\widehat{P}_{rA_d}^{(3)})]\right\}^2}{\widehat{V}(\widehat{P}_{rA_d}^{(1)}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)}) - 2\left\{\widehat{V}[\widehat{P}_{rA_d}^{(1)}] - \widehat{P}_{rA_d}^{(1)}[\widehat{B}(\widehat{P}_{rA_d}^{(3)})]\right\}}. \tag{4.65}$$

Para $\widehat{P}_{rA_d}^{(c_5)}$,

$$\widehat{P}_{rA_d}^{(c_5)} = \widehat{\alpha}_{opt}\widehat{P}_{rA_d}^{(2)} + (1 - \widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(3)}, \tag{4.66}$$

donde

$$\widehat{\alpha}_{opt} = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(3)}) - \widehat{cov}(\widehat{P}_{rA_d}^{(2)}, \widehat{P}_{rA_d}^{(3)})}{\widehat{V}(\widehat{P}_{rA_d}^{(2)}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)}) - 2\widehat{cov}(\widehat{P}_{rA_d}^{(2)}, \widehat{P}_{rA_d}^{(3)})}. \tag{4.67}$$

Su varianza estimada, en caso de que la covarianza sea pequeña, estará dada por

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_5)}) = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(2)})\widehat{V}(\widehat{P}_{rA_d}^{(3)})}{\widehat{V}(\widehat{P}_{rA_d}^{(2)}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)})} \tag{4.68}$$

y si la covarianza es grande

$$\widehat{AV}(\widehat{P}_{rA_d}^{(c_5)}) = \frac{\widehat{V}(\widehat{P}_{rA_d}^{(2)})\widehat{V}(\widehat{P}_{rA_d}^{(3)}) - \left\{\widehat{V}[\widehat{P}_{rA_d}^{(2)}] - \widehat{P}_{rA_d}^{(2)}[\widehat{B}(\widehat{P}_{rA_d}^{(3)})]\right\}^2}{\widehat{V}(\widehat{P}_{rA_d}^{(2)}) + \widehat{V}(\widehat{P}_{rA_d}^{(3)}) - 2\left\{\widehat{V}[\widehat{P}_{rA_d}^{(2)}] - \widehat{P}_{rA_d}^{(2)}[\widehat{B}(\widehat{P}_{rA_d}^{(3)})]\right\}}. \tag{4.69}$$

No se ha propuesto un estimador combinado entre $\widehat{P}_{rA_d}^{(1)}$ y $\widehat{P}_{rA_d}^{(2)}$ porque tienen la misma varianza y no produciría ninguna mejora en la estimación de P_{A_d}.

4.1.6. Estimador combinado de razón óptimo

Observemos que el estimador usual $\widehat{P}_{A_d} = p_{A_d}$ puede obtenerse también como $p_{A_d} = 1 - q_{A_d}$, donde $q_{A_d} = n_d^{-1}\sum_{j\in s_d} A_j^c$, ($A_j^c$ representa las unidades en la submuestra s_d que no poseen el atributo A), entonces p_{A_d} tiene el mismo comportamiento en la estimación de P_{A_d} que q_{A_d} en la estimación de Q_{A_d}. Sin embargo, esta propiedad no se cumple para $\widehat{P}_{rA_d}$, los estimadores de razón dados por (4.13) y (4.18), ya que se puede ver fácilmente que $\widehat{P}_{rA_d} \neq 1 - \widehat{Q}_{rA_d}$, donde $\widehat{Q}_{rA_d} = \widehat{R}_d^c Q_{B_d}$ y $\widehat{R}_d^c = q_{A_d}/q_{B_d}$. El estimador de razón complementario para la estimación de P_A, $\widehat{P}_{r.q} = 1 - \widehat{Q}_r$, definido por Rueda et al. (2011), para

el cual establecieron propiedades a partir de la comparación con el estimador directo de razón simple mediante el criterio de mínima varianza, verifica que si $P_A < P_B$, entonces $AV(\widehat{P}_r) < AV(\widehat{P}r.q)$ y $\widehat{P}_r$ es más eficiente que $\widehat{P}_{r.q}$; en caso contrario, se cumple que es más eficiente $\widehat{P}_{r.q}$.

Esta teoría puede aplicarse a la estimación de razón de P_A en dominios y definimos el estimador complementario de razón como

$$\widehat{P}_{rA_dq} = 1 - \widehat{Q}_{rA_d}, \tag{4.70}$$

El objetivo consiste en definir un nuevo estimador combinado, usando una combinación lineal del estimador de razón dado por (4.13) o (4.18) con $\widehat{P}_{rA_dq}$.

El nuevo estimador combinado es de la forma

$$\widehat{P}_{rA_d\alpha} = \alpha\widehat{P}_{rA_d} + (1-\alpha)\widehat{P}_{rA_dq} \tag{4.71}$$

donde α es el peso que pondera a ambos estimadores y se obtiene de (4.42) si la covarianza es grande o de (4.47) y si la covarianza es pequeña. La expresión para los pesos óptimos en el estimador compuesto (4.71) es

$$\alpha_{opt} = \frac{V(\widehat{P}_{rA_dq}) - cov(\widehat{P}_{rA_d}, \widehat{P}_{rA_dq})}{V(\widehat{P}_{rA_d}) + V(\widehat{P}_{rA_dq}) - 2cov(\widehat{P}_{rA_d}, \widehat{P}_{rA_dq})}, \tag{4.72}$$

y su estimador

$$\widehat{\alpha}_{opt} = \frac{\widehat{V}(\widehat{P}_{rA_dq}) - \widehat{cov}(\widehat{P}_{rA_d}, \widehat{P}_{rA_dq})}{\widehat{V}(\widehat{P}_{rA_d}) + \widehat{V}(\widehat{P}_{rA_dq}) - 2\widehat{cov}(\widehat{P}_{rA_d}.\widehat{P}_{rA_dq})}, \tag{4.73}$$

El término para la covarianza entre los dos estimadores se obtiene como sigue

$$C = Cov(\widehat{P}_{rA_d}, \widehat{P}_{rA_dq}) = Cov(\widehat{P}_{rA_d}, 1 - \widehat{Q}_{rA_d}) = -Cov(\widehat{P}_{rA_d}, \widehat{Q}_{rA_d}) =$$

$$-Cov(\widehat{R}_d P_{B_d}, \widehat{R}^c_d Q_{B_d}) = -P_{B_d}Q_{B_d}Cov(\widehat{R}_d, \widehat{R}^c_d)$$

y en términos de las aproximaciones lineales para R, se obtiene

$$C = -P_{B_d}Q_{B_d}Cov\left[R_d + \frac{1}{P_{B_d}}(\widehat{P}_{A_d} - R_d\widehat{P}_{B_d}), R^c_d + \frac{1}{Q_{B_d}}(\widehat{Q}_{A_d} - R^c_d\widehat{Q}_{B_d})\right]$$

o de forma equivalente

$$-P_{B_d}Q_{B_d}Cov\left[R_d + \frac{1}{P_{B_d}}(\widehat{P}_{A_d} - R_d\widehat{P}_{B_d}), R^c_d + \frac{1}{Q_{B_d}}(1 - \widehat{P}_{A_d} - R^c_d(1 - \widehat{P}_{B_d}))\right]$$

o también

$$-P_{B_d}Q_{B_d}Cov\left[R_d+\frac{1}{P_{B_d}}(\widehat{P}_{A_d}-R_d\widehat{P}_{B_d}),R_d^c+\frac{1}{Q_{B_d}}(1-R_d^c-\widehat{P}_{A_d}+R_d^c\widehat{P}_{B_d})\right]$$

aplicando las propiedades de la covarianza y mediante sencillas operaciones algebraicas obtenemos que

$$C=V(\widehat{P}_{A_d})+R_dR_d^cV(\widehat{P}_{B_d})-(R_d+R_d^c)cov(\widehat{P}_{A_d},\widehat{P}_{B_d}) \tag{4.74}$$

y su estimador

$$\widehat{C}=\widehat{V}(\widehat{P}_{A_d})+\widehat{R}_d\widehat{R}_d^c\widehat{V}(\widehat{P}_{B_d})-(\widehat{R}_d+\widehat{R}_d^c)\widehat{cov}(\widehat{P}_{A_d},\widehat{P}_{B_d}). \tag{4.75}$$

Entonces, el estimador óptimo estará dado por

$$\widehat{P}_{rA_d\alpha}^{opt}=\widehat{\alpha}_{opt}\widehat{P}_{rA_d}+(1-\widehat{\alpha}_{opt})\widehat{P}_{rA_dq}. \tag{4.76}$$

Si la covarianza es grande, esta se obtiene con (4.75) y su varianza estará dada por (4.46); en caso del supuesto de covarianza pequeña o cero, la varianza se obtiene con (4.48). Para ambos casos, $\widehat{V}(\widehat{P}_{rA_d})=\widehat{V}(\widehat{P}_{A_d})+\widehat{R}_d^2\widehat{V}(\widehat{P}_{B_d})-2\widehat{R}_d\widehat{cov}(\widehat{P}_{A_d},\widehat{P}_{B_d})$ y $\widehat{V}(\widehat{P}_{rA_dq})=\widehat{V}(\widehat{P}_{A_d})+(\widehat{R}_d^c)^2\widehat{V}(\widehat{P}_{B_d})-2\widehat{R}_d^c\widehat{cov}(\widehat{P}_{A_d},\widehat{P}_{B_d})$.
La expresión para la varianza del estimador óptimo se escribirá como

$$AV(\widehat{P}_{rA_d\alpha}^{opt})=\frac{V(\widehat{P}_{rA_d})V(\widehat{P}_{rA_dq})-cov^2(\widehat{P}_{rA_d},\widehat{P}_{rA_dq})}{V(\widehat{P}_{rA_d})+V(\widehat{P}_{rA_dq})-2cov(\widehat{P}_{rA_d},\widehat{P}_{rA_dq})}$$

haciendo las correspondientes sustituciones y operando algebraicamente, se obtiene

$$AV(\widehat{P}_{rA_d\alpha}^{opt})=\frac{V(\widehat{P}_{A_d})V(\widehat{P}_{B_d})-cov^2(\widehat{P}_{A_d},\widehat{P}_{B_d})}{V(\widehat{P}_{B_d})} \tag{4.77}$$

cuya demostración completa puede consultarse en un trabajo reciente de J.F. Muñoz et al. (2011).

Bajo un diseño $MAS(N,n)$, la expresión para la varianza del estimador óptimo es

$$AV(\widehat{P}_{rA_d\alpha}^{opt})=\frac{N}{N_d}\frac{1-f}{n}\left[\frac{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}-\phi_d^2P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}{P_{B_d}Q_{B_d}}\right]=$$

$$AV(\widehat{P}_{rA_d\alpha}^{opt})=\frac{N}{N_d}\frac{1-f}{n}P_{A_d}Q_{A_d}(1-\phi_d^2) \tag{4.78}$$

y su estimador

$$\widehat{AV}(\widehat{P}_{rA_d\alpha}^{opt})=\frac{1-f}{n_d}\frac{n}{n-1}P_{A_d}Q_{A_d}(1-\widehat{\phi}_d^2). \tag{4.79}$$

4.1.7. Estimador de razón multivariante

Caso de p atributos auxiliares con información a nivel poblacional

Asumamos ahora que el atributo de interés A está asociado a p atributos auxiliares $B_1, \cdots, B_p$, de los cuales se asume conocida P_{B_i}, $i = 1, \cdots, p$.

El estimador de razón en presencia de atributos auxiliares multivariantes, a nivel de dominio, estará dado por

$$\widehat{P}_{A_{d.rM}} = \sum_{i=1}^{p} w_i \widehat{P}_{r_i d} = \mathbf{w}\widehat{\mathbf{P}}_{rd}^{'}, \tag{4.80}$$

donde $\widehat{P}_{r_i d} = \widehat{R}_{id} P_{B_i}$, $\widehat{R}_{id} = \frac{\widehat{P}_{A_d}}{\widehat{P}_{B_i d}}$, $\mathbf{w} = (w_1, \cdots, w_p)$ y $\widehat{\mathbf{P}}_{rd} = (\widehat{P}_{r_1 d}, \cdots, \widehat{P}_{r_i d})$.

Como se observa en el estimador propuesto, se estimarán las proporciones de razón para cada uno de los p atributos B y se ponderarán mediante los pesos w_i, $i = 1 \cdots, p$, que deberán satisfacer la condición $\sum_p w_i = 1$ y serán calculados de manera que se maximize la precisión del estimador propuesto $\widehat{P}_{A_{d.rM}}$.

La varianza de $\widehat{P}_{A_{d.rM}}$ puede escribirse como

$$AV(\widehat{P}_{A_{d.rM}}) = \mathbf{wCw'}, \tag{4.81}$$

donde $\mathbf{C} = (c_{ik})$ se define como una matriz de dimensión $p \times p$ con $c_{ik} = cov(\widehat{P}_{r_i d}, \widehat{P}_{r_k d})$, $i \neq k$, y $c_{ii} = AV(\widehat{P}_{r_i d})$, con $i, k = 1, \cdots, p$, donde

$$AV(\widehat{P}_{r_i d}) = \left(\frac{P_{B_i}}{P_{B_i d}}\right)^2 \left[V(\widehat{P}_{A_d}) + R_{id}^2 V(\widehat{P}_{B_i d}) - 2R_{id} cov(\widehat{P}_{A_d}, \widehat{P}_{B_i d})\right]$$

y

$$cov(\widehat{P}_{r_i d}, \widehat{P}_{r_k d}) = \left(\frac{P_{B_i}}{P_{B_i d}}\right)^2 \left[V(\widehat{P}_{A_d}) + R_{id} R_{kd} cov(\widehat{P}_{B_i d}, \widehat{P}_{B_k d}) \right.$$
$$\left. - R_{id} cov(\widehat{P}_{A_d}, \widehat{P}_{B_i d}) - R_{kd} cov(\widehat{P}_{A_d}, \widehat{P}_{B_k d})\right].$$

Bajo un diseño $MAS(N, n)$, estas últimas expresiones se escriben

$$AV(\widehat{P}_{r_i d}) = \left(\frac{P_{B_i}}{P_{B_i d}}\right)^2 \frac{N}{N_d} \frac{1-f}{n} \left[P_{A_d} Q_{A_d} + R_{id}^2 P_{B_i d} Q_{B_i d} \right. \tag{4.82}$$
$$\left. - R_{id} \phi_i \sqrt{P_{A_d} Q_{A_d} P_{B_i d} Q_{B_i d}}\right]$$

y

$$cov(\widehat{P}_{r_id}, \widehat{P}_{r_kd}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \frac{N}{N_d}\frac{1-f}{n}[P_{A_d}Q_{A_d} + R_{id}R_{kd}\phi_{ik}\sqrt{P_{B_id}Q_{B_id}P_{B_kd}Q_{B_kd}} - R_{id}\phi_i\sqrt{P_{A_d}Q_{A_d}P_{B_id}Q_{B_id}} - R_{kd}\phi_k\sqrt{P_{A_d}Q_{A_d}P_{B_kd}Q_{B_kd}}].$$

Una expresión para el sesgo de $\widehat{P}_{A_{d.rM}}$ será

$$B(\widehat{P}_{A_{d.rM}}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \sum_{i=1}^{p} w_i P_{A_d}\left[\frac{V(\widehat{P}_{B_id})}{P^2_{B_id}} - \frac{cov(\widehat{P}_{A_d}, \widehat{P}_{B_id})}{P_{A_d}P_{B_id}}\right]$$

su estimador

$$\widehat{B}(\widehat{P}_{A_{d.rM}}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \sum_{i=1}^{p} w_i \widehat{P}_{A_d}\left[\frac{\widehat{V}(\widehat{P}_{B_id})}{\widehat{P}^2_{B_id}} - \frac{\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_id})}{\widehat{P}_{A_d}\widehat{P}_{B_id}}\right]$$

y bajo un diseño $MAS(N,n)$,

$$\widehat{B}(\widehat{P}_{A_{d.rM}}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \frac{1-f}{n_d}\frac{n}{n-1}\sum_{i=1}^{p} w_i\left[\widehat{R}_{id}\widehat{Q}_{B_id} - \widehat{\phi}_i\frac{\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}}}{\widehat{P}_{B_id}}\right] \tag{4.83}$$

Dado que $\mathbf{C}$ es semidefinida positiva, la desigualdad generalizada de Cauchy–Schwarz puede usarse para mostrar que el óptimo $\mathbf{w}$ que minimiza $AV(\widehat{P}_{A_{d.rM}})$ es

$$\mathbf{w}_{opt} = \frac{\mathbf{e}\mathbf{C}^{-1}}{\mathbf{e}\mathbf{C}^{-1}\mathbf{e}'}$$

donde $\mathbf{e} = (1, \cdots, 1)$.

Sustituyendo $\mathbf{w}_{opt}$ en (4.81) obtenemos

$$AV(A_{d.rM})_{min} = \frac{1}{\mathbf{e}\mathbf{C}^{-1}\mathbf{e}'}.$$

Bajo un diseño $MAS(N,n)$, el estimador de P_{A_d} estará dado por

$$\widehat{P}_{A_{d.rM}} = \widehat{\mathbf{w}}_{opt}\widehat{\mathbf{P}}'_{rd} \tag{4.84}$$

donde, $\widehat{\mathbf{w}}_{opt} = \frac{\mathbf{e}\widehat{\mathbf{C}}^{-1}}{\mathbf{e}\widehat{C}^{-1}\mathbf{e}'}$, $\widehat{\mathbf{C}} = (\widehat{c}_{ik})$, $\widehat{c}_{ik} = \widehat{cov}(\widehat{P}_{r_id}, \widehat{P}_{r_kd})$, $i \neq k$, y $\widehat{c}_{ii} = \widehat{V}(\widehat{P}_{r_id})$, con $i, k = 1, \cdots, p$,

$$\widehat{V}(\widehat{P}_{r_id}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \frac{1-f}{n_d}\frac{n}{n-1}\left(\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}^2_{id}\widehat{P}_{B_id}\widehat{Q}_{B_id} - 2\widehat{R}_{id}\widehat{\phi}_i\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}}\right) \tag{4.85}$$

y

$$\widehat{cov}(\widehat{P}_{r_id},\widehat{P}_{r_kd}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \frac{1-f}{n_d}\frac{n}{n-1}\Bigg[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}_{id}\widehat{R}_{kd}\widehat{\phi}_{ik}\sqrt{\widehat{P}_{B_id}\widehat{Q}_{B_id}\widehat{P}_{B_kd}\widehat{Q}_{B_kd}} - \widehat{R}_{id}\widehat{\phi}_i\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}} - \widehat{R}_{kd}\widehat{\phi}_k\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_kd}\widehat{Q}_{B_kd}}\Bigg].$$

Ahora, si $n/(n-1) \approx 1$, entonces

$$\widehat{V}(\widehat{P}_{r_id}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \frac{1-f}{n_d}\left(\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}^2_{id}\widehat{P}_{B_id}\widehat{Q}_{B_id} - 2\widehat{R}_{id}\widehat{\phi}_i\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}}\right)$$

y

$$\widehat{cov}(\widehat{P}_{r_id},\widehat{P}_{r_kd}) = \left(\frac{P_{B_i}}{P_{B_id}}\right)^2 \frac{1-f}{n_d}[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}_{id}\widehat{R}_{kd}\widehat{\phi}_{ik}\sqrt{\widehat{P}_{B_id}\widehat{Q}_{B_id}\widehat{P}_{B_kd}\widehat{Q}_{B_kd}} - \widehat{R}_{id}\widehat{\phi}_i\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}} - \widehat{R}_{kd}\widehat{\phi}_k\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_kd}\widehat{Q}_{B_kd}}\ ],$$

donde $\widehat{\phi}_i$ se estima de la submuestra organizando los datos A y B_j, $i = 1, \cdots, p$, con sus respectivos complementos, en una tabla de doble entrada. Se procede de forma similar en la el cálculo de $\widehat{\phi}_{ik}$, es decir, hacemos una tabla de doble entrada con los datos de B_i y B_k, $i \neq k$, con sus respectivos complementos (ver apéndice A).

Caso de p atributos auxiliares a nivel de dominio

Bajo el supuesto de que se conoce información auxiliar a nivel de dominio, es decir, es conocida P_{B_d}, el estimador de P_{A_d} estará dado por

$$\widehat{P}_{A_{d.rM}} = \sum_{i=1}^{p} w_i\widehat{P}_{r_id} = \mathbf{w}\widehat{\mathbf{P}}'_{rd}, \tag{4.86}$$

donde $\widehat{P}_{r_id} = \widehat{R}_{id}P_{B_id}$, $\widehat{R}_{id} = \frac{\widehat{P}_{A_d}}{\widehat{P}_{B_id}}$, $\mathbf{w} = (w_1, \cdots, w_p)$ y $\widehat{\mathbf{P}}_{rd} = (\widehat{P}_{r_1d}, \cdots, \widehat{P}_{r_pd})$.

Obsérvese que la única diferencia entre este estimador y el dado por(4.80) se da en $\widehat{P}_{r_id}$, el estimador de razón de la proporción a nivel de dominio, toda vez que se conocen P_{B_i} para el caso de (4.80) o P_{B_id} para el caso de (4.86).

Bajo un diseño $MAS(N,n)$ varianza aproximada y su estimador estarán dados por

$$AV(\widehat{P}_{r_id}) = \frac{N}{N_d}\frac{1-f}{n}\left[P_{A_d}Q_{A_d} + R^2_{id}P_{B_id}Q_{B_id} - R_{id}\phi_i\sqrt{P_{A_d}Q_{A_d}P_{B_id}Q_{B_id}}\right] \tag{4.87}$$

y

$$cov(\widehat{P}_{r_id}, \widehat{P}_{r_kd}) = \frac{N}{N_d}\frac{1-f}{n}[P_{A_d}Q_{A_d} + R_{id}R_{kd}\phi_{ik}\sqrt{P_{B_id}Q_{B_id}P_{B_kd}Q_{B_kd}} - R_{id}\phi_i\sqrt{P_{A_d}Q_{A_d}P_{B_id}Q_{B_id}} - R_{kd}\phi_k\sqrt{P_{A_d}Q_{A_d}P_{B_kd}Q_{B_kd}}].$$

$$\widehat{V}(\widehat{P}_{r_id}) = \frac{1-f}{n_d}\frac{n}{n-1}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}^2_{id}\widehat{P}_{B_id}\widehat{Q}_{B_id} - 2\widehat{R}_{id}\widehat{\phi}_i\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}}\right] \tag{4.88}$$

y

$$\widehat{cov}(\widehat{P}_{r_id}, \widehat{P}_{r_kd}) = \frac{1-f}{n_d}\frac{n}{n-1}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \widehat{R}_{id}\widehat{R}_{kd}\widehat{\phi}_{ik}\sqrt{\widehat{P}_{B_id}\widehat{Q}_{B_id}\widehat{P}_{B_kd}\widehat{Q}_{B_kd}} - \widehat{R}_{id}\widehat{\phi}_i\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}} - \widehat{R}_{kd}\widehat{\phi}_k\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_kd}\widehat{Q}_{B_kd}}\right].$$

4.2. Estimadores de regresión propuestos

4.2.1. Información auxiliar de dominio: se conoce P_{B_d}.

Consideremos ahora, un estimador "tipo regresión" para estimar P_{A_d}, la proporción de individuos que poseen el atributo A dentro del dominio d, si contamos con información auxiliar del atributo B asociada al atributo A a nivel de dominio. El estimador propuesto es

$$\widehat{P}^{(1)}_{regA_d} = \widehat{P}_{A_d} + b(P_{B_d} - \widehat{P}_{B_d}) \tag{4.89}$$

donde se asume conocido P_{B_d}, la proporción de individuos que poseen el atributo B en el dominio d.

En este caso, b es una constante arbitraria conocida, que en el modelo de regresión lineal para variables continuas representa la pendiente de la recta de regresión, y $\widehat{P}_{B_d}$ es el estimador de P_{B_d}. Aunque b se supone arbitraria, debería tomar valores preferentemente próximos al valor poblacional R o β, el coeficiente de regresión lineal de mínimos cuadrados.

Una expresión para la varianza de $\widehat{P}^{(1)}_{regA_d}$, se obtiene como sigue

$$V(\widehat{P}^{(1)}_{regA_d}) = V[\widehat{P}_{A_d} + b(P_{B_d} - \widehat{P}_{B_d})] = V[\widehat{P}_{A_d} + bP_{B_d} - b\widehat{P}_{B_d}]$$

$$V(\widehat{P}^{(1)}_{regA_d}) = V(\widehat{P}_{A_d}) + b^2V(\widehat{P}_{B_d}) - 2bcov(\widehat{P}_{A_d}, \widehat{P}_{B_d}) \tag{4.90}$$

y tendrá como estimador

$$\widehat{V}(\widehat{P}^{(1)}_{regA_d}) = \widehat{V}(\widehat{P}_{A_d}) + b^2\widehat{V}(\widehat{P}_{B_d}) - 2b\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_d}). \tag{4.91}$$

Bajo un diseño muestral $MAS(N, n)$, la varianza dada por (4.90) se escribe como

$$V(\widehat{P}^{(1)}_{regA_d}) = \frac{N}{N_d}\frac{1-f}{n}\left[P_{A_d}Q_{A_d} + b^2P_{B_d}Q_{B_d} - 2b\phi_d\sqrt{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}\right] \tag{4.92}$$

y su estimador

$$\widehat{V}(\widehat{P}^{(1)}_{regA_d}) = \frac{1-f}{n_d}\frac{n}{n-1}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + b^2\widehat{P}_{B_d}\widehat{Q}_{B_d} - 2b\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\right], \tag{4.93}$$

o también

$$\widehat{V}(\widehat{P}^{(1)}_{regA_d}) = \frac{1-f}{n_d}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + b^2\widehat{P}_{B_d}\widehat{Q}_{B_d} - 2b\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\right]$$

si $n/(n-1) \approx 1$.

En la práctica b no es conocido y debe estimarse. Una expresión para b puede hallarse minimizando (4.90) como sigue:

$$\frac{\partial V(\widehat{P}^{(1)}_{regA_d})}{\partial b} = 2bV(\widehat{P}_{B_d}) - 2cov(\widehat{P}_{A_d}, \widehat{P}_{B_d}),$$

igualando a cero se consigue

$$2bV(\widehat{P}_{B_d}) - 2cov(\widehat{P}_{A_d}, \widehat{P}_{B_d}) = bV(\widehat{P}_{B_d}) - cov(\widehat{P}_{A_d}, \widehat{P}_{B_d}) = 0$$

$$b = \frac{cov(\widehat{P}_{A_d}, \widehat{P}_{B_d})}{V(\widehat{P}_{B_d})},$$

que es un mínimo porque

$$\frac{\partial^2 V(\widehat{P}^{(1)}_{regA_d})}{\partial b^2} = 2V(\widehat{P}_{B_d}) > 0.$$

Entonces podemos escribir

$$b_{opt} = \frac{cov(\widehat{P}_{A_d}, \widehat{P}_{B_d})}{V(\widehat{P}_{B_d})}, \tag{4.94}$$

y su estimador

$$\widehat{b}_{opt} = \frac{\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_d})}{\widehat{V}(\widehat{P}_{B_d})}, \tag{4.95}$$

Bajo un diseño $MAS(N, n)$, (4.94) se escribe

$$b_{opt} = \frac{\phi_d \sqrt{P_{A_d} Q_{A_d} P_{B_d} Q_{B_d}}}{P_{B_d} Q_{B_d}} = \phi_d \frac{\sqrt{P_{A_d} Q_{A_d}}}{\sqrt{P_{B_d} Q_{B_d}}}$$

y (4.95)

$$\widehat{b}_{opt} = \widehat{\phi}_d \frac{\sqrt{\widehat{P}_{A_d} \widehat{Q}_{A_d}}}{\sqrt{\widehat{P}_{B_d} \widehat{Q}_{B_d}}},$$

y entonces, el estimador de regresión óptimo para P_{A_d} en (4.89) será

$$\widehat{P}^{(1)}_{regA_d} = \widehat{P}_{A_d} + \widehat{b}_{opt}(P_{B_d} - \widehat{P}_{B_d}). \tag{4.96}$$

Bajo muestreo aleatorio simple, la varianza de este estimador viene dada por:

$$AV(\widehat{P}^{(1)}_{regA_d}) = \frac{N}{N_d} \frac{1-f}{n} \left[P_{A_d} Q_{A_d} + \left(\phi_d \frac{\sqrt{P_{A_d} Q_{A_d}}}{\sqrt{P_{B_d} Q_{B_d}}} \right)^2 P_{B_d} Q_{B_d} \right.$$
$$\left. -2 \left(\phi_d \frac{\sqrt{P_{A_d} Q_{A_d}}}{\sqrt{P_{B_d} Q_{B_d}}} \right) \phi_d \sqrt{P_{A_d} Q_{A_d} P_{B_d} Q_{B_d}} \right]$$

$$AV(\widehat{P}^{(1)}_{regA_d}) = \frac{N}{N_d}\frac{1-f}{n}P_{A_d}Q_{A_d}(1-\phi_d^2) \tag{4.97}$$

y un estimador de su varianza viene dado por:

$$\widehat{V}(\widehat{P}^{(1)}_{regA_d}) = \frac{1-f}{n_d-1}\widehat{P}A_d\widehat{Q}_{A_d}(1-\widehat{\phi}_d^2). \tag{4.98}$$

4.2.2. Información auxiliar poblacional: se conoce P_B

Si se asume conocido P_B, la proporción de individuos que poseen el atributo B en la población, el estimador propuesto será

$$\widehat{P}^{(2)}_{regA_d} = \widehat{P}_{A_d} + b(P_B - \widehat{P}_{B_d}), \tag{4.99}$$

cuya expresión para la varianza será

$$V(\widehat{P}^{(2)}_{regA_d}) = V[\widehat{P}_{A_d} + b(P_B - \widehat{P}_{B_d})] = V[\widehat{P}_{A_d} + bP_B - b\widehat{P}_{B_d}]$$

$$V(\widehat{P}^{(2)}_{regA_d}) = V(\widehat{P}_{A_d}) + b^2V(\widehat{P}_{B_d}) - 2bcov(\widehat{P}_{A_d}, \widehat{P}_{B_d}),$$

que es idéntica a la expresión dada por (4.90), por lo que el estimador (4.99) para b óptimo bajo un diseño $MAS(N,n)$ será

$$\widehat{P}^{(2)}_{regA_d} = \widehat{P}_{A_d} + \widehat{b}_{opt}(P_B - \widehat{P}_{B_d}), \tag{4.100}$$

y su varianza estimada

$$\widehat{V}(\widehat{P}^{(2)}_{regA_d}) = \frac{1-f}{n_d-1}\widehat{P}A_d\widehat{Q}_{A_d}(1-\widehat{\phi}_d^2).$$

Observe que las varianzas y covarianzas para (4.89) y (4.99) son las mismas, ya que solo difieren en las constantes P_B y P_{B_d}. Además, los estimadores de las varianzas son idénticos a los estimadores de la varianza del estimador combinado de razón óptimo (4.79).

Estimadores de diferencia

Si $b = 1$ en (4.89) y (4.99), se obtiene el conocido estimador de diferencia. Si se conoce P_{B_d}

$$\widehat{P}^{(1)}_{difA_d} = \widehat{P}_{A_d} + (P_{B_d} - \widehat{P}_{B_d}), \tag{4.101}$$

si se conoce P_B,

$$\widehat{P}^{(2)}_{difA_d} = \widehat{P}_{A_d} + (P_B - \widehat{P}_{B_d}), \tag{4.102}$$

o el de diferencia sintético

$$\widehat{P}^{(3)}_{difA_d} = \widehat{P}_A + (P_B - \widehat{P}_B) \tag{4.103}$$

La varianza bajo un diseño $MAS(N,n)$ para (4.102) y (4.101) es

$$AV(\widehat{P}_{difA_d}) = \frac{N}{N_d}\frac{1-f}{n}\left(P_{A_d}Q_{A_d} + P_{B_d}Q_{B_d} - 2\phi\sqrt{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}\right)$$
$$= \frac{N}{N_d}\frac{1-f}{n}\left(P_{A_d}Q_{A_d} + P_{B_d}Q_{B_d} - 2\frac{\sqrt{P_{B_d}Q_{B_d}}}{\sqrt{P_{A_d}Q_{A_d}}}\sqrt{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}\right)$$

$$AV(\widehat{P}_{difA_d}) = \frac{N}{N_d}\frac{1-f}{n}\left[P_{A_d}Q_{A_d} - P_{B_d}Q_{B_d}\right] \tag{4.104}$$

y su estimador

$$\widehat{AV}(\widehat{P}_{difA_d}) = \frac{1-f}{n_d}\frac{n}{n-1}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} - \widehat{P}_{B_d}\widehat{Q}_{B_d}\right] \tag{4.105}$$

ya que sabemos que si $b=1$, entonces $\widehat{\phi}_d = \frac{\sqrt{\widehat{P}_{B_d}\widehat{Q}_{B_d}}}{\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}}}$.

Si $n/(n-1) \approx 1$, el estimador de la varianza se puede aproximar por

$$\widehat{AV}(\widehat{P}_{difA_d}) = \frac{1-f}{n_d}\left(\widehat{P}_{A_d}\widehat{Q}_{A_d} - \widehat{P}_{B_d}\widehat{Q}_{B_d}\right).$$

Para (4.103), el estimador de diferencia sintético, se tiene

$$AV(\widehat{P}_{difA_d}) = \frac{N-n}{(N-1)n}\frac{1-f}{n}\left[P_AQ_A - P_BQ_B\right] \tag{4.106}$$

y su estimador

$$\widehat{AV}(\widehat{P}_{difA_d}) = \frac{1-f}{n-1}\left[\widehat{P}_A\widehat{Q}_A - \widehat{P}_B\widehat{Q}_B\right]. \tag{4.107}$$

Es posible hallar, también, los correspondientes estimadores combinados de diferencia.

4.2.3. Estimador de regresión sintético

En este caso suponemos que el estimador de regresión basado en la muestra s, puede ser usado para estimar la proporción P_{A_d} correspondiente al dominio d, si se asume conocido, además, información auxiliar a nivel de dominio, es decir, se conoce P_{B_d}. El estimador propuesto es

$$\widehat{P}^{(3)}_{regA_d} = \widehat{P}_A + \widehat{b}(P_{B_d} - \widehat{P}_B) \tag{4.108}$$

siendo $\widehat{b} = \frac{\widehat{cov}(\widehat{P}_A, \widehat{P}_B)}{\widehat{V}(\widehat{P}_B)}$

La varianza del estimador de regresión sintético viene dada por:

$$AV(\widehat{P}^{(3)}_{regA_d}) = \frac{N-n}{(N-1)n} P_A Q_A (1-\phi^2) \tag{4.109}$$

y su estimador

$$\widehat{AV}(\widehat{P}^{(3)}_{regA_d}) = \frac{1-f}{n-1} \widehat{P}_A \widehat{Q}_A (1-\widehat{\phi}^2). \tag{4.110}$$

4.2.4. Extensión de los estimadores de regresión a un diseño general de muestreo

Hemos seleccionado una muestra aleatoria s bajo un diseño general de muestreo, con probabilidades de inclusión π_j y π_{jk}. Usando información auxiliar definimos el estimador de regresión de la forma (4.89) o (4.99), entonces, una expresión aproximada de la varianza se obtiene usando las propiedades de los estimadores de H-T,

$$AV(\widehat{P}_{regA_d}) = V\left(\sum_{j\in U_d} \frac{A_j - bB_j}{\pi_j}\right),$$

con b dada como en (4.94) o de forma equivalente,

$$b = \frac{\sum_{j\in U_d}(A_j - P_{A_d})(B_j - P_{B_d})}{\sum_{j\in U_d}(B_j - P_{B_d})^2},$$

entonces

$$AV(\widehat{P}_{regA_d}) = \frac{1}{N_d^2} \sum_{j,k\in U_d} (\pi_j\pi_k - \pi_{jk}) \left(\frac{E_j}{\pi_j}\right)\left(\frac{E_k}{\pi_k}\right)$$

y su estimador

$$\widehat{AV}(\widehat{P}_{regA_d}) = \frac{1}{\widehat{N}_d^2} \sum_{j,k\in s_d} \left(\frac{\pi_j\pi_k - \pi_{jk}}{\pi_{jk}}\right)\left(\frac{e_j}{\pi_j}\right)\left(\frac{e_k}{\pi_k}\right), \tag{4.111}$$

con $E_j = A_j - bB_j$ y $e_j = A_j - \widehat{b}B_j$.

4.2.5. Estimadores combinados de regresión

Con los estimadores de P_{A_d} dados por (4.6), (4.96), (4.100) y (4.108) se pueden construir los siguientes estimadores combinados

$$\widehat{P}^{(c_1)}_{regA_d} = \alpha \widehat{P}_{A_d} + (1-\alpha)\widehat{P}^{(1)}_{regA_d}, \tag{4.112}$$

$$\widehat{P}^{(c_2)}_{regA_d} = \alpha \widehat{P}_{A_d} + (1-\alpha)\widehat{P}^{(2)}_{regA_d}, \tag{4.113}$$

$$\widehat{P}^{(c_3)}_{regA_d} = \alpha \widehat{P}_{A_d} + (1-\alpha)\widehat{P}^{(3)}_{regA_d}, \tag{4.114}$$

$$\widehat{p}^{(c_4)}_{regA_d} = \alpha \widehat{P}^{(1)}_{regA_d} + (1-\alpha)\widehat{P}^{(3)}_{regA_d}, \tag{4.115}$$

$$\widehat{P}^{(c_2)}_{regA_d} = \alpha \widehat{P}^{(2)}_{regA_d} + (1-\alpha)\widehat{P}^{(3)}_{regA_d}, \tag{4.116}$$

donde $0 \leq \alpha \leq 1$ y $\alpha = \frac{V_2 - C}{V_1 + V_2 - 2C}$.

Si asumimos que la covarianza C es pequeña, entonces la aproximación óptima para los pesos α_{opt} depende solamente de la razón de varianzas, por lo que puede escribirse como

$$\alpha_{opt} = \frac{1}{1+F}$$

donde $F = \frac{V(\widehat{t}_1)}{V(\widehat{t}_2)}$.

Bajo este supuesto los estimadores combinados y los estimadores de sus varianzas, considerando el valor óptimo para α, son los siguientes:

Para $\widehat{P}^{(c_1)}_{regA_d}$,

$$\widehat{P}^{(c_1)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(1)}_{regA_d}, \tag{4.117}$$

y su varianza estimada

$$\widehat{AV}(\widehat{P}^{(c_1)}_{regA_d}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}^{(1)}_{regA_d})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}^{(1)}_{regA_d})}. \tag{4.118}$$

Para $\widehat{P}^{(c_2)}_{regA_d}$,

$$\widehat{P}^{(c_2)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(2)}_{regA_d}, \tag{4.119}$$

y su varianza estimada

$$\widehat{AV}(\widehat{P}^{(c_2)}_{regA_d}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}^{(2)}_{regA_d})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}^{(2)}_{regA_d})}. \tag{4.120}$$

Para $\widehat{P}^{(c_3)}_{regA_d}$,

$$\widehat{P}^{(c_3)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(3)}_{regA_d}, \tag{4.121}$$

y su varianza estimada

$$\widehat{AV}(\widehat{P}^{(c_3)}_{regA_d}) = \frac{\widehat{V}(\widehat{P}_{A_d})\widehat{V}(\widehat{P}^{(3)}_{regA_d})}{\widehat{V}(\widehat{P}_{A_d}) + \widehat{V}(\widehat{P}^{(3)}_{regA_d})}. \tag{4.122}$$

Para $\widehat{P}^{(c_4)}_{regA_d}$,

$$\widehat{P}^{(c_4)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}^{(1)}_{regA_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(3)}_{regA_d}, \tag{4.123}$$

y su varianza estimada

$$\widehat{AV}(\widehat{P}^{(c_4)}_{regA_d}) = \frac{\widehat{V}(\widehat{P}^{(1)}_{regA_d})\widehat{V}(\widehat{P}^{(3)}_{regA_d})}{\widehat{V}(\widehat{P}^{(1)}_{regA_d}) + \widehat{V}(\widehat{P}^{(3)}_{regA_d})}. \tag{4.124}$$

Para $\widehat{P}^{(c_5)}_{regA_d}$,

$$\widehat{P}^{(c_5)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}^{(2)}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(3)}_{regA_d}, \tag{4.125}$$

y su varianza estimada

$$\widehat{AV}(\widehat{P}^{(c_5)}_{regA_d}) = \frac{\widehat{V}(\widehat{P}^{(2)}_{A_d})\widehat{V}(\widehat{P}^{(3)}_{regA_d})}{\widehat{V}(\widehat{P}^{(2)}_{rA_d}) + \widehat{V}(\widehat{P}^{(3)}_{rA_d})}, \tag{4.126}$$

donde $\widehat{V}(\widehat{P}_{A_d})$ está dada por (4.11), $\widehat{V}(\widehat{P}^{(1)}_{regA_d}) = \widehat{V}(\widehat{P}^{(2)}_{regA_d})$ está dada por (4.98) y $\widehat{V}(\widehat{P}^{(3)}_{regA_d})$ está dada por (4.110).

4.2.6. Múltiples atributos auxiliares conocidos a nivel de dominio

Consideraremos conocidos, en el modelo de regresión (4.89), $B = (B_1, \cdots, B_p)$ atributos auxiliares, en el cual b_i, $i = 1, \cdots, p$, se conoce para cada atributo auxiliar o se estima a partir de (4.95). El modelo para estimar P_{A_d} puede escribirse como

$$\widehat{P}_{regA_dM} = \widehat{P}_{A_d} + \sum_{i=1}^{p} b_i(P_{B_id} - \widehat{P}_{B_id}) \tag{4.127}$$

donde $\widehat{P}_{A_d}, \widehat{P}_{B_id}$ son las proporciones muestrales de A en el dominio y B_i en el dominio para el i-ésimo atributo B, respectivamente; P_{B_id} es la proporción poblacional de B_{id}.

Su varianza estimada se obtiene como sigue

$$\begin{aligned} AV(\widehat{P}_{regA_dM}) &= E[\widehat{P}_{regA_dM} - P_{A_d}]^2 \\ &= E[(\widehat{P}_{A_d} - P_{A_d}) + b_1(P_{B_1d} - \widehat{P}_{B_1d}) + \cdots + b_i(P_{B_id} - \widehat{P}_{B_id})]^2 \\ = V(\widehat{P}_{A_d}) + \sum_{i=1}^{p} b_i^2 V(\widehat{P}_{B_id}) &- 2\sum_{i=1}^{p} b_i cov(\widehat{P}_{A_d}, \widehat{P}_{B_id}) + 2\sum_{i\neq k} b_i b_k cov(\widehat{P}_{B_id}, \widehat{P}_{B_kd}). \end{aligned}$$

Entonces,

$$\begin{aligned} AV(\widehat{P}_{regA_dM}) &= V(\widehat{P}_{A_d}) + \sum_{i=1}^{p} b_i^2 V(\widehat{P}_{B_id}) \\ -2\sum_{i=1}^{p} b_i cov(\widehat{P}_{A_d}, \widehat{P}_{B_id}) &+ 2\sum_{i\neq k} b_i b_k cov(\widehat{P}_{B_id}, \widehat{P}_{B_kd}) \end{aligned} \tag{4.128}$$

o de forma matricial

$$AV(\widehat{P}_{regA_dM}) = \mathbf{eVe'}$$

donde $\mathbf{e}= (1, \cdots, 1)$ y $\mathbf{V}= (\nu_{ik}) = cov(\widehat{P}_{A_d}, b_i\widehat{P}_{B_id})$, con $i = 1, \cdots, p$.

Un estimador para la varianza dada por (4.128) se escribe

$$\begin{aligned} \widehat{V}(\widehat{P}_{regA_dM}) &= \widehat{V}(\widehat{P}_{A_d}) + \sum_{i=1}^{p} b_i^2 \widehat{V}(\widehat{P}_{B_id}) \\ -2\sum_{i} b_i \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_id}) &+ 2\sum_{i\neq k} b_i b_k \widehat{cov}(\widehat{P}_{B_id}, \widehat{P}_{B_kd}) \end{aligned} \tag{4.129}$$

Ahora, bajo un diseño $MAS(N, n)$, (4.128) será

$$\begin{aligned} AV(\widehat{P}_{regA_dM}) &= \frac{N}{N_d}\frac{1-f}{n}[P_{A_d}Q_{A_d} + \sum_{i=1}^{p} b_i^2 P_{B_id}Q_{B_id} \\ -2\sum_{i} b_i\phi_i\sqrt{P_{A_d}Q_{A_d}P_{B_id}Q_{B_id}} &+ 2\sum_{i\neq k} b_i b_k \phi_{ik}\sqrt{P_{B_id}Q_{B_id}P_{B_kd}Q_{B_kd}}] \end{aligned} \tag{4.130}$$

y su estimador

$$\begin{aligned} \widehat{V}(\widehat{P}_{regA_dM}) &= \frac{1-f}{n_d}\frac{n}{n-1}[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \sum_{i=1}^{p} b_i^2 \widehat{P}_{B_id}\widehat{Q}_{B_id} \\ -2\sum_{i=1}^{p} b_i\widehat{\phi}_i\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_id}\widehat{Q}_{B_id}} &+ 2\sum_{i\neq k} b_i b_k \widehat{\phi}_{ik}\sqrt{\widehat{P}_{B_id}\widehat{Q}_{B_id}\widehat{P}_{B_kd}\widehat{Q}_{B_kd}}]. \end{aligned} \tag{4.131}$$

Igual que en el caso univariado, b óptimo será de la forma

$$b_{opt} = \frac{cov(\widehat{P}_{A_d}, \widehat{P}_{B_i d})}{V(\widehat{P}_{B_i d})},$$

entonces podemos hallar el estimador óptimo de (4.127) y del estimador de la varianza (4.131) como

$$\widehat{P}^{opt}_{regA_dM} = \widehat{P}_{A_d} + \sum_{i=1}^{p} b_{i_{opt}}(P_{B_i d} - \widehat{P}_{B_i d})$$

y $\widehat{V}(\widehat{P}^{opt}_{regA_dM})$ sustituyendo los correspondientes valores óptimos de b.

Otra opción consiste en considerar pesos ponderados w_i de dominio, en el cual se tienen también p atributos auxiliares B_i, los cuales están disponibles en cada dominio de tamaño N_d. Además, sea R_i, la razón de P_{A_d} y $P_{B_i d}$, con $i = 1, \cdots, p$. Entonces usaremos el siguiente estimador de diferencia ponderado para estimar, de una muestra aleatoria simple de tamaño n, la proporción de individuos que poseen el atributo A dentro del dominio d:

$$\widehat{P}^{pond}_{regA_dM} = \sum_i w_i t_i \tag{4.132}$$

donde $t_i = \widehat{P}_{A_d} - R_i(\widehat{P}_{B_i d} - P_{B_i d})$ y $\widehat{P}_{A_d}, \widehat{P}_{B_i d}$ son las proporciones muestrales de los atributos A y B_i, respectivamente. La cantidad $P_{B_i d}$ es la proporción poblacional de B_i en el dominio y w_i son pesos que satisfacen la condición $\sum_{i=1}^{p} w_i = 1$ y serán calculados de manera que se maximize la precisión del estimador propuesto, es decir, que minimicen la varianza del estimador. El estimador $\widehat{P}^{pond}_{regA_dM}$ es insesgado y su varianza está dada por

$$V(\widehat{P}^{pond}_{regA_dM}) = \sum_i \sum_k w_i w_k cov(t_i, t_k) \tag{4.133}$$

donde

$$cov(t_i, t_k) = cov[\widehat{P}_{A_d} - R_i(\widehat{P}_{B_i d} - P_{B_i d}), \widehat{P}_{A_d} - R_k(\widehat{P}_{B_k d} - P_{B_k d})]$$
$$= V(\widehat{P}_{A_d}) - R_i cov(\widehat{P}_{A_d}, \widehat{P}_{B_i d}) - R_k cov(\widehat{P}_{A_d}, \widehat{P}_{B_k d}) + R_i R_k cov(\widehat{P}_{B_i d}, \widehat{P}_{B_k d})$$

por lo que

$$V(\widehat{p}^{pond}_{regA_dM}) = \sum_i \sum_k w_i w_k [V(\widehat{p}_{A_d})$$
$$-R_i cov(\widehat{p}_{A_d}, \widehat{p}_{B_i d}) - R_k cov(\widehat{p}_{A_d}, \widehat{p}_{B_k d}) + R_i R_k cov(\widehat{p}_{B_i d}, \widehat{p}_{B_k d})]$$
$$V(\widehat{P}^{pond}_{regA_dM}) = \sum_i \sum_k w_i w_k \nu_{ik} = \mathbf{wVw'}, \tag{4.134}$$

donde

$$\nu_{ik} = V(\widehat{P}_{A_d}) - R_i cov(\widehat{P}_{A_d}, \widehat{P}_{B_i d}) - R_k cov(\widehat{P}_{A_d}, \widehat{P}_{B_k d}) + R_i R_k cov(\widehat{P}_{B_i d}, \widehat{P}_{B_k d}),$$

$\mathbf{V} = (\nu_{ik})$ y $\mathbf{w} = (w_1, \cdots, w_p)$, siendo **w'** el traspuesto de **w**.

Un estimador para la varianza será

$$\widehat{V}(\widehat{P}^{pond}_{regA_dM}) = \sum_i \sum_k w_i w_k [\widehat{V}(\widehat{P}_{A_d}) - \widehat{R}_i \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_i d}) - \widehat{R}_k \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_k d}) + \widehat{R}_i \widehat{R}_k \widehat{cov}(\widehat{P}_{B_i d}, \widehat{P}_{B_k d})]. \quad (4.135)$$

Para un diseño $MAS(N, n)$ la varianza dada por (4.134) es

$$V(\widehat{p}^{pond}_{regA_dM}) = \frac{N}{N_d} \frac{1-f}{n} \sum_i \sum_k w_i w_k [P_{A_d} Q_{A_d} - R_i \phi_i \sqrt{P_{A_d} Q_{A_d} P_{B_i d} Q_{B_i d}} - R_k \phi_k \sqrt{P_{A_d} Q_{A_d} P_{B_k d} Q_{B_k d}} + R_i R_k \phi_{ik} \sqrt{P_{B_i d} Q_{B_i d} P_{B_k d} Q_{B_k d}}]$$

$$V(\widehat{P}^{pond}_{regA_dM}) = \frac{N}{N_d} \frac{1-f}{n} \sum_i \sum_k w_i w_k \nu_{ik} = \frac{N}{N_d} \frac{1-f}{n} \mathbf{wVw'}, \quad (4.136)$$

y su estimador

$$\widehat{V}(\widehat{P}^{pond}_{regA_dM}) = \frac{1-f}{n_d} \frac{n}{n-1} \sum_i \sum_k w_i w_k [\widehat{P}_{A_d} \widehat{Q}_{A_d} - \widehat{R}_i \widehat{\phi}_i \sqrt{\widehat{P}_{A_d} \widehat{Q}_{A_d} \widehat{P}_{B_i d} \widehat{Q}_{B_i d}} - \widehat{R}_k \widehat{\phi}_k \sqrt{\widehat{P}_{A_d} \widehat{Q}_{A_d} \widehat{P}_{B_k d} \widehat{Q}_{B_k d}} + \widehat{R}_i \widehat{R}_k \widehat{\phi}_{ik} \sqrt{\widehat{P}_{B_i d} \widehat{Q}_{B_i d} \widehat{P}_{B_k d} \widehat{Q}_{B_k d}}] \quad (4.137)$$

o de forma matricial

$$\widehat{V}(\widehat{p}^{pond}_{regA_dM}) = \left(\frac{1-f}{n_d} \frac{n}{n-1} \right) \sum_i \sum_k w_i w_k \widehat{\nu}_{ik} = \left(\frac{1-f}{n_d} \frac{n}{n-1} \right) \mathbf{w} \widehat{V} \mathbf{w'} \quad (4.138)$$

o sus formas reducidas si suponemos que $n/(n-1) \approx 1$.

Un procedimiento propuesto por Olkin, I. (1958), nos permite establecer que el peso óptimo w_i está dado por

$$w_i = \frac{\text{Suma de los elementos de la i-ésima columna } A^{-1}}{\text{Suma de todos los elementos } i^2 \text{ de } A^{-1}} \quad (4.139)$$

donde A^{-1} es la matriz inversa de A. Usando los pesos óptimos el estimador de la varianza encontrada es

$$\widehat{V}(\widehat{p}_{regA_dM}) = \left(\frac{1-f}{n_d}\frac{n}{n-1}\right) /(\text{Suma de todos los elementos } i^2 \text{ de } A^{-1})$$

o también

$$\widehat{V}(\widehat{p}_{regA_dM}) = \left(\frac{1-f}{n_d}\right) /(\text{Suma de todos los elementos } i^2 \text{ de } A^{-1})$$

si $n/(n-1) \approx 1$.

A partir de los resultados anteriores, el estimador óptimo de P_{A_d} será

$$\widehat{p}^{pond.opt}_{regA_dM} = \sum_i w_i^{opt} t_i \tag{4.140}$$

y un estimador de su varianza estará dada por (4.136), pero con pesos óptimos.

4.3. Estimadores de calibración propuestos

Deville y Särndal (1992), introducen la teoría general de estimadores de calibración para totales poblacionales. El objetivo consiste en estimar el total poblacional, extendiendo una idea de calibración de Lemel (1976), Deville (1988), usada en una población con totales conocidos para modificar los pesos de diseño de muestreo básicos que aparecen en los estimadores de Horvitz-Thompson, por nuevos pesos, tan próximos como sea posible a los pesos de diseño para una métrica dada, que satisfagan una ecuación de calibración. Se minimiza la distancia entre los pesos, se hallan los pesos de calibración y se estima el total usando los pesos calibrados. Theberge (1999), usó la idea de calibración para estimar varianzas; la formulación del problema de calibración es más general que la dada por Deville y Särndal (1992), además de usar medidas arbitrarias de distancia en la estimación. Estevao y Särndal (2000), probaron que las distintas medidas de distancia propuestas en los estimadores de calibración producen resultados aproximadamente idénticos, por esa razón desarrollaron una aproximación alternativa, la forma funcional de los pesos calibrados. Spiegelman, D. et al.(2000), propusieron un estimador eficiente de calibración para regresión logística y otros modelos de regresión lineal generalizados. Kott (2006) definió los pesos de calibración para satisfacer las ecuaciones de calibración y dar un estimador bajo el diseño consistente. Una generalización del procedimiento clásico de calibración la encontramos en Guggemos, F., et al.(2010) y se denomina procedimiento de calibración penalizada. En esta sección abordaremos la estimación de calibración para proporciones a partir de las ideas iniciales y usando la forma funcional de los pesos de calibración.

4.3.1. Estimación de una proporción para el dominio

Como es usual consideremos asociado con A_j un solo atributo auxiliar B_j de forma tal que para los elementos $j \in s$, los vectores (A_j, B_j) son observados y se cumple que $B_j = 1$ si la j-ésima unidad posee el atributo B y $B_j = 0$ en caso contrario. Asumimos conocido el total de B o la proporción P_B a nivel de de dominio, es decir, T_{B_j} o P_{B_j}.

La proporción poblacional del atributo A en el dominio U_d está dado por

$$P_{A_d} = \frac{1}{N_d} \sum_{j \in U_d} A_j \tag{4.141}$$

y si designamos por $d_j = 1/\pi_j$, los pesos obtenidos mediante las probabilidades de inclusión de primer orden (la d con que designamos este peso, no significa

dominio), entonces el estimador de Horvitz-Thompson será

$$\widehat{P}_{A_d} = \frac{1}{\widehat{N}_d} \sum_{j \in s_d} d_j A_j \tag{4.142}$$

que es idéntico al dado por (4.6), en el cual no se incorpora la información auxiliar disponible en la fase de estimación de la proporción de A, a nivel de dominio. Para estimar esta proporción por el método de calibración, lo haremos a partir de la estimación del total de dominio, toda vez que (4.142) es no lineal. Una forma de incorporar la información auxiliar en la estimación de T_{A_d}, consiste en modificar los pesos d_j por pesos nuevos w_j, usando las técnicas de calibración introducidas por Deville y Särndal (1992).

Siguiendo a Deville y Särndal (1992) obtenemos un estimador de calibración para el total del atributo B de la siguiente forma:

Asumimos que

$$T_{B_d} = \sum_{j \in U_d} B_j,$$

se conoce a priori. Dada una muestra s, conocemos B_s y B_{s_d}, los valores del atributo auxiliar para las unidades muestrales y para la submuestra en el dominio de interés. Deseamos hallar los pesos w_j, para $j \in s$, tales que

$$\widehat{T}_{B_d} = \sum_{j \in s_d} w_j B_j = T_{B_d} \tag{4.143}$$

y los w_j's están próximos a los d_j's.

Como una medida de la distancia entre los w_j's y los d_j's, usaremos la distancia Chi-cuadrado y nuestro objetivo es minimizar

$$\phi = \sum_{j \in s_d} \frac{(w_j - d_j)^2}{d_j q_j} \tag{4.144}$$

sujeto a la condición

$$T_{B_d} = \sum_{j \in s_d} w_j B_j$$

donde las q_j son constantes positivas conocidas no relacionadas a las d_j.

Usando el método de multiplicadores de Lagrange para optimización restringida obtenemos la expresión

$$L(w_j, d_j) = \sum_{j \in s_d} \frac{(w_j - d_j)^2}{d_j q_j} - 2 \sum_{j \in s_d} w_j B_j. \tag{4.145}$$

Derivando (4.145) con respecto a w_j e igualando a cero obtenemos

$$w_j = d_j + \lambda d_j q_j B_j. \tag{4.146}$$

Ahora, si multiplicamos (4.146) por B_j y sumamos sobre la submuestra conseguimos

$$\sum_{j\in s_d} w_j B_j = \sum_{j\in s_d} d_j B_j + \lambda \sum_{j\in s_d} d_j q_j B_j^2$$

$$T_{B_d} = \widehat{T}_{B_d} + \lambda \sum_{j\in s_d} d_j q_j B_j$$

donde $B_j^2 = B_j$, por definición de B_j. Entonces resolviendo para λ se obtiene

$$\lambda = \frac{T_{B_d} - \widehat{T}_{B_d}}{\sum_{j\in s_d} d_j q_j B_j},$$

por lo que

$$\widehat{T}_{A_d w} = \sum_{j\in s_d} w_j A_j = \sum_{j\in s_d} \left[d_j + \left(\frac{T_{B_d} - \widehat{T}_{B_d}}{\sum_{j\in s_d} d_j q_j B_j} \right) d_j q_j B_j \right] A_j$$

$$\widehat{T}_{A_d w} = \sum_{j\in s_d} d_j A_j + \sum_{j\in s_d} \left(\frac{T_{B_d} - \widehat{T}_{B_d}}{\sum_{j\in s_d} d_j q_j B_j} \right) d_j q_j B_j A_j$$

$$\widehat{T}_{A_d w} = \widehat{T}_{A_d} + \frac{(T_{B_d} - \widehat{T}_{B_d})}{\sum_{j\in s_d} d_j q_j B_j} \sum_{j\in s_d} d_j q_j B_j A_j. \tag{4.147}$$

El estimador dado por (4.147) tiene la forma de un estimador de diferencia generalizado, donde

$$b = \frac{\sum_{j\in s_d} d_j q_j B_j A_j}{\sum_{j\in s_d} d_j q_j B_j}$$

por lo que su varianza aproximada estará dada por

$$AV(\widehat{T}_{A_d w}) = V(\widehat{T}_{A_d}) + b^2 V(\widehat{T}_{B_d}) - 2bcov(\widehat{T}_{A_d}, \widehat{T}_{B_d}) \tag{4.148}$$

y su estimador

$$\widehat{AV}(\widehat{T}_{A_d w}) = \widehat{V}(\widehat{T}_{A_d}) + \widehat{b}^2 \widehat{V}(\widehat{T}_{B_d}) - 2\widehat{b}\widehat{cov}(\widehat{T}_{A_d}, \widehat{T}_{B_d}) \tag{4.149}$$

Asumiendo un diseño $MAS(N, n)$ con $q_j = 1\ \forall j \in U$, el estimador dado por (4.147) se escribe

$$\widehat{T}_{A_d w} = \widehat{T}_{A_d} + \frac{\widehat{T}_{AB_d}}{\widehat{T}_{B_d}} (T_{B_d} - \widehat{T}_{B_d}), \tag{4.150}$$

donde $\widehat{T}_{AB_d} = \sum_{j \in s_d} A_j B_j$.

Su varianza aproximada será

$$AV(\widehat{T}_{A_d w}) = N_d^2 \left[V(\widehat{P}_{A_d}) + \left(\frac{P_{AB_d}}{P_{B_d}} \right)^2 V(\widehat{P}_{B_d}) - 2 \left(\frac{P_{AB_d}}{P_{B_d}} \right) cov(\widehat{P}_{A_d}, \widehat{P}_{B_d}) \right]$$

con $P_{AB_d} = \frac{1}{N_d} \sum_{j \in U_d} A_j B_j$. De forma equivalente podemos escribir

$$AV(\widehat{T}_{A_d w}) = N_d^2 \frac{N}{N_d} \frac{1-f}{n} [P_{A_d} Q_{A_d} + \left(\frac{P_{AB_d}}{P_{B_d}} \right)^2 P_{B_d} Q_{B_d} - 2 \left(\frac{P_{AB_d}}{P_{B_d}} \right) \phi_d \sqrt{P_{A_d} Q_{A_d} P_{B_d} Q_{B_d}}] \quad (4.151)$$

y su estimador

$$\widehat{AV}(\widehat{T}_{A_d w}) = N_d^2 \frac{n}{n-1} \frac{1-f}{n_d} [\ \widehat{P}_{A_d} \widehat{Q}_{A_d} + \left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}} \right)^2 \widehat{P}_{B_d} \widehat{Q}_{B_d} - 2 \left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}} \right) \widehat{\phi}_d \sqrt{\widehat{P}_{A_d} \widehat{Q}_{A_d} \widehat{P}_{B_d} \widehat{Q}_{B_d}}\ ] \quad (4.152)$$

donde ϕ_d, el coeficiente V de Cramer, se estima de la submuestra s_d.

Sabemos que $T_{A_d w} = N_d P_{A_d w}$ y como nos interesa estimar la proporción obtenemos que

$$\widehat{P}_{A_d w} = \frac{1}{N_d} \widehat{T}_{A_d w} \quad (4.153)$$

si N_d es conocido y

$$\widehat{P}_{A_d w} = \frac{1}{\widehat{N}_d} \widehat{T}_{A_d w} \quad (4.154)$$

si N_d no es conocido, por lo que estamos ante la presencia de un estimador de razón de totales.

La varianza de (4.153) será

$$AV(\widehat{P}_{A_d w}) = V[\frac{1}{N_d} \widehat{T}_{A_d w}] = \left(\frac{1}{N_d} \right)^2 V[\widehat{T}_{A_d w}]$$

$$= \left(\frac{N}{N_d} \right)^2 \left[V(\widehat{P}_{A_d}) + b^2 V(\widehat{P}_{B_d}) - 2bcov(\widehat{P}_{A_d}, \widehat{P}_{B_d}) \right]$$

y su estimador

$$\widehat{AV}(\widehat{P}_{A_d w}) = \left(\frac{N}{N_d}\right)^2 \left[\widehat{V}(\widehat{P}_{A_d}) + \widehat{b}^2\widehat{V}(\widehat{P}_{B_d}) - 2\widehat{b}\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_d})\right]. \quad (4.155)$$

Para (4.154), toda vez que es una razón de estimadores, su varianza aproximada estará dada por

$$AV[\widehat{P}_{A_d w}] = \left(\frac{N}{N_d}\right)^2 \left[V(\widehat{P}_{A_d}) + b^2 V(\widehat{P}_{B_d}) - 2bcov(\widehat{P}_{A_d}, \widehat{P}_{B_d})\right] \quad (4.156)$$

y su estimador

$$\widehat{AV}[\widehat{P}_{A_d w}] = \left(\frac{N}{\widehat{N}_d}\right)^2 \left[\widehat{V}(\widehat{P}_{A_d}) + \widehat{b}^2\widehat{V}(\widehat{P}_{B_d}) - 2\widehat{b}\widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_d})\right]. \quad (4.157)$$

Bajo un diseño $MAS(N, n)$ el estimador de la varianza dado por (4.155) se escribe como

$$\widehat{AV}[\widehat{P}_{A_d w}] = \left(\frac{N}{N_d}\right)^2 \frac{1-f}{n}\frac{N_d - 1}{N-1}\frac{n_d}{n_d - 1}[\ \widehat{P}_{A_d}\widehat{Q}_{A_d} + \left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}}\right)^2 \widehat{P}_{B_d}\widehat{Q}_{B_d} - 2\left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}}\right)\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\ ]$$

$$\widehat{AV}[\widehat{P}_{A_d w}] = \frac{N}{N_d}\frac{1-f}{n}\frac{n_d}{n_d - 1}\left[\widehat{P}_{A_d}\widehat{Q}_{A_d} + \left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}}\right)^2 \widehat{P}_{B_d}\widehat{Q}_{B_d} - 2\left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}}\right)\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\ \right] \quad (4.158)$$

y el estimador de la varianza dado por (4.157)

$$\widehat{AV}[\widehat{P}_{A_d w}] = \frac{1-f}{n_d}\frac{n}{n-1}[\ \widehat{P}_{A_d}\widehat{Q}_{A_d} + \left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}}\right)^2 \widehat{P}_{B_d}\widehat{Q}_{B_d} - 2\left(\frac{\widehat{P}_{AB_d}}{\widehat{P}_{B_d}}\right)\widehat{\phi}_d\sqrt{\widehat{P}_{A_d}\widehat{Q}_{A_d}\widehat{P}_{B_d}\widehat{Q}_{B_d}}\ ]. \quad (4.159)$$

4.3.2. Estimador sintético de calibración

Sabemos ya que un estimador sintético de dominio se obtiene usando un estimador directo apropiado para un área grande, bajo el supuesto que las áreas pequeñas tienen las mismas características que el área grande. Entonces el estimador sintético será de la forma

$$\widehat{T}_{A_dw} = \widehat{T}_A + \frac{(T_B - \widehat{T}_B)}{\sum_{j\in s} d_j q_j B_j} \sum_{j\in s} d_j q_j B_j A_j \tag{4.160}$$

es decir, tiene la forma de un estimador de diferencia generalizado a partir de la muestra s, con $b = \sum_{j\in s} d_j q_j B_j A_j / \sum_{j\in s} d_j q_j B_j$.

La varianza aproximada del estimador es de la forma

$$AV(\widehat{T}_{A_dw}) = V(\widehat{T}_A) + b^2 V(\widehat{T}_B) - 2bCov(\widehat{T}_A, \widehat{T}_B) \tag{4.161}$$

y un estimador de su varianza

$$\widehat{AV}(\widehat{T}_{A_dw}) = \widehat{V}(\widehat{T}_A) + \widehat{b}^2 \widehat{V}(\widehat{T}_B) - 2\widehat{b}\widehat{Cov}(\widehat{T}_A, \widehat{T}_B). \tag{4.162}$$

Bajo un diseño de muestreo $MAS(N, n)$, con $q_j = 1$, el estimador de su varianza será

$$\widehat{AV}(\widehat{T}_{A_dw}) = N^2 \frac{1-f}{n-1} \Bigg[\widehat{P}_A \widehat{Q}_A + \left(\frac{\widehat{P}_{AB}}{\widehat{P}_B} \right)^2 \widehat{P}_B \widehat{Q}_B - 2 \left(\frac{\widehat{P}_{AB}}{\widehat{P}_B} \right) \widehat{\phi} \sqrt{\widehat{P}_A \widehat{Q}_A \widehat{P}_B \widehat{Q}_B} \Bigg] \tag{4.163}$$

donde $\widehat{P}_{AB} = n^{-1} \sum_{j\in s} A_j B_j$ y ϕ es el coeficiente V de Cramer en la muestra.

Los estimadores de la proporción de dominio y su varianza pueden obtenerse si sabemos que $\widehat{T}_{A_dw} = N_d \widehat{P}_{A_dw}$, entonces

$$\widehat{P}_{A_dw} = \frac{1}{N_d} \widehat{T}_{A_dw} \tag{4.164}$$

si se asume N_d conocido.

La varianza de (4.164) será

$$AV(\widehat{P}_{A_dw}) = V[\frac{1}{N_d} \widehat{T}_{A_dw}] = \left(\frac{1}{N_d} \right)^2 V[\widehat{T}_{A_dw}]$$

$$= V(\widehat{P}_A) + b^2 V(\widehat{P}_B) - 2bcov(\widehat{P}_A, \widehat{P}_B)$$

y su estimador

$$\widehat{AV}(\widehat{P}_{A_dw}) = \widehat{V}(\widehat{P}_A) + \widehat{b}^2\widehat{V}(\widehat{P}_B) - 2\widehat{b}\widehat{cov}(\widehat{P}_A, \widehat{P}_B). \tag{4.165}$$

Bajo un diseño $MAS(N,n)$ y $q_j = 1$, el estimador de la varianza dado por (4.165) se escribe como

$$\widehat{AV}[\widehat{P}_{A_dw}] = \frac{1-f}{n-1}\left[\widehat{P}_A\widehat{Q}_A + \left(\frac{\widehat{P}_{AB}}{\widehat{P}_B}\right)^2 \widehat{P}_B\widehat{Q}_B \right.$$
$$\left. -2\left(\frac{\widehat{P}_{AB}}{\widehat{P}_B}\right)\widehat{\phi}\sqrt{\widehat{P}_A\widehat{Q}_A\widehat{P}_B\widehat{Q}_B}\ \right] \tag{4.166}$$

donde $\widehat{P}_{AB} = n^{-1}\sum_{j\in s} A_jB_j$ y ϕ es el coeficiente V de Cramer en la muestra.

Al igual que en los casos anteriores, es posible construir un estimador combinado de estos estimadores.

4.3.3. Estimación calibrada para el caso multivariado

Consideremos ahora un atributo A relacionado con p atributos auxiliares $B_1, \cdots, B_p$, el objetivo es hallar el estimador calibrado del total de dominio T_{A_d} en presencia de dichos atributos auxiliares, a partir de los resultados de Deville y Särndal (1992).

Siguiendo el procedimiento usual de calibración, debemos incorporar la información auxiliar proporcionada por los p atributos auxiliares, considerando nuevos pesos w_j, sujetos a las siguientes condiciones

$$T_{B_id} = \sum_{j\in U_d} B_{ij} = \sum_{j\in s_d} w_jB_{ij} \qquad i = 1, \cdots, p.$$

Sea $\mathbf{T}'_{B_id} = (T_{B_1d}, \cdots, T_{B_pd})$ el vector de totales de cada uno de los p atributos auxiliares, los cuales se asumen conocidos; $\widehat{\mathbf{T}}'_{B_id} = (\widehat{T}_{B_1d}, \cdots, \widehat{T}_{B_pd})$ el vector de estimadores de Horvitz-Thompson de los totales para cada uno de los p atributos auxiliares a nivel de dominio, donde la j-ésima observación tendrá asociado el vector $\mathbf{B}'_d = (B_{1j}, \cdots, B_{pj})$ y el estimador de de H-T del total para el i-ésimo atributo auxiliar será

$$\widehat{T}_{B_id} = \sum_{j\in s_d} d_jB_{ij} \qquad i = 1, \cdots, p.$$

El estimador calibrado del total será de la forma

$$\widehat{T}_{A_dw} = \sum_{j\in s_d} w_jA_j. \tag{4.167}$$

Minimicemos la distancia chi-cuadrado entre los pesos iniciales y los nuevos pesos, bajo las condiciones dadas, usando el método de multiplicadores de Lagrange para optimización restringida y los nuevos pesos serán

$$w_j = d_j + (\mathbf{T}_{B_id} - \widehat{\mathbf{T}}_{B_id})' \left(\sum_{j \in s_d} d_j B_j B_j' \right)^{-1} d_j B_j, \tag{4.168}$$

entonces el estimador calibrado con los nuevos pesos será

$$\widehat{T}_{A_dw} = \sum_{j \in s_d} d_j A_j + (\mathbf{T}_{B_id} - \widehat{\mathbf{T}}_{B_id})' \left(\sum_{j \in s_d} d_j B_j B_j' \right)^{-1} \sum_{j \in s_d} d_j B_j A_j$$

o de forma equivalente

$$\widehat{T}_{A_dw} = \widehat{T}_{A_d} + (\mathbf{T}_{B_id} - \widehat{\mathbf{T}}_{B_id})' \widehat{b} \tag{4.169}$$

donde $\widehat{b} = (\sum_{j \in s_d} d_j B_j B_j')^{-1} \sum_{j \in s_d} d_j B_j A_j$, B_j es una matriz de dimensión $p \times n_d$, A_j es un vector de $n_d \times 1$, la matriz $(\sum_{j \in s_d} d_j B_j B_j')^{-1}$ es de dimensión $p \times p$ y no singular para que el estimador pueda obtenerse.

Una expresión para la varianza del estimador será

$$\begin{aligned} AV(\widehat{T}_{A_dw}) = V(\widehat{T}_{A_d}) + b^2 \sum_i V(\widehat{\mathbf{T}}_{B_id}) - 2b \sum_i cov(\widehat{T}_{A_d}, \widehat{\mathbf{T}}_{B_id}) \\ +2b \sum_{i \neq k} cov(\widehat{\mathbf{T}}_{B_id}, \widehat{\mathbf{T}}_{B_kd}) \qquad i = 1, \cdots, p \text{ y } i \neq k, \end{aligned} \tag{4.170}$$

donde $b = (\sum_{j \in U_d} d_j B_j B_j')^{-1} \sum_{j \in U_d} d_j B_j A_j$.

Un estimador de la varianza será

$$\begin{aligned} \widehat{AV}(\widehat{T}_{A_dw}) = \widehat{V}(\widehat{T}_{A_d}) + \widehat{b}^2 \sum_i V(\widehat{\mathbf{T}}_{B_id}) - 2\widehat{b} \sum_i \widehat{cov}(\widehat{T}_{A_d}, \widehat{\mathbf{T}}_{B_id}) \\ +2\widehat{b} \sum_{i \neq k} \widehat{cov}(\widehat{\mathbf{T}}_{B_id}, \widehat{\mathbf{T}}_{B_kd}) \qquad i = 1, \cdots, p \text{ y } i \neq k, . \end{aligned} \tag{4.171}$$

El objetivo es estimar la proporción del atributo A relacionado a los $B_1, \cdots, B_p$ atributos auxiliares, entonces

$$\begin{aligned} \widehat{P}_{A_dw} = \frac{\widehat{T}_{A_dw}}{N_d} = \frac{1}{N_d} \left[\widehat{T}_{A_d} + (\mathbf{T}_{B_id} - \widehat{\mathbf{T}}_{B_id})' \, \widehat{b} \right] \\ \widehat{P}_{A_dw} = \widehat{P}_{A_d} + (P_{B_id} - \widehat{P}_{B_id})' \, \widehat{b} \end{aligned} \tag{4.172}$$

si N_d es conocido y

$$\widehat{P}_{A_dw} = \frac{\widehat{T}_{A_dw}}{\widehat{N}_d} = \frac{1}{\widehat{N}_d}\left[\widehat{T}_{A_d} + (\mathbf{T}_{B_id} - \widehat{\mathbf{T}}_{B_id})'\ \widehat{b}\right]$$
$$\widehat{P}_{A_dw} = \frac{N_d}{\widehat{N}_d}\left[\widehat{P}_{A_d} + (P_{B_id} - \widehat{P}_{B_id})'\ \widehat{b}\right] \tag{4.173}$$

si N_d es desconocido.

Las expresiones para las varianzas de (4.172) y (4.173) serán

$$AV(\widehat{P}_{A_dw}) = V(\widehat{P}_{A_d}) + b^2 \sum_i V(\widehat{P}_{B_id}) - 2b \sum_i cov(\widehat{P}_{A_d}, \widehat{P}_{B_id}) + 2b \sum_{i \neq k} cov(\widehat{P}_{B_id}, \widehat{P}_{B_kd}),$$

$$AV(\widehat{P}_{A_dw}) = \left(\frac{N_d}{\widehat{N}_d}\right)^2 [V(\widehat{P}_{A_d}) + b^2 \sum_i V(\widehat{P}_{B_id})$$
$$-2b \sum_i cov(\widehat{P}_{A_d}, \widehat{P}_{B_id}) + 2b \sum_{i \neq k} cov(\widehat{P}_{B_id}, \widehat{P}_{B_kd})],$$

respectivamente, y sus estimadores

$$\widehat{AV}(\widehat{P}_{A_dw}) = \widehat{V}(\widehat{P}_{A_d}) + \widehat{b}^2 \sum_i \widehat{V}(\widehat{P}_{B_id}) - 2\widehat{b} \sum_i \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_id}) + 2\widehat{b} \sum_{i \neq k} \widehat{cov}(\widehat{P}_{B_id}, \widehat{P}_{B_kd}), \tag{4.174}$$

$$\widehat{AV}(\widehat{P}_{A_dw}) = \left(\frac{N_d}{\widehat{N}_d}\right)^2 [\widehat{V}(\widehat{P}_{A_d}) + \widehat{b}^2 \sum_i \widehat{V}(\widehat{P}_{B_id}) - 2\widehat{b} \sum_i \widehat{cov}(\widehat{P}_{A_d}, \widehat{P}_{B_id})$$
$$+2\widehat{b} \sum_{i \neq k} \widehat{cov}(\widehat{P}_{B_id}, \widehat{P}_{B_kd})] \tag{4.175}$$

4.3.4. Estimación de proporciones a partir de variables instrumentales

Consideraremos ahora el enfoque propuesto por Estevao y Särndal (2000), que consiste en modificar el requerimiento de minimización de la medida de distancia en la estimación de calibración y adoptar la forma funcional de los pesos de calibración

$$w_j = d_j(1 + \lambda^T z_j) \tag{4.176}$$

para algún vector instrumental z_j, donde λ se determina de la restricción.

La estimación del total poblacional se hace de forma similar a la utilizada hasta ahora y se obtiene la expresión

$$\widehat{T}_w = \widehat{T}_A + (T_B - \widehat{T}_B)\widehat{b}_z \tag{4.177}$$

donde $\widehat{b}_z = (\sum_d d_j z_j \mathbf{B}_j^T)^{-1} \sum_d z_j A_j$.

Como antes, si $\mathbf{B}_d = (B_{1j}, \cdots, B_{pj})^T$ el vector de atributos auxiliares para la j-ésima unidad. La información auxiliar consiste del vector de totales $T_{B_d} = \sum_{j \in U_d} B_j$, compuesto de los p totales conocidos $\sum_{U_d} B_{ij}$ para $i = 1, \cdots, p$. Adicionalmente definimos un vector instrumental $z_j = (z_{1j}, \cdots, z_{pj})^T$ para todo $j \in s$, tal que

(a) $dim(z_j) = p = dim(\mathbf{B}_j)$, y

(b) la matriz $\sum_s d_j z_j \mathbf{B}_j^T$, de dimensión $p \times p$ es no singular.

En el presente trabajo estudiaremos la estimación del total y proporciones de áreas pequeñas usando como variables instrumentales $z_j = (1, \mathbf{B}_j)$ y $z_j = (X_j, \mathbf{B}_j)$.

Para la estimación a nivel de dominio debemos hallar los pesos calibrados $w_{d.j}$ que satisfagan las ecuaciones de calibración

$$\sum_{j \in s_d} w_{d.j} z_j = \sum_{j \in U_d} z_j = z_{U_d} \text{ donde } z_j = (1, \mathbf{B}_j)^T$$

o de forma equivalente

$$\sum_{j \in s_d} w_{d.j} = N_d \text{ y } \sum_{j \in s_d} w_{d.j} B_j = \sum_{j \in U_d} B_j.$$

Entonces, el estimador de calibración para el total a nivel de dominio será

$$\widehat{T}_{A_d w} = \widehat{T}_{A_d} + (T_{B_d} - \widehat{T}_{B_d})^T \widehat{b}_z \tag{4.178}$$

donde $\widehat{b}_z = (\sum_{j \in s_d} d_j z_j \mathbf{B}_d^T)^{-1} \sum_{j \in s_d} d_j z_j A_j$.

En este caso la variable instrumental $z_j = \mathbf{B}_j$, por lo que el problema se reduce al caso univariado que hemos trabajado en el apartado anterior y los estimadores del total y su varianza, de la proporción y su varianza a nivel de dominio, estarían dadas por (4.150) y (4.152), (4.153) y (4.155) para N_d conocido y (4.154) y (4.157) para N_d desconocido.

Ahora, asumimos que las ecuaciones de calibración son

$$\sum_{j \in s} w_j z_j = \sum_{j \in U} z_j = z_U \text{ donde } z_j = (X_j, \mathbf{B}_j)^T$$

que pueden ser expresadas como

$$\sum_{j\in s} w_j = N_d \text{ y } \sum_{j\in s} w_j B_j = \sum_{j\in U} B_j,$$

entonces, el estimador calibrado del total será

$$\widehat{T}_{A_w} = \widehat{T}_A + (T_B - \widehat{T}_B)^T \widehat{b}_z, \tag{4.179}$$

donde $\widehat{b}_z = (\sum_{j\in s} d_j z_j \mathbf{B}_j^T)^{-1} \sum_{j\in s} d_j z_j A_j$, y entonces, el estimador calibrado (4.179) se reduce al estimador sintético del total de dominio.

Una expresión para la varianza será

$$AV(\widehat{T}_{A_w}) = N^2[V(\widehat{P}_A) + b_z^2 V(\widehat{P}_B) - 2bcov(\widehat{P}_A, \widehat{P}_B)]$$

y su estimador

$$\widehat{AV}(\widehat{T}_{A_w}) = N^2[\widehat{V}(\widehat{P}_A) + \widehat{b}_z^2 \widehat{V}(\widehat{P}_B) - 2\widehat{b}\widehat{cov}(\widehat{P}_A, \widehat{P}_B)]. \tag{4.180}$$

El estimador de la proporción a nivel de dominio será

$$\widehat{P}_{A_w} = \widehat{P}_{A_d} + (P_{B_d} - \widehat{P}_{B_d})\widehat{b}_z \tag{4.181}$$

teniendo en cuenta que N_d puede ser conocido o no, en la estimación de P_{A_d} y P_{B_d}.

Siguiendo los trabajos de Estevao y Särndal (2006), Lehtonen, Särndal y Veijanen (2008) y Kim y Park (2009), podemos usar la variable instrumental $z_j = (1, \widehat{A}_j)^T$, donde $\widehat{A}_j$ se estima por regresión logística binaria. El estimador del total a nivel de dominio será

$$\widehat{T}_{A_w} = \widehat{T}_{A_d} + (T_{B_d} - \widehat{T}_{B_d})^T \widehat{b}_z, \tag{4.182}$$

donde los pesos calibrados deben satisfacer la ecuación de calibración

$$\sum_{j\in s_d} w_{d.j} z_j = \sum_{j\in U_d} z_j = z_{U_d}, \text{ con } z_j = (1, \widehat{A}_j)^T$$

y las restricciones de calibración

$$\sum_{j\in s_d} w_{dj} = N_d \text{ y } \sum_{j\in s_d} w_{dj}\widehat{A}_j = \sum_{j\in U_d} \widehat{A}_j = \widehat{T}_{A_j}.$$

Para este caso b se estima por

$$\widehat{b}_z = \left[(\widehat{A}_1, \cdots, \widehat{A}_{n_d}) \begin{pmatrix} B_1 \\ \vdots \\ B_{n_d} \end{pmatrix}\right]^{-1} (\widehat{A}_1, \cdots, \widehat{A}_{n_d}) \begin{pmatrix} A_1 \\ \vdots \\ A_{n_d} \end{pmatrix}.$$

Para hallar $\widehat{A}_j$, asumimos que $A_j = A_{dj}|u_d \sim Bin(n, P_{A_j})$ y denotemos por $\mu_{A_j} = P_{A_j}$ y $\sigma_{A_j} = P_{A_j}(1 - P_{A_j})$, la media y varianza de A_j dado μ_j, respectivamente. La distribución condicional de A_j pertenece a la familia exponencial natural, y entonces

$$P_{A_j} = \frac{\exp\{\eta_j\}}{1 + \exp\{\eta_j\}}, \quad \eta_j = \mathbf{B}_j\boldsymbol{\beta} + u_d, \quad j = 1, \cdots, N_d,$$

que tendrá por estimador

$$\widehat{P}_{A_j} = \frac{\exp\{\widehat{\eta}_j\}}{1 + \exp\{\widehat{\eta}_j\}}, \quad \widehat{\eta}_j = \mathbf{B}_j\widehat{\boldsymbol{\beta}}, \quad j = 1, \cdots, N_d,$$

donde $\boldsymbol{\beta}$ se estima por mínimos cuadrados ponderados o por máxima verosimilitud iterativa con Newton-Raphson de la muestra s.

Ahora, $\widehat{P}_{A_j} = \widehat{\mu}_{A_j} = \widehat{A}_j$ es la probabilidad de que el individuo j tome el valor 0 o 1 y estamos en condiciones, ahora, de estimar el total y la proporción a nivel de dominio usando (4.182).

Capítulo 5

Estudio de simulación. Aplicación a datos de dengue

5.1. Descripción del estudio de simulación

Descripción de la población simulada

Con el fin de comprobar el comportamiento real de los estimadores propuestos se ha realizado un estudio completo de simulación. Centramos nuestro estudio en una población simulada que denominamos PopSIM. La población simulada de tamaño $N = 30000$ está dividida en 30 dominios (6 de tamaño 500, 6 de tamaño 750, 6 de tamaño 1000, 6 de tamaño 1250 y 6 de tamaño 1500) de forma que estos incluyan diversos escenarios en función de las proporciones poblacionales y el coeficiente V de Cramer entre la variable de interés y las variables auxiliares. En concreto, las poblaciones correspondientes a cada dominio se generaron de forma que, $x_i \sim B(N, p), y_{is} \sim B(N, p)$ son independientes e $y_i = x_i y_{is}$, donde p varía entre $\frac{10}{41}$ y $\frac{39}{41}$. Ver tabla (5.1).

Descripción del proceso de simulación

El procedimiento a seguir para esta población será realizar 1000 iteraciones del siguiente proceso:

- Seleccionar una muestra aleatoria simple de tamaño 900.
- Evaluar los estimadores propuestos así como el estimador directo para el dominio de interés considerado y una estimación de su varianza.

Los resultados de las 1000 iteraciones nos permiten evaluar la eficiencia relativa porcentual (RE) con respecto al estimador directo para el dominio,

Tabla 5.1: Características de la población simulada.
Población generada U, de tamaño $N = 30000$, con $P_A = 0{,}4882$, $P_B = 0{,}6706$ y $\phi = 0{,}68456$. Tamaño N_d, proporción del atributo de interés P_{A_d}, del atributo auxiliar P_{B_d} y coeficiente de Cramer ϕ_d entre atributos por dominio U_d

U_d	N_d	P_{A_d}	P_{B_d}	ϕ_d
U_1	500	0.0400000	0.2340000	0.3693183
U_2	500	0.0660000	0.2760000	0.4305398
U_3	500	0.1040000	0.2980000	0.5229054
U_4	500	0.0820000	0.3000000	0.4565349
U_5	500	0.1120000	0.3620000	0.4714749
U_6	500	0.1320000	0.3780000	0.5002377
U_7	750	0.1720000	0.3826667	0.5788934
U_8	750	0.1920000	0.4333333	0.5574395
U_9	750	0.2000000	0.4546667	0.5475887
U_{10}	750	0.1973333	0.4786667	0.5174564
U_{11}	750	0.2373333	0.4733333	0.5884321
U_{12}	750	0.2466667	0.5066667	0.5646388
U_{13}	1000	0.2980000	0.5610000	0.5763555
U_{14}	1000	0.3350000	0.5710000	0.6152083
U_{15}	1000	0.3660000	0.5920000	0.6307614
U_{16}	1000	0.3410000	0.5850000	0.6058716
U_{17}	1000	0.4040000	0.6450000	0.6108040
U_{18}	1000	0.4310000	0.6520000	0.6358407
U_{19}	1250	0.4672000	0.6784000	0.6447397
U_{20}	1250	0.4824000	0.6968000	0.6368205
U_{21}	1250	0.5544000	0.7424000	0.6570418
U_{22}	1250	0.5744000	0.7592000	0.6542693
U_{23}	1250	0.5840000	0.7776000	0.6336494
U_{24}	1250	0.6768000	0.8128000	0.6944733
U_{25}	1500	0.6653333	0.8220000	0.6561264
U_{26}	1500	0.7440000	0.8546667	0.7029925
U_{27}	1500	0.7646667	0.8700000	0.6967972
U_{28}	1500	0.8073333	0.8946667	0.7023853
U_{29}	1500	0.8513333	0.9193333	0.7088487
U_{30}	1500	0.8973333	0.9433333	0.7245918

definida como la razón entre los errores cuadráticos medios empíricos de los estimadores directo y el estimador con el cual se compara y, el sesgo relativo porcentual (RB) de cada uno de los estimadores evaluados, definido a partir de la razón entre el sesgo promedio en la estimación de cada estimador y la proporción estimada del atributo A en el dominio. Es decir,

$RE =$ ECME (directo)/ECME (estimador)$\times 100$.

$RB =$ SE (estimador)$/P_{A_d} \times 100$.

Así, para cada simulación r, para cada dominio d y para cada estimador E, se calcula:

- $n_d^{(r)}$, tamaño de muestra en el dominio,
- $\widehat{E}_d^{(r)}$, estimación en el dominio de P_A,
- $\widehat{E}_d^{(r)} - P_A$, sesgo en la estimación,
- $(\widehat{E}_d^{(r)} - P_A)^2$, cuadrado del sesgo en la estimación,
- $\widehat{V}(\widehat{E}_d^{(r)})$, estimación de la varianza del estimador.

Se repite el proceso anterior para $r = 1, ..., R = 1000$ y se promedian sobre estas R repeticiones los cálculos anteriores, para definir, en cada dominio:

- $\overline{n}_d = \sum_{r=1}^{R} n_d^{(r)}/R$, tamaño muestral medio,
- $\overline{E}_d = \sum_{r=1}^{R} \widehat{E}_d^{(r)}/R$, estimación media en el dominio de P_A,
- $\text{SE} = \sum_{r=1}^{R} (\widehat{E}_d^{(r)} - P_A)/R$, sesgo empírico en la estimación,
- $\text{RECME} = (\sum_{r=1}^{R} (\widehat{E}_d^{(r)} - P_A)^2/R)^{1/2}$, raíz del error cuadrático medio empírico, y
- $\text{REVE} = (\sum_{r=1}^{R} \widehat{V}(\widehat{E}_d^{(r)})/R)^{1/2}$, raíz de la estimación de la varianza empírica.

En las tablas de resultados se han incluido el tamaño muestral medio, $\overline{n}_d$, la estimación promedio, $\overline{E}_d$, el sesgo empírico, SE, la raíz del error cuadrático medio empírico, RECME y la raíz de la estimación de la varianza empírica, REVE, por dominio U_d. Además, también se incluyen los promedios en la población completa y según el tamaño del dominio.

El estimador directo

Resultados

La tabla 5.2 muestra los resultados para el estimador directo.

Tabla 5.2: Resultados de la simulación para el estimador directo, $\widehat{P}_A = p_A$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.090	0.050	0.065	0.075
U_2	500	15.728	0.104	0.038	0.065	0.079
U_3	500	15.366	0.132	0.028	0.074	0.087
U_4	500	15.449	0.113	0.031	0.066	0.082
U_5	500	15.250	0.138	0.026	0.079	0.089
U_6	500	15.157	0.152	0.020	0.081	0.092
U_7	750	22.503	0.177	0.005	0.077	0.079
U_8	750	22.146	0.191	-0.001	0.082	0.082
U_9	750	22.511	0.202	0.002	0.081	0.083
U_{10}	750	22.484	0.204	0.007	0.082	0.084
U_{11}	750	22.187	0.240	0.002	0.088	0.089
U_{12}	750	22.284	0.246	-0.001	0.090	0.090
U_{13}	1000	29.771	0.294	-0.004	0.081	0.082
U_{14}	1000	29.846	0.331	-0.004	0.085	0.085
U_{15}	1000	30.055	0.368	0.002	0.090	0.086
U_{16}	1000	29.854	0.338	-0.003	0.086	0.085
U_{17}	1000	29.817	0.404	0.000	0.091	0.088
U_{18}	1000	29.853	0.433	0.002	0.093	0.089
U_{19}	1250	37.726	0.467	-0.001	0.079	0.080
U_{20}	1250	37.180	0.481	-0.002	0.081	0.081
U_{21}	1250	37.701	0.552	-0.002	0.080	0.080
U_{22}	1250	37.294	0.570	-0.004	0.081	0.080
U_{23}	1250	37.707	0.584	0.000	0.080	0.079
U_{24}	1250	37.483	0.675	-0.002	0.077	0.075
U_{25}	1500	44.443	0.664	-0.001	0.069	0.070
U_{26}	1500	44.686	0.746	0.002	0.066	0.064
U_{27}	1500	44.952	0.766	0.001	0.061	0.062
U_{28}	1500	45.303	0.806	-0.001	0.060	0.058
U_{29}	1500	44.985	0.850	-0.001	0.052	0.053
U_{30}	1500	44.825	0.892	-0.005	0.044	0.046
Todos		30	0.407	0.006	0.076	0.079
Pequeño		15.401	0.121	0.032	0.072	0.084
Mediano		29.911	0.375	0.000	0.084	0.083
Grande		44.866	0.787	-0.001	0.059	0.059

Comentarios

El estimador directo de dominio p_A (4.9) será considerado como el estimador base, por lo que con respecto a él compararemos nuestros estimadores, con objeto de obtener una medida de eficiencia. Como puede observarse en los resultados de la tabla 5.2, el mayor sesgo empírico en valor absoluto es de

0.05, es decir, es muy poco sesgado y con precisión moderada toda vez que la mayor raíz cuadrada del error cuadrático medio empírico es 0.093 y la más pequeña es 0.044. Como se indicó antes, las últimas cuatro filas promedian sobre todos los dominios y según el tamaño del dominio, obteniendo que el estimador directo tiene una mayor precisión para los dominios grandes con un sesgo empírico promedio en valor absoluto de 0.001, aunque su sesgo es más pequeño para dominios medianos y más sesgado para dominios pequeños. La última columna corresponde a la aproximación de la varianza dada por (4.11), cuyo mayor valor en los dominios es 0.092 y el menor 0.046, y se observa que no tiene diferencias significativas con los valores de la raíz del error cuadrático medio empírico. Observe que el menor valor de la varianza aproximada se obtiene en dominios grandes, por lo que el estimador directo será apropiado cuando el tamaño del dominio sea grande.

5.2. Estimadores de razón

Resultados

Las tablas de esta sección muestran el comportamiento de los estimadores de razón propuestos en la población simulada.

Tabla 5.3: Resultados de la simulación para el estimador de razón (ratio), $\widehat{P}_{rA_d}^{(1)}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.094	0.040	0.064	0.007
U_2	500	15.728	0.115	0.033	0.069	0.011
U_3	500	15.366	0.114	0.034	0.068	0.013
U_4	500	15.449	0.122	0.026	0.066	0.013
U_5	500	15.250	0.132	0.026	0.070	0.021
U_6	500	15.157	0.162	0.020	0.076	0.024
U_7	750	22.503	0.148	0.012	0.063	0.020
U_8	750	22.146	0.177	0.003	0.067	0.028
U_9	750	22.511	0.203	-0.001	0.068	0.031
U_{10}	750	22.484	0.202	0.002	0.074	0.036
U_{11}	750	22.187	0.277	0.001	0.077	0.035
U_{12}	750	22.284	0.227	0.005	0.074	0.042
U_{13}	1000	29.771	0.296	-0.001	0.070	0.046
U_{14}	1000	29.846	0.286	-0.001	0.069	0.048
U_{15}	1000	30.055	0.329	-0.002	0.073	0.051
U_{16}	1000	29.854	0.357	0.000	0.073	0.051
U_{17}	1000	29.817	0.396	0.000	0.070	0.064
U_{18}	1000	29.853	0.433	0.002	0.070	0.064
U_{19}	1250	37.726	0.459	0.000	0.065	0.062
U_{20}	1250	37.180	0.491	-0.001	0.065	0.067
U_{21}	1250	37.701	0.544	0.000	0.062	0.074
U_{22}	1250	37.294	0.590	0.002	0.059	0.077
U_{23}	1250	37.707	0.628	-0.001	0.061	0.082
U_{24}	1250	37.483	0.636	0.001	0.059	0.079
U_{25}	1500	44.443	0.683	-0.003	0.052	0.080
U_{26}	1500	44.686	0.756	-0.001	0.047	0.073
U_{27}	1500	44.952	0.771	0.001	0.045	0.074
U_{28}	1500	45.303	0.824	0.001	0.040	0.073
U_{29}	1500	44.985	0.867	0.001	0.036	0.070
U_{30}	1500	44.825	0.903	0.000	0.031	0.061
Todos		30	0.407	0.007	0.063	0.061
Pequeño		15.401	0.123	0.030	0.069	0.014
Mediano		29.911	0.371	0.001	0.068	0.052
Grande		44.866	0.801	0.000	0.042	0.073

Comentarios

De la tabla 5.3 se desprende que el estimador $\widehat{P}^{(1)}_{rA_d}$ (4.13), que utiliza la información auxiliar a nivel de dominio:

$$\widehat{P}^{(1)}_{rA_d} = \frac{\widehat{P}_{A_d}}{\widehat{P}_{B_d}} P_{B_d}$$

es muy poco sesgado (el mayor sesgo empírico en valor absoluto es 0.04) y bastante preciso (la mayor raíz del error cuadrático medio empírico es 0.077). La aproximación de la varianza dada en (4.16) es bastante precisa, aunque en media refleja una ligera subestimación.

Las últimas filas de la tabla 5.3 promedian las anteriores medidas sobre todos los dominios y sobre los dominios según su tamaño. El estimador $\widehat{P}^{(1)}_{rA_d}$ da una precisión mayor para dominios grandes, aunque también es bueno para dominios medianos y pequeños, siendo ligeramente más sesgado para estos.

De la tabla 5.4 se desprende que el estimador $\widehat{P}^{(2)}_{rA_d}$ (4.18), que utiliza la información auxiliar a nivel de población :

$$\widehat{P}^{(2)}_{rA_d} = \frac{\widehat{P}_{A_d}}{\widehat{P}_{B_d}} P_B$$

es más sesgado (el mayor sesgo empírico en valor absoluto es 0.262) y menos preciso (la mayor raíz del error cuadrático medio empírico es 0.268 y la menor 0.065). La aproximación de la varianza dada en (4.23) es poco precisa y la subestima a la vista de la última columna. De las últimas filas de la tabla 5.4 se deduce que el estimador $\widehat{P}^{(2)}_{rA_d}$ da una precisión mayor y un sesgo menor para dominios de tamaño mediano.

De la tabla 5.5 se desprende que el estimador sintético $\widehat{P}^{(3)}_{rA_d}$ (4.26):

$$\widehat{P}^{(3)}_{rA_d} = \widehat{R} P_{B_d} = \frac{\widehat{P}_A}{\widehat{P}_B} P_{B_d}$$

tiene mayor precisión y menor sesgo para dominios medianos, se sitúa en precisión y sesgo entre los dos estimadores anteriores. La aproximación de la varianza dada por (4.32), subestima la varianza del estimador.

Tabla 5.4: Resultados de la simulación para el estimador razón (ratio1), $\widehat{P}^{(2)}_{rA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.276	0.222	0.268	0.058
U_2	500	15.728	0.238	0.156	0.200	0.068
U_3	500	15.366	0.260	0.180	0.225	0.065
U_4	500	15.449	0.248	0.152	0.196	0.069
U_5	500	15.250	0.256	0.150	0.196	0.071
U_6	500	15.157	0.284	0.142	0.191	0.076
U_7	750	22.503	0.268	0.132	0.173	0.060
U_8	750	22.146	0.283	0.109	0.153	0.065
U_9	750	22.511	0.310	0.106	0.148	0.067
U_{10}	750	22.484	0.277	0.077	0.127	0.070
U_{11}	750	22.187	0.371	0.095	0.140	0.071
U_{12}	750	22.284	0.322	0.099	0.144	0.069
U_{13}	1000	29.771	0.370	0.073	0.114	0.065
U_{14}	1000	29.846	0.355	0.068	0.110	0.065
U_{15}	1000	30.055	0.375	0.044	0.094	0.067
U_{16}	1000	29.854	0.403	0.046	0.094	0.067
U_{17}	1000	29.817	0.426	0.030	0.081	0.068
U_{18}	1000	29.853	0.447	0.016	0.074	0.068
U_{19}	1250	37.726	0.454	-0.004	0.065	0.061
U_{20}	1250	37.180	0.461	-0.032	0.069	0.063
U_{21}	1250	37.701	0.492	-0.052	0.077	0.062
U_{22}	1250	37.294	0.527	-0.060	0.080	0.057
U_{23}	1250	37.707	0.523	-0.106	0.117	0.060
U_{24}	1250	37.483	0.542	-0.092	0.105	0.056
U_{25}	1500	44.443	0.558	-0.129	0.135	0.050
U_{26}	1500	44.686	0.586	-0.171	0.175	0.046
U_{27}	1500	44.952	0.588	-0.182	0.185	0.046
U_{28}	1500	45.303	0.610	-0.214	0.216	0.041
U_{29}	1500	44.985	0.625	-0.241	0.242	0.036
U_{30}	1500	44.825	0.641	-0.262	0.263	0.030
Todos		30	0.412	0.012	0.149	0.061
Pequeño		15.401	0.260	0.167	0.213	0.068
Mediano		29.911	0.400	0.030	0.109	0.064
Grande		44.866	0.601	-0.200	0.203	0.041

Tabla 5.5: Resultados de la simulación para el estimador de razón (ratio2), $\widehat{P}^{(3)}_{rA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.166	0.112	0.112	0.099
U_2	500	15.728	0.238	0.156	0.156	0.071
U_3	500	15.366	0.214	0.134	0.134	0.061
U_4	500	15.449	0.241	0.145	0.145	0.060
U_5	500	15.250	0.254	0.148	0.148	0.041
U_6	500	15.157	0.280	0.138	0.138	0.038
U_7	750	22.503	0.270	0.134	0.134	0.037
U_8	750	22.146	0.306	0.133	0.133	0.029
U_9	750	22.511	0.321	0.117	0.117	0.026
U_{10}	750	22.484	0.359	0.159	0.159	0.024
U_{11}	750	22.187	0.365	0.089	0.090	0.024
U_{12}	750	22.284	0.346	0.123	0.124	0.021
U_{13}	1000	29.771	0.392	0.095	0.096	0.017
U_{14}	1000	29.846	0.394	0.107	0.108	0.017
U_{15}	1000	30.055	0.430	0.099	0.100	0.015
U_{16}	1000	29.854	0.434	0.077	0.077	0.016
U_{17}	1000	29.817	0.456	0.060	0.061	0.013
U_{18}	1000	29.853	0.474	0.043	0.045	0.013
U_{19}	1250	37.726	0.495	0.037	0.039	0.012
U_{20}	1250	37.180	0.522	0.030	0.032	0.011
U_{21}	1250	37.701	0.542	-0.002	0.013	0.010
U_{22}	1250	37.294	0.548	-0.039	0.041	0.009
U_{23}	1250	37.707	0.588	-0.041	0.043	0.009
U_{24}	1250	37.483	0.574	-0.060	0.062	0.008
U_{25}	1500	44.443	0.600	-0.087	0.088	0.008
U_{26}	1500	44.686	0.632	-0.125	0.126	0.007
U_{27}	1500	44.952	0.643	-0.127	0.128	0.007
U_{28}	1500	45.303	0.662	-0.161	0.162	0.007
U_{29}	1500	44.985	0.679	-0.187	0.188	0.006
U_{30}	1500	44.825	0.690	-0.213	0.214	0.006
Todos		30	0.437	0.036	0.107	0.012
Pequeño		15.401	0.232	0.139	0.139	0.057
Mediano		29.911	0.434	0.065	0.082	0.015
Grande		44.866	0.651	-0.150	0.151	0.007

5.3. Estimadores de regresión

Resultados

El comportamiento de los estimadores de regresión propuestos en la población simulada se muestran en las tablas de esta sección.

Tabla 5.6: Resultados de la simulación para el estimador de regresión, $\widehat{P}^{(1)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.064	0.010	0.093	0.058
U_2	500	15.728	0.093	0.011	0.109	0.068
U_3	500	15.366	0.088	0.008	0.108	0.065
U_4	500	15.449	0.104	0.008	0.104	0.069
U_5	500	15.250	0.113	0.007	0.109	0.071
U_6	500	15.157	0.150	0.008	0.107	0.076
U_7	750	22.503	0.142	0.006	0.092	0.060
U_8	750	22.146	0.172	-0.002	0.095	0.065
U_9	750	22.511	0.202	-0.002	0.086	0.067
U_{10}	750	22.484	0.198	-0.002	0.095	0.070
U_{11}	750	22.187	0.276	0.000	0.090	0.071
U_{12}	750	22.284	0.229	0.007	0.091	0.069
U_{13}	1000	29.771	0.292	-0.005	0.084	0.065
U_{14}	1000	29.846	0.288	0.001	0.082	0.065
U_{15}	1000	30.055	0.331	0.000	0.086	0.067
U_{16}	1000	29.854	0.359	0.002	0.082	0.067
U_{17}	1000	29.817	0.395	-0.001	0.081	0.068
U_{18}	1000	29.853	0.433	0.002	0.081	0.068
U_{19}	1250	37.726	0.460	0.001	0.070	0.061
U_{20}	1250	37.180	0.491	-0.002	0.072	0.063
U_{21}	1250	37.701	0.543	-0.001	0.067	0.062
U_{22}	1250	37.294	0.588	0.001	0.062	0.057
U_{23}	1250	37.707	0.629	0.000	0.063	0.060
U_{24}	1250	37.483	0.635	0.000	0.063	0.056
U_{25}	1500	44.443	0.683	-0.004	0.054	0.050
U_{26}	1500	44.686	0.756	-0.001	0.049	0.046
U_{27}	1500	44.952	0.772	0.002	0.047	0.046
U_{28}	1500	45.303	0.824	0.000	0.043	0.041
U_{29}	1500	44.985	0.871	0.005	0.038	0.036
U_{30}	1500	44.825	0.899	-0.004	0.033	0.030
Todos		30	0.403	0.002	0.078	0.060
Pequeño		15.401	0.102	0.009	0.105	0.068
Mediano		29.911	0.370	0.000	0.080	0.064
Grande		44.866	0.801	0.000	0.044	0.041

Comentarios

De la tabla 5.6 se observa que el estimador de regresión $\widehat{P}^{(1)}_{regA_d}$ (4.96), con información auxiliar de dominio,

$$\widehat{P}^{(1)}_{regA_d} = \widehat{P}_{A_d} + \widehat{b}_{opt}(P_{B_d} - \widehat{P}_{B_d}),$$

es poco sesgado y preciso. Tiene como mayor sesgo empírico promedio 0.011 y una precisión (RECME) que varía entre los valores 0.033 y 0.109. El estimador da una precisión mayor para dominios grandes, aunque la precisión para dominios medianos y pequeños puede considerarse buena, con sesgos cercanos a cero. La varianza aproximada obtenida a partir de (4.98) refleja una subestimación de la varianza del estimador.

El estimador de regresión (4.100), con información auxiliar a nivel de población, según se desprende de la tabla 5.7

$$\widehat{P}^{(2)}_{regA_d} = \widehat{P}_{A_d} + \widehat{b}_{opt}(P_B - \widehat{P}_{B_d}),$$

es sesgado y menos preciso que $\widehat{P}^{(1)}_{regA_d}$, toda vez que su sesgo empírico promedio toma como valor mayor 0.424 y valor menor 0.05, lo cual conduce a valores grandes de la raíz cuadrada del error cuadrático medio empírico que varían entre 0.07 y 0.449. La varianza aproximada calculada a partir de (4.98) subestima la varianza del estimador. Respecto a los tamaños de dominio, este estimador tiene una mayor precisión para dominios de tamaño mediano, como se comprueba al observar el valor de la raíz del error cuadrático medio empírico (0.153) y el sesgo empírico promedio(0.071).

Como puede observarse en la tabla 5.8, el estimador de regresión sintético (4.108),

$$\widehat{P}^{(3)}_{regA_d} = \widehat{P}_A + \widehat{b}_{opt}(P_{B_d} - \widehat{P}_B),$$

es un poco sesgado, debido a que los valores absolutos de su sesgo empírico varían entre 0.002 y 0.214, lo cual permite situarlo entre los dos estimadores de regresión anteriores, lo que se confirma al observar la RECME. Si consideramos el tamaño de dominio se observa mayor precisión en los dominios de tamaño mediano y aunque la precisión para los dominios de tamaño pequeño es inferior a la de los dominios medianos, ésta es superior a la precisión de los dominios grandes. La varianza aproximada a partir de (4.110) subestima la varianza del estimador y es constante para todos los dominios.

Tabla 5.7: Resultados de la simulación para el estimador de (regresión1), $\widehat{P}^{(2)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.478	0.424	0.449	0.058
U_2	500	15.728	0.415	0.333	0.363	0.068
U_3	500	15.366	0.440	0.360	0.389	0.065
U_4	500	15.449	0.422	0.326	0.356	0.069
U_5	500	15.250	0.415	0.309	0.339	0.071
U_6	500	15.157	0.419	0.277	0.307	0.076
U_7	750	22.503	0.422	0.286	0.312	0.060
U_8	750	22.146	0.407	0.233	0.261	0.065
U_9	750	22.511	0.418	0.214	0.239	0.067
U_{10}	750	22.484	0.366	0.166	0.199	0.070
U_{11}	750	22.187	0.435	0.159	0.190	0.071
U_{12}	750	22.284	0.413	0.190	0.220	0.069
U_{13}	1000	29.771	0.417	0.120	0.151	0.065
U_{14}	1000	29.846	0.410	0.123	0.152	0.065
U_{15}	1000	30.055	0.408	0.077	0.119	0.067
U_{16}	1000	29.854	0.431	0.074	0.112	0.067
U_{17}	1000	29.817	0.438	0.042	0.093	0.068
U_{18}	1000	29.853	0.453	0.022	0.084	0.068
U_{19}	1250	37.726	0.453	-0.005	0.070	0.061
U_{20}	1250	37.180	0.449	-0.044	0.084	0.063
U_{21}	1250	37.701	0.476	-0.068	0.097	0.062
U_{22}	1250	37.294	0.514	-0.073	0.099	0.057
U_{23}	1250	37.707	0.503	-0.125	0.145	0.060
U_{24}	1250	37.483	0.527	-0.107	0.128	0.056
U_{25}	1500	44.443	0.543	-0.144	0.159	0.050
U_{26}	1500	44.686	0.575	-0.182	0.197	0.046
U_{27}	1500	44.952	0.577	-0.193	0.208	0.046
U_{28}	1500	45.303	0.604	-0.220	0.234	0.041
U_{29}	1500	44.985	0.629	-0.237	0.245	0.036
U_{30}	1500	44.825	0.644	-0.259	0.278	0.030
Todos		30	0.470	0.069	0.209	0.060
Pequeño		15.401	0.432	0.338	0.367	0.068
Mediano		29.911	0.441	0.071	0.153	0.064
Grande		44.866	0.595	-0.206	0.220	0.041

Tabla 5.8: Resultados de la simulación para el estimador de (Regresión2), $\widehat{P}^{(3)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.166	0.112	0.112	0.012
U_2	500	15.728	0.238	0.156	0.156	0.012
U_3	500	15.366	0.214	0.134	0.134	0.012
U_4	500	15.449	0.241	0.145	0.145	0.012
U_5	500	15.250	0.254	0.148	0.148	0.012
U_6	500	15.157	0.280	0.138	0.138	0.012
U_7	750	22.503	0.270	0.134	0.134	0.012
U_8	750	22.146	0.306	0.133	0.133	0.012
U_9	750	22.511	0.321	0.117	0.117	0.012
U_{10}	750	22.484	0.359	0.159	0.159	0.012
U_{11}	750	22.187	0.365	0.089	0.090	0.012
U_{12}	750	22.284	0.346	0.123	0.124	0.012
U_{13}	1000	29.771	0.392	0.095	0.096	0.012
U_{14}	1000	29.846	0.394	0.107	0.108	0.012
U_{15}	1000	30.055	0.430	0.099	0.100	0.012
U_{16}	1000	29.854	0.434	0.077	0.077	0.012
U_{17}	1000	29.817	0.456	0.060	0.061	0.012
U_{18}	1000	29.853	0.474	0.043	0.045	0.012
U_{19}	1250	37.726	0.495	0.037	0.039	0.012
U_{20}	1250	37.180	0.522	0.030	0.032	0.012
U_{21}	1250	37.701	0.542	-0.002	0.013	0.012
U_{22}	1250	37.294	0.548	-0.039	0.041	0.012
U_{23}	1250	37.707	0.588	-0.041	0.043	0.012
U_{24}	1250	37.483	0.574	-0.060	0.062	0.012
U_{25}	1500	44.443	0.600	-0.087	0.088	0.012
U_{26}	1500	44.686	0.632	-0.125	0.126	0.012
U_{27}	1500	44.952	0.643	-0.127	0.128	0.012
U_{28}	1500	45.303	0.662	-0.161	0.162	0.012
U_{29}	1500	44.985	0.679	-0.187	0.188	0.012
U_{30}	1500	44.825	0.690	-0.213	0.214	0.012
Todos		30	0.437	0.036	0.107	0.012
Pequeño		15.401	0.232	0.139	0.139	0.012
Mediano		29.911	0.434	0.065	0.082	0.012
Grande		44.866	0.651	-0.150	0.151	0.012

5.4. Estimadores de diferencia

Resultados

Las tablas de esta sección muestran el comportamiento de los estimadores de diferencia propuestos en la población simulada.

Tabla 5.9: Resultados de la simulación para el estimador de diferencia, $\widehat{P}^{(1)}_{difA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.062	0.008	0.095	0.092
U_2	500	15.728	0.092	0.010	0.110	0.106
U_3	500	15.366	0.087	0.007	0.109	0.101
U_4	500	15.449	0.101	0.005	0.107	0.105
U_5	500	15.250	0.112	0.006	0.111	0.106
U_6	500	15.157	0.149	0.007	0.108	0.107
U_7	750	22.503	0.142	0.006	0.092	0.087
U_8	750	22.146	0.170	-0.004	0.094	0.090
U_9	750	22.511	0.204	0.000	0.085	0.088
U_{10}	750	22.484	0.198	-0.002	0.096	0.094
U_{11}	750	22.187	0.275	-0.001	0.089	0.086
U_{12}	750	22.284	0.230	0.007	0.092	0.089
U_{13}	1000	29.771	0.294	-0.003	0.084	0.077
U_{14}	1000	29.846	0.287	0.000	0.081	0.078
U_{15}	1000	30.055	0.330	-0.001	0.083	0.079
U_{16}	1000	29.854	0.359	0.002	0.079	0.077
U_{17}	1000	29.817	0.395	-0.001	0.077	0.076
U_{18}	1000	29.853	0.432	0.001	0.077	0.074
U_{19}	1250	37.726	0.460	0.002	0.069	0.066
U_{20}	1250	37.180	0.491	-0.001	0.070	0.067
U_{21}	1250	37.701	0.543	-0.001	0.066	0.065
U_{22}	1250	37.294	0.589	0.002	0.060	0.060
U_{23}	1250	37.707	0.629	0.000	0.060	0.062
U_{24}	1250	37.483	0.636	0.001	0.061	0.058
U_{25}	1500	44.443	0.683	-0.003	0.053	0.051
U_{26}	1500	44.686	0.756	0.000	0.047	0.046
U_{27}	1500	44.952	0.772	0.002	0.046	0.046
U_{28}	1500	45.303	0.824	0.001	0.041	0.041
U_{29}	1500	44.985	0.867	0.001	0.036	0.036
U_{30}	1500	44.825	0.903	0.000	0.031	0.030
Todos		30	0.402	0.002	0.077	0.075
Pequeño		15.401	0.100	0.007	0.107	0.103
Mediano		29.911	0.370	0.001	0.079	0.076
Grande		44.866	0.801	0.000	0.042	0.042

Comentarios

La tabla 5.9 evidencia que el estimador de diferencia $\widehat{P}^{(1)}_{difA_d}$ (4.101) con información auxiliar a nivel de dominio

$$\widehat{P}^{(1)}_{difA_d} = \widehat{P}_{A_d} + (P_{B_d} - \widehat{P}_{B_d}),$$

es poco sesgado, toda vez que el mayor valor del sesgo absoluto empírico es 0.01, con precisión moderada con valor mayor de la RECME 0.111 y valor menor 0.031. Este estimador es más preciso para dominios grandes, aunque también es algo preciso para dominios medianos y pequeños, cuyos sesgos son cercanos a cero. La varianza aproximada por la fórmula (4.105) es bastante precisa, aunque se aprecia una ligera subestimación.

De la tabla 5.10 se desprende que el estimador de diferencia $\widehat{P}^{(2)}_{difA_d}$ (4.102), con información auxiliar a nivel de población,

$$\widehat{P}^{(2)}_{difA_d} = \widehat{P}_{A_d} + (P_B - \widehat{P}_{B_d}),$$

es bastante sesgado, con un mayor valor del sesgo absoluto empírico de 0.453 y menor de 0.005, y bastante impreciso puesto que la RECME toma valores que varían entre 0.069 y 0.462. Aun siendo impreciso, se observa mayor precisión en dominios medianos y la varianza aproximada a partir de la fórmula (4.105), subestima la varianza del estimador.

Los resultados de la tabla 5.11 indican que el estimador de diferencia sintético (4.103)

$$\widehat{P}^{(3)}_{difA_d} = \widehat{P}_A + (P_{B_d} - \widehat{P}_B),$$

es sesgado, su sesgo absoluto empírico toma valores, en su mayoría menores que los del estimador $\widehat{P}^{(1)}_{difA_d}$, entre 0.004 y 0.139, con precisión entre 0.014 y 0.14. Tiene un mejor comportamiento para los dominios pequeños, aunque también tiene buen comportamiento en los dominios medianos y grandes, situándose en precisión ligeramente por abajo del estimador $\widehat{P}^{(1)}_{difA_d}$. La varianza aproximada mediante la fórmula (4.107), es adecuada, sobre todo en dominios medianos.

Tabla 5.10: Resultados de la simulación para el estimador de (diferencia1), $\widehat{P}^{(2)}_{difA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.507	0.453	0.462	0.092
U_2	500	15.728	0.438	0.356	0.373	0.106
U_3	500	15.366	0.465	0.385	0.400	0.101
U_4	500	15.449	0.443	0.347	0.363	0.105
U_5	500	15.250	0.436	0.330	0.348	0.106
U_6	500	15.157	0.437	0.295	0.314	0.107
U_7	750	22.503	0.444	0.308	0.321	0.087
U_8	750	22.146	0.422	0.249	0.266	0.090
U_9	750	22.511	0.436	0.232	0.247	0.088
U_{10}	750	22.484	0.378	0.178	0.203	0.094
U_{11}	750	22.187	0.446	0.170	0.192	0.086
U_{12}	750	22.284	0.428	0.205	0.225	0.089
U_{13}	1000	29.771	0.429	0.132	0.156	0.077
U_{14}	1000	29.846	0.419	0.132	0.155	0.078
U_{15}	1000	30.055	0.412	0.081	0.116	0.079
U_{16}	1000	29.854	0.436	0.079	0.112	0.077
U_{17}	1000	29.817	0.442	0.046	0.090	0.076
U_{18}	1000	29.853	0.454	0.023	0.080	0.074
U_{19}	1250	37.726	0.453	-0.005	0.069	0.066
U_{20}	1250	37.180	0.447	-0.046	0.084	0.067
U_{21}	1250	37.701	0.471	-0.073	0.098	0.065
U_{22}	1250	37.294	0.509	-0.078	0.098	0.060
U_{23}	1250	37.707	0.494	-0.135	0.147	0.062
U_{24}	1250	37.483	0.520	-0.114	0.129	0.058
U_{25}	1500	44.443	0.532	-0.154	0.163	0.051
U_{26}	1500	44.686	0.561	-0.195	0.201	0.046
U_{27}	1500	44.952	0.562	-0.208	0.213	0.046
U_{28}	1500	45.303	0.588	-0.236	0.239	0.041
U_{29}	1500	44.985	0.607	-0.259	0.261	0.036
U_{30}	1500	44.825	0.629	-0.275	0.276	0.030
Todos		30	0.475	0.074	0.213	0.075
Pequeño		15.401	0.454	0.361	0.377	0.103
Mediano		29.911	0.447	0.077	0.155	0.076
Grande		44.866	0.580	-0.221	0.226	0.042

Tabla 5.11: Resultados de la simulación para el estimador de (diferencia2), $\widehat{P}^{(3)}_{difA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.046	-0.008	0.015	0.057
U_2	500	15.728	0.144	0.062	0.063	0.057
U_3	500	15.366	0.112	0.032	0.034	0.057
U_4	500	15.449	0.148	0.052	0.053	0.057
U_5	500	15.250	0.166	0.060	0.061	0.057
U_6	500	15.157	0.202	0.060	0.061	0.057
U_7	750	22.503	0.188	0.052	0.054	0.057
U_8	750	22.146	0.238	0.064	0.066	0.057
U_9	750	22.511	0.258	0.054	0.055	0.057
U_{10}	750	22.484	0.310	0.110	0.110	0.057
U_{11}	750	22.187	0.319	0.043	0.045	0.057
U_{12}	750	22.284	0.292	0.070	0.071	0.057
U_{13}	1000	29.771	0.356	0.059	0.060	0.057
U_{14}	1000	29.846	0.359	0.072	0.073	0.057
U_{15}	1000	30.055	0.408	0.077	0.078	0.057
U_{16}	1000	29.854	0.413	0.056	0.057	0.057
U_{17}	1000	29.817	0.444	0.048	0.049	0.057
U_{18}	1000	29.853	0.469	0.038	0.040	0.057
U_{19}	1250	37.726	0.497	0.038	0.041	0.057
U_{20}	1250	37.180	0.534	0.042	0.044	0.057
U_{21}	1250	37.701	0.562	0.018	0.022	0.057
U_{22}	1250	37.294	0.570	-0.018	0.022	0.057
U_{23}	1250	37.707	0.625	-0.004	0.014	0.057
U_{24}	1250	37.483	0.606	-0.029	0.032	0.057
U_{25}	1500	44.443	0.641	-0.046	0.047	0.057
U_{26}	1500	44.686	0.685	-0.072	0.073	0.057
U_{27}	1500	44.952	0.700	-0.070	0.072	0.057
U_{28}	1500	45.303	0.726	-0.097	0.098	0.057
U_{29}	1500	44.985	0.750	-0.116	0.117	0.057
U_{30}	1500	44.825	0.764	-0.139	0.140	0.057
Todos		30	0.418	0.017	0.059	0.057
Pequeño		15.401	0.136	0.043	0.048	0.057
Mediano		29.911	0.414	0.044	0.052	0.057
Grande		44.866	0.711	-0.090	0.091	0.057

5.5. Estimadores combinados de razón

Resultados

Las tablas de esta sección muestran el comportamiento de los estimadores combinados de razón propuestos en la población simulada.

Tabla 5.12: Resultados de la simulación para el estimador combinado (cra1), $\widehat{P}_{rA_d}^{(c_1)}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.098	0.044	0.066	0.007
U_2	500	15.728	0.118	0.036	0.070	0.009
U_3	500	15.366	0.117	0.037	0.069	0.011
U_4	500	15.449	0.125	0.029	0.068	0.011
U_5	500	15.250	0.135	0.029	0.072	0.017
U_6	500	15.157	0.165	0.023	0.079	0.019
U_7	750	22.503	0.149	0.013	0.064	0.016
U_8	750	22.146	0.179	0.005	0.069	0.021
U_9	750	22.511	0.204	0.000	0.071	0.023
U_{10}	750	22.484	0.204	0.004	0.076	0.025
U_{11}	750	22.187	0.279	0.003	0.080	0.025
U_{12}	750	22.284	0.227	0.005	0.076	0.028
U_{13}	1000	29.771	0.298	0.001	0.071	0.030
U_{14}	1000	29.846	0.286	-0.001	0.071	0.031
U_{15}	1000	30.055	0.330	-0.001	0.076	0.032
U_{16}	1000	29.854	0.357	0.000	0.077	0.032
U_{17}	1000	29.817	0.398	0.002	0.073	0.037
U_{18}	1000	29.853	0.434	0.003	0.073	0.037
U_{19}	1250	37.726	0.459	0.000	0.068	0.035
U_{20}	1250	37.180	0.492	0.000	0.067	0.037
U_{21}	1250	37.701	0.546	0.002	0.065	0.038
U_{22}	1250	37.294	0.592	0.005	0.063	0.039
U_{23}	1250	37.707	0.629	0.000	0.065	0.040
U_{24}	1250	37.483	0.637	0.003	0.062	0.039
U_{25}	1500	44.443	0.685	-0.002	0.054	0.037
U_{26}	1500	44.686	0.757	0.000	0.050	0.034
U_{27}	1500	44.952	0.773	0.003	0.048	0.034
U_{28}	1500	45.303	0.825	0.002	0.043	0.032
U_{29}	1500	44.985	0.867	0.001	0.038	0.030
U_{30}	1500	44.825	0.903	-0.001	0.032	0.026
Todos		30	0.409	0.008	0.065	0.034
Pequeño		15.401	0.126	0.033	0.071	0.012
Mediano		29.911	0.372	0.002	0.070	0.032
Grande		44.866	0.802	0.001	0.044	0.033

Comentarios

De la tabla 5.12 se observa que el estimador combinado de razón (4.50), a partir de los estimadores directo y de razón con información auxiliar a nivel de dominio,

$$\widehat{P}_{rA_d}^{(c_1)} = \widehat{\alpha}_{op}\widehat{P}_{A_d} + (1-\widehat{\alpha})_{opt}\widehat{P}_{rA_d}^{(1)},$$

es muy poco sesgado, toda vez que el valor mayor del sesgo empírico absoluto es 0.044 y el menor 0.001 y bastante preciso, ya que el mayor valor de la RECME es 0.08 y el menor es 0.032. La precisión de este estimador es mejor para dominios grandes aunque también tiene una precisión buena para los dominios mediano y pequeño, con sesgo empírico también pequeño. La varianza aproximada a partir de la ecuación (4.52), subestima la varianza del estimador.

El estimador combinado de razón (4.54), a partir de los estimadores directo y de razón con información auxiliar a nivel de población,

$$\widehat{P}_{rA_d}^{(c_2)} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(2)},$$

es sesgado y tiene como mayor sesgo empírico 0.191 y menor 0.003, como se observa en la tabla 5.13. La precisión de este estimador es mejor para los dominios medianos debido a que el valor de la RECME es 0.094, menor que los asociados a los dominios grandes y pequeños, y su sesgo es de apenas 0.02. La varianza aproximada dada por (4.56) hace evidente una severa subestimación de la varianza del estimador para dominios pequeños y medianos.

El estimador combinado (4.58), a partir de los estimadores directo y de razón sintético,

$$\widehat{P}_{rA_d}^{(c_3)} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(3)},$$

es sesgado y tiene como mayor valor absoluto del sesgo empírico 0.195 y valor menor 0.002, según se desprende de la tabla 5.14. Tiene menor sesgo para los dominios medianos y aproximadamente igual para los dominios grandes y pequeños. Observemos que la precisión del estimador medida a partir de la RECME es mejor para los dominios medianos, con una varianza aproximada a partir de la expresión (4.60), que subestima la varianza del estimador para cualquier tamaño de dominio.

De la tabla 5.15 se desprende que el estimador combinado (4.62), a partir de los estimadores de razón con información auxiliar de dominio y de razón sintético,

$$\widehat{P}_{rA_d}^{(c_4)} = \widehat{\alpha}_{opt}\widehat{P}_{rA_d}^{(1)} + (1-\widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(3)},$$

es un poco sesgado, toda vez que el mayor valor del sesgo empírico absoluto es 0.153 y el menor es 0.001. Es poco sesgado para los dominios medianos y se observa una precisión similar en los dominios grandes y pequeños, aunque

el sesgo para dominios pequeños es mayor. La varianza aproximada calculada a partir de (4.64) evidencia una severa subestimación de la varianza del estimador para cualquier tamaño de dominio.

El estimador combinado (4.66), de los estimadores de razón con información auxiliar a nivel de población y de razón sintético,

$$\widehat{P}_{rA_d}^{(c_5)} = \widehat{\alpha}_{opt}\widehat{P}_{rA_d}^{(2)} + (1 - \widehat{\alpha}_{opt})\widehat{P}_{rA_d}^{(3)},$$

es sesgado y tiene un mayor sesgo empírico absoluto que los combinados de razón anteriores y menor precisión. Como se observa en la tabla 5.16, el mayor sesgo absoluto empírico es 0.222 y el menor es 0.003. Su precisión varía entre 0.014 y 0.223, por lo que es algo impreciso. Tiene una mayor precisión para dominios medianos y la varianza aproximada por la fórmula (4.68) evidencia subestimación de la varianza del estimador.

Tabla 5.13: Resultados de la simulación para el estimador combinado (cra2),$\widehat{P}_{rA_d}^{(c_2)}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.224	0.170	0.227	0.045
U_2	500	15.728	0.195	0.113	0.164	0.052
U_3	500	15.366	0.213	0.133	0.186	0.050
U_4	500	15.449	0.204	0.108	0.158	0.053
U_5	500	15.250	0.214	0.108	0.161	0.054
U_6	500	15.157	0.243	0.101	0.159	0.058
U_7	750	22.503	0.224	0.088	0.135	0.046
U_8	750	22.146	0.244	0.071	0.123	0.050
U_9	750	22.511	0.270	0.066	0.120	0.052
U_{10}	750	22.484	0.248	0.048	0.107	0.053
U_{11}	750	22.187	0.340	0.064	0.121	0.056
U_{12}	750	22.284	0.286	0.064	0.119	0.054
U_{13}	1000	29.771	0.344	0.047	0.097	0.051
U_{14}	1000	29.846	0.329	0.042	0.094	0.050
U_{15}	1000	30.055	0.358	0.027	0.088	0.052
U_{16}	1000	29.854	0.386	0.029	0.089	0.053
U_{17}	1000	29.817	0.417	0.021	0.080	0.053
U_{18}	1000	29.853	0.443	0.012	0.076	0.053
U_{19}	1250	37.726	0.456	-0.003	0.067	0.048
U_{20}	1250	37.180	0.473	-0.020	0.067	0.049
U_{21}	1250	37.701	0.512	-0.032	0.067	0.049
U_{22}	1250	37.294	0.550	-0.037	0.067	0.046
U_{23}	1250	37.707	0.561	-0.067	0.087	0.047
U_{24}	1250	37.483	0.575	-0.060	0.079	0.045
U_{25}	1500	44.443	0.601	-0.085	0.096	0.040
U_{26}	1500	44.686	0.643	-0.114	0.120	0.037
U_{27}	1500	44.952	0.652	-0.118	0.124	0.036
U_{28}	1500	45.303	0.681	-0.142	0.145	0.032
U_{29}	1500	44.985	0.701	-0.165	0.167	0.029
U_{30}	1500	44.825	0.712	-0.191	0.194	0.024
Todos		30	0.410	0.009	0.120	0.047
Pequeño		15.401	0.215	0.122	0.176	0.052
Mediano		29.911	0.390	0.020	0.094	0.050
Grande		44.866	0.665	-0.136	0.141	0.033

Tabla 5.14: Resultados de la simulación para el estimador combinado (cra3), $\widehat{P}_{rA_d}^{(c_3)}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.163	0.109	0.109	0.043
U_2	500	15.728	0.233	0.151	0.151	0.037
U_3	500	15.366	0.211	0.131	0.131	0.036
U_4	500	15.449	0.236	0.140	0.141	0.035
U_5	500	15.250	0.250	0.144	0.144	0.028
U_6	500	15.157	0.277	0.135	0.135	0.027
U_7	750	22.503	0.265	0.129	0.129	0.025
U_8	750	22.146	0.302	0.128	0.129	0.021
U_9	750	22.511	0.317	0.113	0.113	0.020
U_{10}	750	22.484	0.354	0.154	0.154	0.018
U_{11}	750	22.187	0.363	0.087	0.088	0.019
U_{12}	750	22.284	0.342	0.120	0.120	0.017
U_{13}	1000	29.771	0.389	0.092	0.093	0.014
U_{14}	1000	29.846	0.391	0.104	0.105	0.014
U_{15}	1000	30.055	0.427	0.096	0.097	0.013
U_{16}	1000	29.854	0.432	0.075	0.075	0.013
U_{17}	1000	29.817	0.455	0.059	0.060	0.011
U_{18}	1000	29.853	0.473	0.042	0.044	0.011
U_{19}	1250	37.726	0.494	0.036	0.038	0.010
U_{20}	1250	37.180	0.522	0.029	0.032	0.010
U_{21}	1250	37.701	0.542	-0.002	0.013	0.009
U_{22}	1250	37.294	0.549	-0.038	0.040	0.008
U_{23}	1250	37.707	0.589	-0.039	0.042	0.008
U_{24}	1250	37.483	0.576	-0.058	0.060	0.007
U_{25}	1500	44.443	0.603	-0.084	0.085	0.007
U_{26}	1500	44.686	0.637	-0.119	0.120	0.007
U_{27}	1500	44.952	0.649	-0.121	0.123	0.006
U_{28}	1500	45.303	0.671	-0.152	0.153	0.006
U_{29}	1500	44.985	0.692	-0.174	0.175	0.006
U_{30}	1500	44.825	0.709	-0.195	0.196	0.005
Todos		30	0.437	0.036	0.103	0.010
Pequeño		15.401	0.228	0.135	0.135	0.034
Mediano		29.911	0.432	0.063	0.080	0.013
Grande		44.866	0.660	-0.141	0.142	0.006

Tabla 5.15: Resultados de la simulación para el estimador combinado (cra4), $\widehat{P}_{rA_d}^{(c_4)}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.167	0.113	0.114	0.007
U_2	500	15.728	0.235	0.153	0.154	0.009
U_3	500	15.366	0.213	0.133	0.134	0.011
U_4	500	15.449	0.238	0.142	0.142	0.011
U_5	500	15.250	0.251	0.145	0.146	0.014
U_6	500	15.157	0.278	0.136	0.138	0.014
U_7	750	22.503	0.264	0.128	0.129	0.013
U_8	750	22.146	0.301	0.128	0.128	0.014
U_9	750	22.511	0.316	0.112	0.113	0.014
U_{10}	750	22.484	0.353	0.153	0.153	0.014
U_{11}	750	22.187	0.364	0.088	0.089	0.014
U_{12}	750	22.284	0.342	0.119	0.120	0.014
U_{13}	1000	29.771	0.389	0.092	0.092	0.013
U_{14}	1000	29.846	0.390	0.103	0.104	0.012
U_{15}	1000	30.055	0.427	0.096	0.096	0.012
U_{16}	1000	29.854	0.431	0.074	0.075	0.012
U_{17}	1000	29.817	0.455	0.059	0.061	0.011
U_{18}	1000	29.853	0.475	0.044	0.048	0.011
U_{19}	1250	37.726	0.494	0.036	0.038	0.010
U_{20}	1250	37.180	0.522	0.029	0.032	0.010
U_{21}	1250	37.701	0.543	-0.001	0.015	0.009
U_{22}	1250	37.294	0.552	-0.036	0.040	0.008
U_{23}	1250	37.707	0.591	-0.037	0.042	0.008
U_{24}	1250	37.483	0.580	-0.054	0.059	0.007
U_{25}	1500	44.443	0.607	-0.080	0.082	0.007
U_{26}	1500	44.686	0.645	-0.112	0.115	0.007
U_{27}	1500	44.952	0.656	-0.114	0.117	0.007
U_{28}	1500	45.303	0.685	-0.138	0.144	0.006
U_{29}	1500	44.985	0.716	-0.150	0.160	0.006
U_{30}	1500	44.825	0.756	-0.148	0.169	0.006
Todos		30	0.441	0.040	0.102	0.010
Pequeño		15.401	0.230	0.137	0.138	0.011
Mediano		29.911	0.433	0.063	0.080	0.012
Grande		44.866	0.677	-0.124	0.131	0.006

Tabla 5.16: Resultados de la simulación para el estimador combinado (cra5), $\widehat{P}_{rA_d}^{(c_5)}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.207	0.153	0.202	0.037
U_2	500	15.728	0.249	0.167	0.182	0.033
U_3	500	15.366	0.234	0.154	0.179	0.032
U_4	500	15.449	0.249	0.153	0.166	0.031
U_5	500	15.250	0.264	0.158	0.171	0.026
U_6	500	15.157	0.289	0.147	0.159	0.025
U_7	750	22.503	0.272	0.136	0.140	0.023
U_8	750	22.146	0.307	0.133	0.136	0.020
U_9	750	22.511	0.321	0.117	0.118	0.019
U_{10}	750	22.484	0.356	0.156	0.157	0.018
U_{11}	750	22.187	0.369	0.093	0.098	0.018
U_{12}	750	22.284	0.345	0.123	0.123	0.016
U_{13}	1000	29.771	0.392	0.095	0.095	0.014
U_{14}	1000	29.846	0.393	0.106	0.107	0.013
U_{15}	1000	30.055	0.428	0.097	0.098	0.012
U_{16}	1000	29.854	0.433	0.076	0.077	0.013
U_{17}	1000	29.817	0.456	0.060	0.062	0.011
U_{18}	1000	29.853	0.475	0.044	0.049	0.011
U_{19}	1250	37.726	0.494	0.035	0.037	0.010
U_{20}	1250	37.180	0.521	0.028	0.031	0.009
U_{21}	1250	37.701	0.541	-0.003	0.014	0.008
U_{22}	1250	37.294	0.548	-0.039	0.042	0.008
U_{23}	1250	37.707	0.586	-0.043	0.045	0.008
U_{24}	1250	37.483	0.574	-0.060	0.063	0.007
U_{25}	1500	44.443	0.598	-0.089	0.090	0.007
U_{26}	1500	44.686	0.630	-0.127	0.128	0.006
U_{27}	1500	44.952	0.640	-0.130	0.131	0.006
U_{28}	1500	45.303	0.658	-0.165	0.166	0.006
U_{29}	1500	44.985	0.673	-0.193	0.193	0.005
U_{30}	1500	44.825	0.681	-0.222	0.223	0.005
Todos		30	0.439	0.039	0.116	0.010
Pequeño		15.401	0.249	0.155	0.176	0.031
Mediano		29.911	0.434	0.064	0.083	0.012
Grande		44.866	0.647	-0.154	0.155	0.006

5.6. Estimadores combinados de regresión

Resultados

Las tablas de esta sección muestran el comportamiento de los estimadores de regresión combinados propuestos en la población simulada.

Tabla 5.17: Resultados de la simulación para el estimador combinado (crega1), $\widehat{P}^{(c_1)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.082	0.028	0.070	0.045
U_2	500	15.728	0.107	0.025	0.080	0.052
U_3	500	15.366	0.104	0.024	0.078	0.050
U_4	500	15.449	0.116	0.020	0.076	0.053
U_5	500	15.250	0.126	0.020	0.080	0.054
U_6	500	15.157	0.159	0.017	0.083	0.058
U_7	750	22.503	0.146	0.010	0.069	0.046
U_8	750	22.146	0.177	0.003	0.074	0.050
U_9	750	22.511	0.204	0.000	0.070	0.052
U_{10}	750	22.484	0.202	0.002	0.077	0.053
U_{11}	750	22.187	0.279	0.003	0.079	0.056
U_{12}	750	22.284	0.229	0.006	0.074	0.054
U_{13}	1000	29.771	0.296	-0.001	0.072	0.051
U_{14}	1000	29.846	0.288	0.001	0.070	0.050
U_{15}	1000	30.055	0.332	0.001	0.075	0.052
U_{16}	1000	29.854	0.359	0.002	0.074	0.053
U_{17}	1000	29.817	0.398	0.002	0.073	0.053
U_{18}	1000	29.853	0.435	0.004	0.074	0.053
U_{19}	1250	37.726	0.460	0.001	0.066	0.048
U_{20}	1250	37.180	0.492	-0.001	0.067	0.049
U_{21}	1250	37.701	0.546	0.002	0.065	0.049
U_{22}	1250	37.294	0.591	0.004	0.062	0.046
U_{23}	1250	37.707	0.629	0.001	0.064	0.047
U_{24}	1250	37.483	0.637	0.002	0.061	0.045
U_{25}	1500	44.443	0.685	-0.002	0.054	0.040
U_{26}	1500	44.686	0.757	0.000	0.050	0.037
U_{27}	1500	44.952	0.773	0.003	0.048	0.036
U_{28}	1500	45.303	0.825	0.002	0.044	0.032
U_{29}	1500	44.985	0.870	0.004	0.039	0.029
U_{30}	1500	44.825	0.900	-0.003	0.033	0.024
Todos		30	0.407	0.006	0.067	0.047
Pequeño		15.401	0.115	0.022	0.078	0.052
Mediano		29.911	0.372	0.002	0.070	0.050
Grande		44.866	0.802	0.001	0.045	0.033

Comentarios

De acuerdo con la tabla 5.17, el estimador(4.117), combinado de los estimadores directo y de regresión con información auxiliar a nivel de dominio,

$$\widehat{P}^{(c_1)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(1)}_{regA_d},$$

es poco sesgado, toda vez que el mayor valor absoluto del sesgo empírico es 0.028 y el menor es 0.001; el mayor valor de la RECME es 0.083 y el menor es 0.033, lo que nos indica buena precisión. A la vista de la última columna, la varianza estimada dada por la ecuación (4.118) produce una significativa subestimación de la varianza del estimador para cualquier tamaño de dominio. La precisión de este estimador es mejor para dominios grandes, siendo en estos menos sesgado, aunque también es buena para dominios medianos y pequeños, en los cuales el sesgo es pequeño pero ligeramente mayor que el que corresponde a los dominios grandes y la precisión puede considerarse buena.

Según tabla 5.18 el estimador (4.119), combinado de los estimadores directo y de regresión con información auxiliar a nivel de población,

$$\widehat{P}^{(c_2)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(2)}_{regA_d},$$

es sesgado, como mayor sesgo absoluto empírico de 0.285 y menor valor de 0.003. Es impreciso, dado que su RECME varía entre 0.066 y 0.314. La varianza aproximada dada por la ecuación (4.120) subestima severamente la varianza del estimador, como se puede comprobar en la última columna. El estimador tiene un mejor comportamiento para los dominios medianos, como se constata al observar su RECME y sesgo empírico.

La tabla 5.19 indica que el estimador (4.121), combinado de los estimadores directo y de regresión sintético,

$$\widehat{P}^{(c_3)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(3)}_{regA_d},$$

es sesgado, pero menos que $\widehat{P}^{(c_2)}_{regA_d}$, su mayor sesgo absoluto empírico es 0.196 y el menor es 0.002. Es impreciso, porque su RECME toma valores entre 0.013 y 0.197. El estimador tiene mejor precisión para dominios medianos, seguido de dominios pequeños y grandes, en ese orden. La varianza aproximada calculada con (4.122) es constante para los distintos tamaños de dominio y subestima la varianza del estimador.

De la tabla 5.20 se desprende que el estimador (4.123), combinado de los estimadores de regresión con información auxiliar de dominio y de regresión sintético,

$$\widehat{P}^{(c_4)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}^{(1)}_{regA_d} + (1-\widehat{\alpha}_{opt})\widehat{P}^{(3)}_{regA_d},$$

es sesgado, toda vez que el mayor valor del sesgo absoluto empírico es 0.154 y el menor valor es 0.001, reflejando un comportamiento similar a $\widehat{P}^{(c_3)}_{regA_d}$. Es impreciso, con valores de la RECME que varían entre 0.015 y 0.171. El estimador tiene un mejor comportamiento para dominios medianos y su varianza estimada por la ecuación (4.124) es constante y subestima la varianza del estimador.

De la tabla 5.21 se observa que el estimador (4.125), combinado de los estimadores de regresión con información auxiliar de nivel poblacional y de regresión sintético,

$$\widehat{P}^{(c_5)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}^{(2)}_{A_d} + (1 - \widehat{\alpha}_{opt})\widehat{P}^{(3)}_{regA_d},$$

es sesgado, tiene como mayor sesgo absoluto empírico 0.21 y menor 0.004; es impreciso, debido a que su RECME toma valores entre 0.015 y 0.212. De los estimadores combinados de regresión, es el de peor comportamiento. Su comportamiento es mejor para dominios de tamaño mediano y la varianza aproximada por la fórmula (4.126) es constante y subestima la varianza del estimador.

Tabla 5.18: Resultados de la simulación para el estimador combinado (crega2), $\widehat{P}^{(c_2)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.339	0.285	0.314	0.045
U_2	500	15.728	0.297	0.215	0.245	0.052
U_3	500	15.366	0.316	0.236	0.266	0.050
U_4	500	15.449	0.304	0.208	0.238	0.053
U_5	500	15.250	0.305	0.199	0.230	0.054
U_6	500	15.157	0.322	0.180	0.213	0.058
U_7	750	22.503	0.313	0.177	0.203	0.046
U_8	750	22.146	0.316	0.143	0.173	0.050
U_9	750	22.511	0.335	0.131	0.159	0.052
U_{10}	750	22.484	0.300	0.100	0.135	0.053
U_{11}	750	22.187	0.379	0.103	0.142	0.056
U_{12}	750	22.284	0.340	0.118	0.150	0.054
U_{13}	1000	29.771	0.373	0.076	0.111	0.051
U_{14}	1000	29.846	0.362	0.075	0.109	0.050
U_{15}	1000	30.055	0.378	0.047	0.093	0.052
U_{16}	1000	29.854	0.404	0.047	0.092	0.053
U_{17}	1000	29.817	0.425	0.029	0.081	0.053
U_{18}	1000	29.853	0.448	0.017	0.077	0.053
U_{19}	1250	37.726	0.456	-0.003	0.066	0.048
U_{20}	1250	37.180	0.466	-0.026	0.070	0.049
U_{21}	1250	37.701	0.503	-0.041	0.075	0.049
U_{22}	1250	37.294	0.542	-0.045	0.075	0.046
U_{23}	1250	37.707	0.550	-0.079	0.102	0.047
U_{24}	1250	37.483	0.565	-0.069	0.091	0.045
U_{25}	1500	44.443	0.592	-0.095	0.110	0.040
U_{26}	1500	44.686	0.636	-0.121	0.134	0.037
U_{27}	1500	44.952	0.645	-0.125	0.137	0.036
U_{28}	1500	45.303	0.678	-0.145	0.156	0.032
U_{29}	1500	44.985	0.699	-0.167	0.179	0.029
U_{30}	1500	44.825	0.721	-0.183	0.190	0.024
Todos		30	0.444	0.043	0.147	0.047
Pequeño		15.401	0.314	0.221	0.251	0.052
Mediano		29.911	0.414	0.044	0.111	0.050
Grande		44.866	0.662	-0.139	0.151	0.033

Tabla 5.19: Resultados de la simulación para el estimador combinado (crega3), $\widehat{P}^{(c_3)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.163	0.109	0.110	0.012
U_2	500	15.728	0.234	0.152	0.152	0.012
U_3	500	15.366	0.211	0.131	0.131	0.012
U_4	500	15.449	0.237	0.141	0.141	0.012
U_5	500	15.250	0.250	0.144	0.144	0.012
U_6	500	15.157	0.277	0.135	0.135	0.012
U_7	750	22.503	0.265	0.129	0.129	0.012
U_8	750	22.146	0.302	0.129	0.129	0.012
U_9	750	22.511	0.317	0.113	0.114	0.012
U_{10}	750	22.484	0.354	0.154	0.154	0.012
U_{11}	750	22.187	0.363	0.087	0.088	0.012
U_{12}	750	22.284	0.343	0.120	0.120	0.012
U_{13}	1000	29.771	0.390	0.093	0.093	0.012
U_{14}	1000	29.846	0.391	0.104	0.105	0.012
U_{15}	1000	30.055	0.428	0.097	0.097	0.012
U_{16}	1000	29.854	0.432	0.075	0.076	0.012
U_{17}	1000	29.817	0.455	0.059	0.060	0.012
U_{18}	1000	29.853	0.474	0.043	0.044	0.012
U_{19}	1250	37.726	0.494	0.036	0.038	0.012
U_{20}	1250	37.180	0.522	0.029	0.032	0.012
U_{21}	1250	37.701	0.542	-0.002	0.013	0.012
U_{22}	1250	37.294	0.549	-0.038	0.040	0.012
U_{23}	1250	37.707	0.589	-0.039	0.042	0.012
U_{24}	1250	37.483	0.576	-0.059	0.060	0.012
U_{25}	1500	44.443	0.603	-0.084	0.085	0.012
U_{26}	1500	44.686	0.637	-0.120	0.121	0.012
U_{27}	1500	44.952	0.648	-0.122	0.123	0.012
U_{28}	1500	45.303	0.671	-0.153	0.154	0.012
U_{29}	1500	44.985	0.691	-0.175	0.176	0.012
U_{30}	1500	44.825	0.707	-0.196	0.197	0.011
Todos		30	0.437	0.036	0.103	0.012
Pequeño		15.401	0.229	0.135	0.135	0.012
Mediano		29.911	0.433	0.063	0.080	0.012
Grande		44.866	0.659	-0.142	0.143	0.012

Tabla 5.20: Resultados de la simulación para el estimador combinado (crega4), $\widehat{P}^{(c_4)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.165	0.111	0.113	0.011
U_2	500	15.728	0.235	0.153	0.154	0.012
U_3	500	15.366	0.213	0.133	0.134	0.011
U_4	500	15.449	0.238	0.142	0.143	0.012
U_5	500	15.250	0.251	0.145	0.146	0.012
U_6	500	15.157	0.278	0.136	0.138	0.012
U_7	750	22.503	0.265	0.129	0.130	0.012
U_8	750	22.146	0.302	0.129	0.129	0.012
U_9	750	22.511	0.317	0.113	0.113	0.012
U_{10}	750	22.484	0.354	0.154	0.154	0.012
U_{11}	750	22.187	0.364	0.088	0.091	0.012
U_{12}	750	22.284	0.343	0.120	0.120	0.012
U_{13}	1000	29.771	0.389	0.092	0.093	0.012
U_{14}	1000	29.846	0.391	0.104	0.105	0.012
U_{15}	1000	30.055	0.427	0.096	0.097	0.012
U_{16}	1000	29.854	0.432	0.075	0.076	0.012
U_{17}	1000	29.817	0.455	0.059	0.061	0.012
U_{18}	1000	29.853	0.475	0.044	0.048	0.012
U_{19}	1250	37.726	0.494	0.036	0.038	0.012
U_{20}	1250	37.180	0.522	0.029	0.033	0.012
U_{21}	1250	37.701	0.543	-0.001	0.015	0.012
U_{22}	1250	37.294	0.551	-0.036	0.040	0.012
U_{23}	1250	37.707	0.591	-0.037	0.042	0.012
U_{24}	1250	37.483	0.580	-0.055	0.059	0.012
U_{25}	1500	44.443	0.606	-0.080	0.082	0.012
U_{26}	1500	44.686	0.644	-0.113	0.115	0.011
U_{27}	1500	44.952	0.655	-0.115	0.118	0.011
U_{28}	1500	45.303	0.684	-0.139	0.145	0.011
U_{29}	1500	44.985	0.715	-0.151	0.162	0.011
U_{30}	1500	44.825	0.753	-0.150	0.171	0.010
Todos		30	0.441	0.040	0.102	0.012
Pequeño		15.401	0.230	0.137	0.138	0.012
Mediano		29.911	0.433	0.063	0.080	0.012
Grande		44.866	0.676	-0.125	0.132	0.011

Tabla 5.21: Resultados de la simulación para el estimador combinado (crega5), $\widehat{P}^{(c_5)}_{regA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.212	0.158	0.199	0.011
U_2	500	15.728	0.255	0.173	0.188	0.012
U_3	500	15.366	0.239	0.159	0.180	0.011
U_4	500	15.449	0.256	0.160	0.173	0.012
U_5	500	15.250	0.269	0.163	0.175	0.012
U_6	500	15.157	0.293	0.151	0.162	0.012
U_7	750	22.503	0.279	0.143	0.147	0.012
U_8	750	22.146	0.312	0.139	0.142	0.012
U_9	750	22.511	0.325	0.121	0.122	0.012
U_{10}	750	22.484	0.359	0.159	0.160	0.012
U_{11}	750	22.187	0.371	0.095	0.102	0.012
U_{12}	750	22.284	0.349	0.126	0.127	0.012
U_{13}	1000	29.771	0.393	0.096	0.097	0.012
U_{14}	1000	29.846	0.395	0.108	0.109	0.012
U_{15}	1000	30.055	0.430	0.099	0.099	0.012
U_{16}	1000	29.854	0.434	0.077	0.078	0.012
U_{17}	1000	29.817	0.457	0.061	0.062	0.012
U_{18}	1000	29.853	0.476	0.045	0.050	0.012
U_{19}	1250	37.726	0.494	0.035	0.038	0.012
U_{20}	1250	37.180	0.520	0.028	0.031	0.012
U_{21}	1250	37.701	0.540	-0.004	0.015	0.012
U_{22}	1250	37.294	0.548	-0.040	0.042	0.012
U_{23}	1250	37.707	0.586	-0.043	0.046	0.012
U_{24}	1250	37.483	0.574	-0.061	0.063	0.012
U_{25}	1500	44.443	0.598	-0.089	0.090	0.012
U_{26}	1500	44.686	0.629	-0.127	0.129	0.011
U_{27}	1500	44.952	0.639	-0.131	0.132	0.011
U_{28}	1500	45.303	0.658	-0.165	0.166	0.011
U_{29}	1500	44.985	0.668	-0.198	0.200	0.011
U_{30}	1500	44.825	0.693	-0.210	0.213	0.010
Todos		30	0.442	0.041	0.118	0.012
Pequeño		15.401	0.254	0.161	0.179	0.012
Mediano		29.911	0.436	0.066	0.085	0.012
Grande		44.866	0.648	-0.153	0.155	0.011

5.7. Estimadores de regresión logística

Resultados

Las tablas de esta sección muestran el comportamiento de los estimadores de regresión logística propuestos en la población simulada.

Tabla 5.22: Resultados de la simulación para el estimador (logistic), $\widehat{P}_{logA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.180	0.126	0.133	0.090
U_2	500	15.728	0.178	0.096	0.112	0.091
U_3	500	15.366	0.186	0.106	0.119	0.091
U_4	500	15.449	0.193	0.097	0.114	0.090
U_5	500	15.250	0.197	0.091	0.111	0.091
U_6	500	15.157	0.223	0.081	0.110	0.092
U_7	750	22.503	0.211	0.075	0.097	0.074
U_8	750	22.146	0.235	0.062	0.094	0.075
U_9	750	22.511	0.255	0.051	0.091	0.074
U_{10}	750	22.484	0.246	0.046	0.089	0.074
U_{11}	750	22.187	0.320	0.044	0.096	0.074
U_{12}	750	22.284	0.271	0.049	0.092	0.075
U_{13}	1000	29.771	0.332	0.035	0.084	0.064
U_{14}	1000	29.846	0.317	0.030	0.081	0.064
U_{15}	1000	30.055	0.342	0.011	0.080	0.064
U_{16}	1000	29.854	0.362	0.005	0.083	0.064
U_{17}	1000	29.817	0.401	0.005	0.080	0.064
U_{18}	1000	29.853	0.429	-0.002	0.079	0.064
U_{19}	1250	37.726	0.449	-0.010	0.074	0.057
U_{20}	1250	37.180	0.473	-0.019	0.075	0.057
U_{21}	1250	37.701	0.523	-0.021	0.075	0.057
U_{22}	1250	37.294	0.570	-0.017	0.075	0.057
U_{23}	1250	37.707	0.596	-0.033	0.080	0.056
U_{24}	1250	37.483	0.608	-0.026	0.076	0.056
U_{25}	1500	44.443	0.648	-0.038	0.074	0.051
U_{26}	1500	44.686	0.716	-0.041	0.071	0.051
U_{27}	1500	44.952	0.730	-0.040	0.069	0.051
U_{28}	1500	45.303	0.779	-0.045	0.067	0.051
U_{29}	1500	44.985	0.815	-0.051	0.068	0.051
U_{30}	1500	44.825	0.846	-0.057	0.069	0.051
Todos		30	0.421	0.020	0.087	0.067
Pequeño		15.401	0.193	0.100	0.116	0.091
Mediano		29.911	0.386	0.016	0.083	0.065
Grande		44.866	0.756	-0.045	0.070	0.051

Comentarios

Los cuatro estimadores de regresión logística incluyen información completa, y son variantes del modelo LGREG propuesto por Lehtonen, Särndal y Veijanen (2009). El estimador *LGREG* para el total dado por (**??**) y para la proporción dada por (**??**), cuyas variantes simuladas fueron:

- $\widehat{P}_{logA_d}$, cuyos coeficientes de regresión β se generaron por regresión lineal (gl) y son comunes a todos los dominios, estimando el argumento de la exponencial en el modelo de regresión logística mediante $x_{U_d} * \widehat{\beta}$, con información auxiliar poblacional completa.
- $\widehat{P}^{(1)}_{logA_d}$, cuyos coeficientes de regresión β se generaron por regresión lineal (gl) para cada dominio, estimando el argumento de la exponencial en el modelo de regresión logística mediante $x_{U_d} * \widehat{\beta}_d$, con información auxiliar completa de dominio.
- $\widehat{P}^{(2)}_{logA_d}$, cuyos coeficientes de regresión β se generaron por (glm) para cada dominio, es decir, en el modelo de regresión logística se usó $x_{U_d} * \widehat{\beta}_d$ para estimar el argumento de la exponencial, con información auxiliar completa de dominio.
- $\widehat{P}^{(3)}_{logA_d}$, en el cual coeficientes de regresión β se generaron por (glm) comunes a todos los dominios, es decir, en el modelo de regresión logística se usó $x_{U_d} * \widehat{\beta}$ para estimar el argumento de la exponencial, con información auxiliar poblacional completa.

Asumimos que estos estimadores son adecuados para hacer las comparaciones con nuestros estimadores, toda vez que incorporan en la fase de estimación, información auxiliar completa. A la vista de los resultados de las tablas 5.22 a 5.25, encontramos que sólo los estimadores $\widehat{P}_{logA_d}$ y $\widehat{P}^{(1)}_{logA_d}$ tienen mejores propiedades, aunque son un poco sesgados y con precisión aceptable, ya que el mayor sesgo empírico absoluto del primero es 0.126 y el menor es 0.002, mientras que para el segundo el mayor sesgo es 0.097 y el menor es 0.002. Ambos tienen una precisión similar al estimador directo (4.9) y mejor comportamiento para dominios medianos y pequeños. La última columna de la tabla, evidencia subestimación de la varianza del estimador.

Tabla 5.23: Resultados de la simulación para el estimador (logistic1), $\widehat{P}^{(1)}_{logA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.151	0.097	0.114	0.091
U_2	500	15.728	0.154	0.072	0.099	0.092
U_3	500	15.366	0.160	0.080	0.105	0.092
U_4	500	15.449	0.166	0.070	0.101	0.092
U_5	500	15.250	0.173	0.067	0.101	0.092
U_6	500	15.157	0.204	0.062	0.107	0.093
U_7	750	22.503	0.188	0.052	0.091	0.075
U_8	750	22.146	0.217	0.043	0.092	0.075
U_9	750	22.511	0.240	0.036	0.093	0.075
U_{10}	750	22.484	0.233	0.033	0.091	0.075
U_{11}	750	22.187	0.313	0.037	0.099	0.074
U_{12}	750	22.284	0.259	0.036	0.094	0.076
U_{13}	1000	29.771	0.326	0.029	0.086	0.065
U_{14}	1000	29.846	0.311	0.024	0.084	0.065
U_{15}	1000	30.055	0.339	0.008	0.084	0.064
U_{16}	1000	29.854	0.360	0.003	0.086	0.065
U_{17}	1000	29.817	0.400	0.004	0.082	0.064
U_{18}	1000	29.853	0.429	-0.002	0.080	0.064
U_{19}	1250	37.726	0.449	-0.010	0.074	0.057
U_{20}	1250	37.180	0.474	-0.019	0.074	0.057
U_{21}	1250	37.701	0.523	-0.021	0.073	0.057
U_{22}	1250	37.294	0.569	-0.018	0.074	0.056
U_{23}	1250	37.707	0.595	-0.034	0.079	0.056
U_{24}	1250	37.483	0.606	-0.028	0.074	0.056
U_{25}	1500	44.443	0.645	-0.042	0.074	0.051
U_{26}	1500	44.686	0.710	-0.047	0.072	0.050
U_{27}	1500	44.952	0.724	-0.046	0.071	0.050
U_{28}	1500	45.303	0.770	-0.053	0.071	0.050
U_{29}	1500	44.985	0.804	-0.062	0.075	0.050
U_{30}	1500	44.825	0.833	-0.070	0.079	0.049
Todos		30	0.411	0.010	0.086	0.068
Pequeño		15.401	0.168	0.075	0.104	0.092
Mediano		29.911	0.379	0.010	0.084	0.065
Grande		44.866	0.748	-0.053	0.074	0.050

Tabla 5.24: Resultados de la simulación para el estimador (logistic2), $\widehat{P}^{(2)}_{logA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.079	0.025	0.136	0.089
U_2	500	15.728	0.089	0.007	0.113	0.090
U_3	500	15.366	0.094	0.014	0.123	0.090
U_4	500	15.449	0.095	-0.001	0.123	0.090
U_5	500	15.250	0.113	0.007	0.121	0.090
U_6	500	15.157	0.157	0.015	0.132	0.091
U_7	750	22.503	0.128	-0.008	0.118	0.074
U_8	750	22.146	0.169	-0.004	0.118	0.074
U_9	750	22.511	0.203	-0.001	0.119	0.075
U_{10}	750	22.484	0.198	-0.002	0.112	0.074
U_{11}	750	22.187	0.298	0.022	0.119	0.074
U_{12}	750	22.284	0.230	0.007	0.118	0.075
U_{13}	1000	29.771	0.315	0.018	0.099	0.065
U_{14}	1000	29.846	0.297	0.010	0.098	0.065
U_{15}	1000	30.055	0.334	0.003	0.094	0.064
U_{16}	1000	29.854	0.356	-0.001	0.095	0.065
U_{17}	1000	29.817	0.399	0.003	0.087	0.064
U_{18}	1000	29.853	0.429	-0.002	0.081	0.063
U_{19}	1250	37.726	0.449	-0.010	0.074	0.056
U_{20}	1250	37.180	0.473	-0.020	0.070	0.056
U_{21}	1250	37.701	0.518	-0.026	0.069	0.055
U_{22}	1250	37.294	0.559	-0.028	0.069	0.053
U_{23}	1250	37.707	0.582	-0.047	0.078	0.053
U_{24}	1250	37.483	0.593	-0.042	0.072	0.052
U_{25}	1500	44.443	0.622	-0.064	0.081	0.046
U_{26}	1500	44.686	0.679	-0.078	0.089	0.044
U_{27}	1500	44.952	0.689	-0.081	0.091	0.043
U_{28}	1500	45.303	0.726	-0.097	0.104	0.041
U_{29}	1500	44.985	0.752	-0.114	0.118	0.040
U_{30}	1500	44.825	0.772	-0.131	0.134	0.041
Todos		30	0.380	-0.021	0.102	0.065
Pequeño		15.401	0.104	0.011	0.125	0.090
Mediano		29.911	0.363	-0.007	0.094	0.064
Grande		44.866	0.707	-0.094	0.103	0.043

Tabla 5.25: Resultados de la simulación para el estimador (logistic3), $\widehat{P}^{(3)}_{logA_d}$.

U_d	N_d	$\overline{n}_d$	$\overline{E}_d$	SE	RECME	REVE
U_1	500	15.454	0.204	0.150	0.156	0.089
U_2	500	15.728	0.195	0.113	0.127	0.090
U_3	500	15.366	0.205	0.125	0.137	0.090
U_4	500	15.449	0.212	0.116	0.131	0.089
U_5	500	15.250	0.214	0.108	0.125	0.090
U_6	500	15.157	0.238	0.096	0.121	0.090
U_7	750	22.503	0.229	0.093	0.111	0.073
U_8	750	22.146	0.250	0.076	0.103	0.073
U_9	750	22.511	0.268	0.064	0.097	0.073
U_{10}	750	22.484	0.256	0.056	0.094	0.073
U_{11}	750	22.187	0.331	0.055	0.099	0.073
U_{12}	750	22.284	0.283	0.060	0.097	0.074
U_{13}	1000	29.771	0.340	0.043	0.086	0.063
U_{14}	1000	29.846	0.325	0.038	0.082	0.063
U_{15}	1000	30.055	0.346	0.015	0.079	0.062
U_{16}	1000	29.854	0.367	0.010	0.081	0.063
U_{17}	1000	29.817	0.403	0.007	0.078	0.062
U_{18}	1000	29.853	0.430	-0.001	0.077	0.062
U_{19}	1250	37.726	0.448	-0.010	0.072	0.055
U_{20}	1250	37.180	0.470	-0.023	0.075	0.055
U_{21}	1250	37.701	0.518	-0.026	0.075	0.055
U_{22}	1250	37.294	0.565	-0.022	0.074	0.055
U_{23}	1250	37.707	0.589	-0.040	0.082	0.055
U_{24}	1250	37.483	0.602	-0.033	0.076	0.055
U_{25}	1500	44.443	0.639	-0.047	0.077	0.050
U_{26}	1500	44.686	0.706	-0.051	0.076	0.050
U_{27}	1500	44.952	0.719	-0.051	0.074	0.050
U_{28}	1500	45.303	0.766	-0.057	0.075	0.049
U_{29}	1500	44.985	0.801	-0.065	0.078	0.049
U_{30}	1500	44.825	0.832	-0.072	0.081	0.049
Todos		30	0.425	0.024	0.093	0.066
Pequeño		15.401	0.211	0.118	0.133	0.090
Mediano		29.911	0.390	0.020	0.085	0.064
Grande		44.866	0.744	-0.057	0.077	0.049

5.8. Comparación de estimadores

Resultados

Las tablas de esta sección muestran comparativas del comportamiento de todos los estimadores propuestos en la población simulada.

Tabla 5.26: Sesgo relativo empírico de los estimadores comparados. $\text{SE}/P_{A_d} \times 100\,\%$ en total y por tamaño del dominio

	Inf. auxiliar	Estimador	Total	Pequeño	Mediano	Grande
Directo		$\widehat{P}_{A_d}$	7.88	38.06	0.57	-0.37
Razón	Dominio	$\widehat{P}^{(1)}_{rA_d}$	7.81	36.96	0.71	-0.04
	Población	$\widehat{P}^{(2)}_{rA_d}$	47.80	204.22	19.8	-24.62
	Dominio sint.	$\widehat{P}^{(3)}_{rA_d}$	45.95	158.69	29.81	-18.39
Regresión	Dominio	$\widehat{P}^{(1)}_{regA_d}$	2.24	10.57	0.22	-0.02
	Población	$\widehat{P}^{(2)}_{regA_d}$	102.09	411.30	41.54	-25.47
	Dominio sint.	$\widehat{P}^{(3)}_{regA_d}$	45.95	158.69	29.81	-18.39
Diferencia	Dominio	$\widehat{P}^{(1)}_{difA_d}$	1.87	8.60	0.25	-0.02
	Población	$\widehat{P}^{(2)}_{difA_d}$	109.15	439.19	44.65	-27.36
	Dominio sint.	$\widehat{P}^{(3)}_{difA_d}$	16.84	41.88	17.76	-10.98
Comb. Razón	Dominio	$\widehat{P}^{(c_1)}_{rA_d}$	8.9	41.08	1.11	0.07
	Población	$\widehat{P}^{(c_2)}_{rA_d}$	34.53	150.65	12.9	-16.71
	Dominio sint.	$\widehat{P}^{(c_3)}_{rA_d}$	44.76	154.48	28.88	-17.31
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{rA_d}$	45.67	157.07	28.86	-15.29
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{rA_d}$	50.41	181.79	29.74	-18.93
Comb. Regresión	Dominio	$\widehat{P}^{(c_1)}_{regA_d}$	6.05	27.14	1.00	0.09
	Población	$\widehat{P}^{(c_2)}_{regA_d}$	65.90	269.56	25.71	-17.17
	Dominio sint.	$\widehat{P}^{(c_3)}_{regA_d}$	44.85	154.80	28.95	-17.39
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{regA_d}$	45.63	156.51	29.02	-15.42
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{regA_d}$	52.32	188.36	30.68	-18.82
Reg. Logística	Población	$\widehat{P}_{logA_d}$	29.22	121.35	10.14	-5.65
	Dominio	$\widehat{P}^{(1)}_{logA_d}$	21.09	91.08	6.99	-6.61
	Dominio	$\widehat{P}^{(2)}_{logA_d}$	0.03	14.65	-0.97	-11.59
	Población	$\widehat{P}^{(3)}_{logA_d}$	34.91	143.78	12.63	-7.11

Tabla 5.27: Eficiencia relativa empírica de los estimadores comparados. ECME(directo)/ECME(estimador)$\times$ 100 % de los estimadores comparados respecto al estimador directo en total y por tamaño del dominio.

	Inf. auxiliar	Estimador	Total	Pequeño	Mediano	Grande
Directo		$\widehat{P}_{A_d}$	100	100	100	100
Razón	Dominio	$\widehat{P}^{(1)}_{rA_d}$	122.96	108.65	123.25	136.40
	Población	$\widehat{P}^{(2)}_{rA_d}$	64.99	36.25	85.73	31.52
	Dominio sint.	$\widehat{P}^{(3)}_{rA_d}$	131.34	57.13	185.35	43.54
Regresión	Dominio	$\widehat{P}^{(1)}_{regA_d}$	103.26	70.56	105.39	129.55
	Población	$\widehat{P}^{(2)}_{regA_d}$	50.16	20.94	67.28	28.02
	Dominio sint.	$\widehat{P}^{(3)}_{regA_d}$	131.34	57.13	185.35	43.54
Diferencia	Dominio	$\widehat{P}^{(1)}_{difA_d}$	106.51	71.07	108.49	136.03
	Población	$\widehat{P}^{(2)}_{difA_d}$	50.44	20.64	67.99	27.62
	Dominio sint.	$\widehat{P}^{(3)}_{difA_d}$	204.25	216.84	243.08	75.17
Comb. Razón	Dominio	$\widehat{P}^{(c_1)}_{rA_d}$	117.92	105.05	118.48	129.09
	Población	$\widehat{P}^{(c_2)}_{rA_d}$	75.16	43.95	95.34	45.85
	Dominio sint.	$\widehat{P}^{(c_3)}_{rA_d}$	133.39	58.69	187.51	45.73
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{rA_d}$	133.43	57.16	187.19	48.40
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{rA_d}$	127.42	45.20	183.10	42.59
Comb. Regresión	Dominio	$\widehat{P}^{(c_1)}_{regA_d}$	115.89	95.43	118.90	127.32
	Población	$\widehat{P}^{(c_2)}_{regA_d}$	65.54	30.64	85.26	41.29
	Dominio sint.	$\widehat{P}^{(c_3)}_{regA_d}$	133.27	58.57	187.40	45.56
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{regA_d}$	132.76	57.04	186.23	48.06
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{regA_d}$	125.81	43.80	180.96	42.41
Reg. Logística	Población	$\widehat{P}_{logA_d}$	89.93	63.58	101.08	82.81
	Dominio	$\widehat{P}^{(1)}_{logA_d}$	89.83	67.67	100.26	80.70
	Dominio	$\widehat{P}^{(2)}_{logA_d}$	79.40	57.02	92.66	61.99
	Población	$\widehat{P}^{(3)}_{logA_d}$	86.06	53.21	99.94	77.24

Comentarios

Para la interpretación de los resultados, consideremos los dominios divididos en 3 categorías. Dominios pequeños (de tamaño 500), medianos (de tamaños 750, 1000 y 1250) y grandes (de tamaño 1500). Las tablas (5.26) y (5.27), (5.29) y (5.30), muestran los resultados obtenidos. En concreto, se presentan los datos correspondientes al sesgo relativo y eficiencia relativa media para cada uno de los estimadores en el total de la población y en las categorías consideradas en función del tamaño del dominio. Se han expresado el sesgo relativo en porcentaje (a la verdadera proporción en el dominio) y los errores cuadráticos medios relativos al estimador directo.

Los resultados obtenidos muestran que los estimadores propuestos $\widehat{P}^{(1)}_{rA_d}$ (4.13), $\widehat{P}^{(3)}_{rA_d}$ (4.26), $\widehat{P}^{(1)}_{regA_d}$ (4.96), $\widehat{P}^{(3)}_{regA_d}$ (4.108), $\widehat{P}^{(1)}_{difA_d}$ (4.101), $\widehat{P}^{(3)}_{difA_d}$ (4.103), $\widehat{P}^{(c_1)}_{rA_d}$(4.50), $\widehat{P}^{(c_3)}_{rA_d}$ (4.58), $\widehat{P}^{(c_4)}_{rA_d}$ (4.62), $\widehat{P}^{(c_5)}_{rA_d}$ (4.66), $\widehat{P}^{(c_1)}_{regA_d}$ (4.117), $\widehat{P}^{(c_3)}_{regA_d}$ (4.121), $\widehat{P}^{(c_4)}_{regA_d}$ (4.123) y $\widehat{P}^{(c_5)}_{regA_d}$ (4.125), suponen un aumento en la precisión media si consideramos todos los posibles dominios. Todos ellos mejoran las estimaciones en el caso de los dominios medianos, suponiendo en los casos $\widehat{P}^{(3)}_{rA_d}$ (4.26), $\widehat{P}^{(3)}_{regA_d}$ (4.108), $\widehat{P}^{(3)}_{difA_d}$ (4.103), $\widehat{P}^{(c_3)}_{rA_d}$ (4.58), $\widehat{P}^{(c_4)}_{rA_d}$ (4.62), $\widehat{P}^{(c_5)}_{rA_d}$ (4.66), $\widehat{P}^{(c_3)}_{regA_d}$ (4.121), $\widehat{P}^{(c_4)}_{regA_d}$ (4.123) y $\widehat{P}^{(c_5)}_{regA_d}$ (4.125), un aumento en precisión de más del 80 %. Mención especial merece el estimador $\widehat{P}^{(3)}_{difA_d}$ (4.103) , toda vez que aumenta a más del doble la eficiencia del estimador directo, sesgún se puede ver en la tabla 5.27.

El comportamiento de los dominios pequeños y grandes es distinto. Para los dominios pequeños se aprecia mejora en la precisión solo con los estimadores $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(3)}_{difA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$. Los estimadores $\widehat{P}^{(3)}_{rA_d}$ y $\widehat{P}^{(3)}_{regA_d}$ tienen comportamiento similar y mejoran la eficiencia del estimador directo en un 85 %, mientras que el estimador $\widehat{P}^{(3)}_{difA_d}$ duplica la eficiencia y los estimadores $\widehat{P}^{(c1)}_{rA_d}$, $\widehat{P}^{(c1)}_{regA_d}$, tienen la misma eficiencia, cuyo valor es del 18 %.

Para los dominios grandes existe un aumento en la precisión de los estimadores $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(1)}_{regA_d}$, $\widehat{P}^{(1)}_{difA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$. Por otra parte, el comportamiento de $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$, (4.13), (4.50) y (4.117), respectivamente, resultan ser buenos estimadores, toda vez que mejoran siempre la calidad de las estimaciones, en la población global simulada y en los distintos tamaños de dominio. Similar comportamiento se observa en el estimador $\widehat{P}^{(3)}_{difA_d}$, aunque su comportamiento en dominios grandes no es muy bueno.

De los modelos de regresión logística considerados, tan solo los estimadores $\widehat{P}_{logA_d}$ y $\widehat{P}^{(1)}_{logA_d}$, en el caso de dominios medianos, muestran una eficiencia relativa similar al estimador base y en cualquier caso inferior a los estimadores $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(3)}_{rA_d}$, $\widehat{P}^{(1)}_{regA_d}$, $\widehat{P}^{(3)}_{regA_d}$, $\widehat{P}^{(1)}_{difA_d}$, $\widehat{P}^{(3)}_{difA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$, $\widehat{P}^{(c_3)}_{rA_d}$, $\widehat{P}^{(c_4)}_{rA_d}$, $\widehat{P}^{(c_5)}_{rA_d}$, $\widehat{P}^{(c_1)}_{regA_d}$,

$\widehat{P}^{(c_3)}_{regA_d}$, $\widehat{P}^{(c_4)}_{regA_d}$ y $\widehat{P}^{(c_5)}_{regA_d}$.

Si consideramos ahora los valores del sesgo relativo, observamos que los estimadores $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$, no sólo mejoran el estimador base con respecto al error de estimación, sino que también mantienen, en media, el valor del sesgo, lo cual confirma que su uso proporcionará estimaciones más potentes sin repercutir en el valor del sesgo.

Ahora, si comparamos los estimadores de acuerdo con la información auxiliar disponible, encontramos que son más eficientes aquellos que hacen uso de información a nivel de dominio, en este caso, P_{B_d}, que aquellos que hacen uso de información auxiliar de nivel poblacional. Los estimadores que tienen un mejor comportamiento son $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(1)}_{regA_d}$, $\widehat{P}^{(1)}_{difA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$, tanto para la población global como para los distintos tamaños de dominio.

Si comparamos los estimadores de razón y los estimadores de regresión, encontramos que para la población global es mas eficiente $\widehat{P}^{(1)}_{rA_d}$ que $\widehat{P}^{(1)}_{regA_d}$, además, $\widehat{P}^{(3)}_{rA_d}$ y $\widehat{P}^{(3)}_{regA_d}$ son igual de eficientes. Para los dominios pequeños el estimador de razón $\widehat{P}^{(1)}_{rA_d}$ resulta ser mejor que los estimadores de regresión; para los dominios de tamaño mediano los estimadores $\widehat{P}^{(1)}_{rA_d}$ y $\widehat{P}^{(1)}_{regA_d}$, superan en eficiencia al estimador directo y $\widehat{P}^{(1)}_{rA_d}$ es superior a $\widehat{P}^{(1)}_{regA_d}$; además, los estimadores $\widehat{P}^{(3)}_{rA_d}$ y $\widehat{P}^{(3)}_{regA_d}$ son igual de eficientes y significan una aumento de aproximadamente el 70 % respecto del estimador directo. En el caso de dominios grandes se mantiene la superioridad de los estimadores de razón sobre los de regresión.

Comparando ahora los estimadores combinados de razón con los combinados de regresión, encontramos que para la población global solo los estimadores $\widehat{P}^{(c_2)}_{rA_d}$ y $\widehat{P}^{(c_2)}_{rA_d}$, no superan en eficiencia al estimador base, mientras que en los otros el comportamiento es similar, es decir, los estimadores combinados de razón y los estimadores combinados de regresión en la población global y en los diferentes tamaños de dominio, son igual de eficientes y superiores al estimador directo.

Si comparamos los estimadores de razón simple con los estimadores combinados de razón, se observa un aumento significativo en la eficiencia de los estimadores combinados y la consecuente disminución del sesgo. Este comportamiento se observa también al comparar los estimadores simples de regresión con los estimadores combinados de regresión. Entonces, se infiere que los estimadores combinados de razón y regresión mejoran la eficiencia de los estimadores de razón y regresión simples y disminuyen el sesgo. Respecto de los estimadores sintéticos $\widehat{P}^{(3)}_{rA_d}$, $\widehat{P}^{(3)}_{regA_d}$ y $\widehat{P}^{(3)}_{difA_d}$, mejoran la eficiencia del estimador directo para la población global y para dominios de tamaño mediano, aunque habrá que notar que solo el estimador de diferencia tiene un buen desempeño,

además, en dominios pequeños. Los estimadores compuestos $\widehat{P}^{(c1)}_{rA_d}$ y $\widehat{P}^{(c1)}_{regA_d}$, evidencian una ligera superioridad del primero sobre el segundo, $\widehat{P}^{(c4)}_{rA_d}$ es similar a $\widehat{P}^{(c4)}_{regA_d}$, $\widehat{P}^{(c5)}_{rA_d}$ y $\widehat{P}^{(c5)}_{regA_d}$, tienen también comportamiento similar.

Resultados de la segunda población simulada

En vista que en la población generada, según tabla 5.1, P_{A_d}, P_{B_d} y ϕ_d crecen con N_d, lo cual permite suponer que están correlados positivamente, hemos ejecutado el estudio de simulación en diferentes situaciones. Entre ellas mostramos una población (Tabla 5.28), obtenida siguiendo el mismo proceso de simulación descrito antes, pero con valores de P_{A_d} decreciendo desde $\frac{39}{41}$ hasta $\frac{10}{41}$.

Tabla 5.28: Características de la segunda población simulada.
Población generada U, de tamaño $N = 30000$, con $P_A = 0{,}3172$, $P_B = 0{,}5078$ y $\phi = 0{,}6496$. Tamaño N_d, proporción del atributo de interés P_{A_d}, del atributo auxiliar P_{B_d} y coeficiente de Cramer ϕ_d entre atributos por dominio U_d

U_d	N_d	P_{A_d}	P_{B_d}	ϕ_d
U_1	500	0.90600000	0.9560000	0.6660358
U_2	500	0.86200000	0.9360000	0.6535309
U_3	500	0.84000000	0.9060000	0.7380393
U_4	500	0.76600000	0.8600000	0.7299973
U_5	500	0.75000000	0.8640000	0.6871843
U_6	500	0.68000000	0.8220000	0.6783496
U_7	750	0.68533333	0.8000000	0.7378967
U_8	750	0.61866667	0.7866667	0.6632999
U_9	750	0.58666667	0.7493333	0.6890588
U_{10}	750	0.55866667	0.7680000	0.6183814
U_{11}	750	0.49466667	0.6800000	0.6787157
U_{12}	750	0.48666667	0.7026667	0.6333783
U_{13}	1000	0.47100000	0.6850000	0.6398713
U_{14}	1000	0.39500000	0.6430000	0.6020735
U_{15}	1000	0.39900000	0.6320000	0.6217483
U_{16}	1000	0.32900000	0.5730000	0.6044680
U_{17}	1000	0.29800000	0.5490000	0.5905296
U_{18}	1000	0.28100000	0.5330000	0.5851720
U_{19}	1250	0.26560000	0.5112000	0.5880549
U_{20}	1250	0.22640000	0.4888000	0.5532352
U_{21}	1250	0.21440000	0.4544000	0.5724396
U_{22}	1250	0.19760000	0.4360000	0.5644095
U_{23}	1250	0.16800000	0.3976000	0.5531110
U_{24}	1250	0.16240000	0.3912000	0.5493038
U_{25}	1500	0.13000000	0.3673333	0.5073052
U_{26}	1500	0.11066667	0.3420000	0.4893009
U_{27}	1500	0.09933333	0.2980000	0.5097132
U_{28}	1500	0.07800000	0.2753333	0.4718692
U_{29}	1500	0.07200000	0.2886667	0.4372507
U_{30}	1500	0.05933333	0.2600000	0.4237022

Tabla 5.29: Sesgo relativo empírico de los estimadores comparados en la segunda población.
SE/$P_{A_d}\times$ 100 % en total y por tamaño del dominio

	Inf. auxiliar	Estimador	Total	Pequeño	Mediano	Grande
Directo		$\widehat{P}_{A_d}$	0.09	-3.01	0.32	2.51
Razón	Dominio	$\widehat{P}^{(1)}_{rA_d}$	0.55	0.04	0.23	2.02
	Población	$\widehat{P}^{(2)}_{rA_d}$	3.04	-40.87	-7.40	78.31
	Dominio sint.	$\widehat{P}^{(3)}_{rA_d}$	18.47	-32.45	4.94	109.97
Regresión	Dominio	$\widehat{P}^{(1)}_{regA_d}$	0.09	-2.26	0.26	1.96
	Población	$\widehat{P}^{(2)}_{regA_d}$	9.56	-12.65	-0.90	63.17
	Dominio sint.	$\widehat{P}^{(3)}_{regA_d}$	18.47	-32.45	4.94	109.97
Diferencia	Dominio	$\widehat{P}^{(1)}_{difA_d}$	-0.17	0.34	0.01	-1.21
	Población	$\widehat{P}^{(2)}_{difA_d}$	41.34	-45.20	-5.05	267.06
	Dominio sint.	$\widehat{P}^{(3)}_{difA_d}$	2.13	-14.43	6.99	4.08
Comb. Razón	Dominio	$\widehat{P}^{(c_1)}_{rA_d}$	0.02	-0.54	-0.28	1.48
	Población	$\widehat{P}^{(c_2)}_{rA_d}$	-7.64	-30.57	-6.11	10.72
	Dominio sint.	$\widehat{P}^{(c_3)}_{rA_d}$	17.54	-32.27	4.78	105.61
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{rA_d}$	21.08	-20.12	5.35	109.47
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{rA_d}$	18.29	-33.27	4.97	109.78
Comb. Regresión	Dominio	$\widehat{P}^{(c_1)}_{regA_d}$	0.25	-2.34	0.41	2.36
	Población	$\widehat{P}^{(c_2)}_{regA_d}$	5.10	-10.14	-0.51	37.19
	Dominio sint.	$\widehat{P}^{(c_3)}_{regA_d}$	15.96	-31.92	4.50	98.22
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{regA_d}$	18.02	-20.92	4.86	96.45
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{regA_d}$	18.76	-24.05	4.84	103.34
Reg. Logística	Población	$\widehat{P}_{logA_d}$	4.64	-7.52	-0.64	32.65
	Dominio	$\widehat{P}^{(1)}_{logA_d}$	-0.13	-10.01	-1.58	14.08
	Dominio	$\widehat{P}^{(2)}_{logA_d}$	-17.30	-17.40	-4.56	-55.40
	Población	$\widehat{P}^{(3)}_{logA_d}$	3.07	-5.88	-0.41	22.47

Tabla 5.30: Eficiencia relativa empírica de los estimadores comparados en la segunda población.
ECME(directo)/ECME(estimador)$\times$ 100 % de los estimadores comparados respecto al estimador directo en total y por tamaño del dominio.

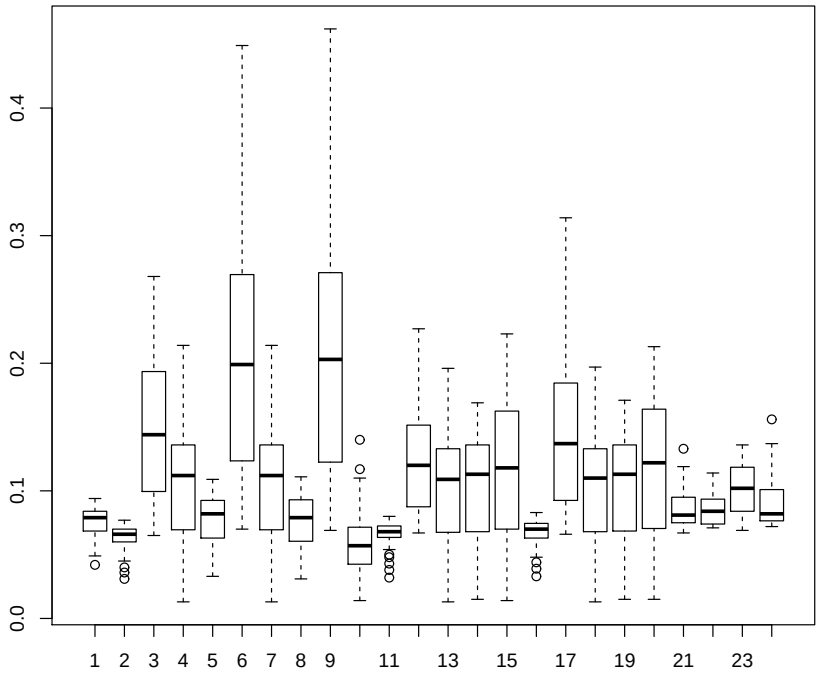
	Información auxiliar	Estimador	Total	Pequeño	Mediano	Grande
Directo		$\widehat{P}_{A_d}$	100	100	100	100
Razón	Dominio	$\widehat{P}^{(1)}_{rA_d}$	124.44	134.74	125.06	112.29
	Población	$\widehat{P}^{(2)}_{rA_d}$	66.49	31.47	85.10	45.70
	Dominio sint.	$\widehat{P}^{(3)}_{rA_d}$	122.24	40.36	175.26	45.05
Regresión	Dominio	$\widehat{P}^{(1)}_{regA_d}$	112.63	112.56	113.20	110.97
	Población	$\widehat{P}^{(2)}_{regA_d}$	89.08	72.74	103.71	61.55
	Dominio sint.	$\widehat{P}^{(3)}_{regA_d}$	122.24	40.36	175.26	45.05
Diferencia	Dominio	$\widehat{P}^{(1)}_{difA_d}$	108.19	139.53	111.10	68.13
	Población	$\widehat{P}^{(2)}_{difA_d}$	47.28	27.72	63.13	19.28
	Dominio sint.	$\widehat{P}^{(3)}_{difA_d}$	232.07	95.00	286.14	206.91
Comb. Razón	Dominio	$\widehat{P}^{(c_1)}_{rA_d}$	113.96	121.70	114.57	104.40
	Población	$\widehat{P}^{(c_2)}_{rA_d}$	87.74	42.95	101.87	90.16
	Dominio sint.	$\widehat{P}^{(c_3)}_{rA_d}$	123.23	40.57	176.30	46.67
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{rA_d}$	121.94	46.02	172.82	45.23
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{rA_d}$	121.47	39.88	174.14	45.02
Comb. Regresión	Dominio	$\widehat{P}^{(c_1)}_{regA_d}$	108.60	109.13	108.58	108.15
	Población	$\widehat{P}^{(c_2)}_{regA_d}$	95.09	82.75	105.03	77.62
	Dominio sint.	$\widehat{P}^{(c_3)}_{regA_d}$	124.84	40.97	177.88	49.62
	Dominio-sint.(D)	$\widehat{P}^{(c_4)}_{regA_d}$	126.11	46.88	177.72	50.51
	Población-sint.(D)	$\widehat{P}^{(c_5)}_{regA_d}$	124.83	46.64	176.64	47.58
Reg. Logística	Población	$\widehat{P}_{logA_d}$	100.16	90.38	107.52	87.85
	Dominio	$\widehat{P}^{(1)}_{logA_d}$	99.48	84.98	107.73	89.22
	Dominio	$\widehat{P}^{(2)}_{logA_d}$	87.53	68.54	106.11	50.79
	Población	$\widehat{P}^{(3)}_{logA_d}$	101.49	95.15	106.10	94.02

Comentarios sobre la segunda población

Se desprende de las tablas 5.29 y 5.30 que los estimadores estudiados en esta segunda población simulada, tienen un comportamiento similar en ambas poblaciones, aunque aquí surge mayor evidencia para descartar el estimador de diferencia sintético $\widehat{P}^{(3)}_{difA_d}$ (4.103), como estimador adecuado para dominios pequeños, toda vez que su eficiencia es menor que la del estimador de diferencia $\widehat{P}^{(1)}_{difA_d}$ (4.101). A la vista de los resultados, identificamos como los mejores estimadores para los diferentes tamaños de dominio a $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(1)}_{regA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$, cuyas eficiencias alcanzan hasta el 34 % y son los de menor sesgo relativo. Se ha incluido el estimador $\widehat{P}^{(1)}_{regA_d}$, que tiene una eficiencia menor a los otros tres estimadores, porque en todas la simulaciones realizadas refleja siempre una eficiencia superior al del estimador directo.

Por lo anterior, los comentarios válidos para los estimadores de la primera población simulada, valen para los estimadores en la segunda población, con excepción del estimador $\widehat{P}^{(3)}_{difA_d}$ (4.103).

Figura 5.1: Gráfico de cajas de las RECME de los estimadores. El gráfico ofrece una imágen visual de la variabilidad de los estimadores y se confirma que la varianza empírica es menor para los estimadores 2, 5, 11 y 16, que corresponden a los estimadores $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(1)}_{regA_d}$, $\widehat{P}^{(c1)}_{rA_d}$ y $\widehat{P}^{(c1)}_{regA_d}$.



5.9. Conclusiones del estudio de simulación

Toda vez que nuestro objetivo consistió en hallar estimadores para proporciones en áreas pequeñas, debido a que es un importante tema en muchas aplicaciones prácticas, propusimos varios estimadores que hacen uso de información auxiliar a nivel de población y a nivel de dominio, en vista de que la estimación directa de dominios no proporciona buenas estimaciones. El estudio de simulación realizado con los estimadores propuestos, aporta evidencia para afirmar que los estimadores $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(c1)}_{rA_d}$, $\widehat{P}^{(1)}_{regA_d}$ y $\widehat{P}^{(c1)}_{regA_d}$, son eficientes y muy poco sesgados, por tanto, pueden usarse en la estimación de proporciones de áreas pequeñas y en cualquier tamaño de dominio. Debemos destacar que estos tres estimadores hacen uso de información auxiliar en forma de proporciones de dominio, es decir, se conoce P_{B_d}, la proporción del atributo auxiliar B a nivel de dominio, que se asume asociado al atributo A.

5.10. Estimación con datos de dengue

Los datos

La aplicación de los estimadores propuestos a un problema con datos reales, se hace con una base de datos de casos confirmados de dengue por el laboratorio estatal de salud pública del Estado de Guerrero (LESPEG), México (2006). La información contiene registros de los pacientes diagnosticados con dengue y características asociadas a los síntomas de un cuadro típico de dengue, organizados para las 7 regiones del estado. La tabla (5.31) muestra los tamaños poblacionales de esta regiones, que serán consideradas como dominios. Como variable principal consideraremos el tipo de dengue sufrido por el paciente (clásico o hemorrágico), donde $Y = 1$ si el paciente presenta dengue clásico y $Y = 0$, si presenta dengue hemorrágico. Para la selección de la variable auxiliar, tomamos en consideración la relación entre la variable principal y las distintas variables auxiliares, a partir de un modelo de regresión logística binaria. Las variables que mejor clasifican a los pacientes son sexo, enfermos similares en la localidad, cefalea, mialgias, artralgias, vómito, náuseas, dolor abdominal, diarrea, congestión nasal, escape de líquidos, hemorragias, petequias y melena. De entre las variables con mayor correlación, seleccionamos artralgias y realizamos las estimaciones con una muestra disponible de tamaño 300. La metodología, en este caso, al tratarse de una muestra real, consiste en estimar el error cuadrático medio de cada uno de los estimadores mediante Jackknife. Posteriormente, definimos la eficiencia relativa tomando como base el estimador directo.

Tabla 5.31: Tamaños poblacionales de la población "dengue" en las regiones.

Región	Tamaño poblacional.
1. Tierra caliente	40
2. Zona norte	211
3. Zona centro	160
4. Montaña	412
5. Costa grande	354
6. Costa chica	74
7. Acapulco	964

Resultados

Seleccionada la muestra, encontramos que todas las submuestras correspondientes a las regiones eran no vacías. Los tamaños de las submuestras para cada dominio fueron $n_1 = 2$, $n_2 = 37$, $n_3 = 18$, $n_4 = 52$, $n_5 = 46$, $n_6 = 7$ y $n_7 = 138$. El dominio 1 solo tiene casos de dengue clásico y el dominio 6 tiene de los dos tipos, pero con solo 5 casos de dengue hemorrágico. Centraremos nuestro estudio en el dominio 6, que es uno de los casos extremos que se pueden presentar en la selección de la muestra, toda vez que todos los valores correspondientes a la variable respuesta en la submuestra, resultaron ser unos, es decir, todos los pacientes en la submuestra están infectados de dengue clásico. Existen muchos casos en la práctica en los que podemos encontrar esta situación debido a la ocurrencia de factores relacionados a las características específicas del dominio, tales como una proporción a nivel dominio muy cercana a 1. En este caso, el uso de estimadores de regresión logística no es factible. Por otra parte, cabe esperar que el estimador base, que solo utiliza la información muestral de dominio, no ofrezca estimaciones precisas. La tabla (5.32) muestra las eficiencias relativas de los estimadores propuestos.

Como puede observarse la mayoría de los estimadores propuestos mejoran al estimador base con respecto a la eficiencia; en el caso de los estimadores $\widehat{P}^{(c_4)}_{regA_d}$ y $\widehat{P}^{(c_5)}_{regA_d}$, la ganancia en eficiencia alcanza el 56.34 %. De especial interés para nosotros es el estimador $\widehat{P}^{(1)}_{rA_d}$, toda vez que lo clasificamos como eficiente, y con esta muestra mejora la eficiencia en un 47.29 %, que sabemos usa información auxiliar a nivel de dominio. Los estimadores $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$, aunque su eficiencia no es tan grande como en la población simulada, aquí habría que destacar que, aún con este caso extremo de muestra y con submuestras por dominio muy pequeñas, son ligeramente más eficientes que el estimador directo.

Esta situación nos hace deducir que precisamente la inclusión de la razón entre la proporción poblacional y muestral de la variable auxiliar en el dominio es lo que mejora la estimación. En resumen, podemos concluir, que el problema inicial de tener una submuestra en el dominio compuesta solo por unos, puede

solventarse incluyendo en las estimaciones la información proporcionada por la variable auxiliar.

Tabla 5.32: Eficiencia relativa estimada de los estimadores para la población "dengue" en el dominio U_6 (Costa chica).
Población U de tamaño $N = 2215$, con $n = 300$.

	Estimador	Eficiencia relativa
Directo	$\widehat{P}_{A_d}$	100.00
Razón	$\widehat{P}^{(1)}_{rA_d}$	144.29
	$\widehat{P}^{(2)}_{rA_d}$	147.16
	$\widehat{P}^{(3)}_{rA_d}$	8.41
Regresión	$\widehat{P}^{(1)}_{regA_d}$	108.06
	$\widehat{P}^{(2)}_{regA_d}$	108.48
	$\widehat{P}^{(3)}_{regA_d}$	12.45
Comb. Razón	$\widehat{P}^{(c_1)}_{rA_d}$	100.73
	$\widehat{P}^{(c_2)}_{rA_d}$	100.77
	$\widehat{P}^{(c_3)}_{rA_d}$	5.03
	$\widehat{P}^{(c_4)}_{rA_d}$	105.61
	$\widehat{P}^{(c_5)}_{rA_d}$	105.61
Comb. Regresión	$\widehat{P}^{(c_1)}_{regA_d}$	100.17
	$\widehat{P}^{(c_2)}_{regA_d}$	100.17
	$\widehat{P}^{(c_3)}_{regA_d}$	5.03
	$\widehat{P}^{(c_4)}_{regA_d}$	156.34
	$\widehat{P}^{(c_5)}_{regA_d}$	156.34

Estimación de proporciones en dominios

Se usaron los estimadores con mejores propiedades en la estimación de proporciones de pacientes con dengue clásico por regiones. La tabla 5.33 permite observar las proporciones estimadas para cada dominio. La estimación de proporciones en el dominio 1, que corresponde a la región Tierra caliente, no se ha realizado, toda vez que la muestra de dominio es demasiado pequeña ($n_1 = 2$) y todos los casos fueron confirmados con dengue clásico.

Tabla 5.33: Estimación de proporciones y varianzas por regiones.

	Estimador	p	$\widehat{\sigma}^2$
Tierra caliente	$\widehat{P}^{(1)}_{rA_d}$	NA	NA
	$\widehat{P}^{(1)}_{regA_d}$	NA	NA
	$\widehat{P}^{(c1)}_{rA_d}$	NA	NA
	$\widehat{P}^{(c1)}_{regA_d}$	NA	NA
Zona norte	$\widehat{P}^{(1)}_{rA_d}$	0.800	0.007
	$\widehat{P}^{(1)}_{regA_d}$	0.802	0.003
	$\widehat{P}^{(c1)}_{rA_d}$	0.841	0.002
	$\widehat{P}^{(c1)}_{regA_d}$	0.832	0.002
Zona centro	$\widehat{P}^{(1)}_{rA_d}$	0.791	0.009
	$\widehat{P}^{(1)}_{regA_d}$	0.792	0.005
	$\widehat{P}^{(c1)}_{rA_d}$	0.838	0.003
	$\widehat{P}^{(c1)}_{regA_d}$	0.827	0.002
Montaña	$\widehat{P}^{(1)}_{rA_d}$	0.994	0.004
	$\widehat{P}^{(1)}_{regA_d}$	0.991	0.002
	$\widehat{P}^{(c1)}_{rA_d}$	0.984	0.000
	$\widehat{P}^{(c1)}_{regA_d}$	0.999	0.000
Costa grande	$\widehat{P}^{(1)}_{rA_d}$	0.878	0.003
	$\widehat{P}^{(1)}_{regA_d}$	0.883	0.001
	$\widehat{P}^{(c1)}_{rA_d}$	0.910	0.001
	$\widehat{P}^{(c1)}_{regA_d}$	0.902	0.000
Costa chica	$\widehat{P}^{(1)}_{rA_d}$	0.925	0.008
	$\widehat{P}^{(1)}_{regA_d}$	0.923	0.004
	$\widehat{P}^{(c1)}_{rA_d}$	0.916	0.004
	$\widehat{P}^{(c1)}_{regA_d}$	0.918	0.002
Acapulco	$\widehat{P}^{(1)}_{rA_d}$	0.494	0.002
	$\widehat{P}^{(1)}_{regA_d}$	0.488	0.002
	$\widehat{P}^{(c1)}_{rA_d}$	0.500	0.001
	$\widehat{P}^{(c1)}_{regA_d}$	0.496	0.001

5.11. Conclusiones generales

Los estimadores de áreas pequeñas para variables cuantitativas, han demostrado su eficiencia en distintas aplicaciones debido a la naturaleza de las variables de estudio y a las variables auxiliares involucradas. De hecho, existe toda una fundamentación teórica sobre los métodos y sobre los modelos supuestos en las inferencias asociadas. Rao (2003) hace un recuento de estos métodos e incluye algunos que se han pensado para variables binarias, pero con información auxiliar de naturaleza continua, bajo un enfoque bayesiano. Se han probado distintos modelos de enlace entre las unidades de los dominios y las distribuciones a posteriori, que hará falta probar para datos que se distribuyen según alguna función de distribución continua no estudiada, para agotar la aplicación de los modelos jerárquicos de bayes a la estimación en áreas pequeñas. En general, las estimaciones de áreas pequeñas con variables cuantitativas funcionan muy bien, siempre que el tamaño de la submuestra en el dominio sea mayor que 2 y se cuente con información auxiliar proveniente de registros administrativos, un censo, estudios previos o de otros dominios con características similares, en forma de totales poblacionales o de dominio.

Hasta ahora, no hay ninguna discusión completa sobre cual es la especificación lineal o no lineal más apropiada cuando se necesitan estimaciones de áreas pequeñas para variables no continuas, es decir, para datos de frecuencias o datos categóricos. Los especialistas en encuestas saben que estas variables surgen con mucha frecuencia y, aún más, las variables auxiliares asociadas a tales variables también pueden proceder de datos no continuos. El problema de estimación se complica porque la aplicación de los métodos desarrollados para variables continuas no son necesariamente válidos para variables no continuas. Se justifica el cálculo de medias, que para este tipo de variables son proporciones de atributos, correlaciones, covarianzas, disimilaridades y otras medidas, pero, resulta complicado y hasta incompresible, la justificación teórica de un modelos lineal para este tipo de variables, sin antes haber realizado con ellos alguna transformación. Sin embargo, en años recientes, se ha justificado la aplicación de estimadores en la estimación de proporciones, como se puede ver en Murgi, S.J. et al. (1995) y una serie de trabajos de Rueda, et al. (2008, 2009, 2010 y 2011).

Según nuestros objetivos, hemos propuesto estimadores de razón, de regresión, de diferencia y de calibración para la estimación de proporciones en dominios, incorporando información auxiliar en forma de totales o proporciones en la fase de estimación. La ventaja de los estimadores propuestos, frente a los métodos de estimación indirecta basada en modelos bajo el enfoque clásico o bayesiano, predominantes en la teoría estadística de las últimas décadas para la estimación en áreas pequeñas, es que se obtienen buenas estimaciones con

la información auxiliar disponible en forma de totales o proporciones, provenientes de censos, estudios previos o registros administrativos, sin demandar el uso de información completa. Su utilidad es manifiesta cuando por razones ajenas al investigador, por ejemplo, confidencialidad de la información relacionada con personas o instituciones públicas o privadas, no es posible contar con información auxiliar completa relacionada a la variable de interés. Uno de los inconveniente principales que podemos señalar, es que la precisión de las estimaciones es menor que de aquellos que usan información auxiliar completa, por lo que, para que nuestros estimadores sean al menos igual de precisos que los existentes en la literatura actual para la estimación en áreas pequeñas, es necesario incorporar a nuestros estimadores información auxiliar de varias variables auxiliares relacionadas con la variable de interés o información completa de tales variables auxiliares.

Se ha comprobado la eficiencia de los estimadores propuestos mediante un estudio de simulación y, según evidencia proporcionada por las simulaciones, únicamente los estimadores (4.13), (4.50), (4.96) y (4.117)

$$\widehat{P}^{(1)}_{rA_d} = \widehat{R}_d P_{B_d} = \frac{\widehat{P}_{A_d}}{\widehat{P}_{B_d}} P_{B_d}$$

$$\widehat{P}^{(1)}_{regA_d} = \widehat{P}_{A_d} + \widehat{b}_{opt}(P_{B_d} - \widehat{P}_{B_d})$$

$$\widehat{P}^{(c1)}_{rA_d} = \widehat{\alpha}_{op}\widehat{P}_{A_d} + (1 - \widehat{\alpha})_{opt}\widehat{P}^{(1)}_{rA_d},$$

y

$$\widehat{P}^{(c1)}_{regA_d} = \widehat{\alpha}_{opt}\widehat{P}_{A_d} + (1 - \widehat{\alpha}_{opt})\widehat{P}^{(1)}_{regA_d},$$

mejoran la eficiencia en la estimación, respecto del estimador base, en los dominios pequeños, aunque merece mención especial el estimador de razón, toda vez que su eficiencia es superior al estimador de regresión. En los dominios clasificados como medianos funcionan muy bien la mayoría de estimadores, con excepción de $\widehat{P}^{(2)}_{rA_d}$, $\widehat{P}^{(2)}_{regA_d}$, $\widehat{P}^{(1)}_{difA_d}$, $\widehat{P}^{(c_2)}_{rA_d}$ y $\widehat{P}^{(c_2)}_{regA_d}$. El resultado obtenido en los dominios clasificados como grandes, refuerza la idea que los estimadores propuestos funcionan mucho mejor para dominios medianos y pequeños que para dominios grandes.

Entonces, si disponemos de información auxiliar de dominio, los estimadores apropiados serán $\widehat{P}^{(1)}_{rA_d}$, $\widehat{P}^{(1)}_{regA_d}$, $\widehat{P}^{(c_1)}_{rA_d}$ y $\widehat{P}^{(c_1)}_{regA_d}$, porque se ha demostrado que mejoran la eficiencia como mínimo en un 27 % y como máximo un 36 %, mientras que el resto, tiene un comportamiento menos eficiente.

En situaciones prácticas, no es fácil tener acceso a información auxiliar a nivel de dominio y solo se dispone de algunas medidas de posición a nivel de la población. En tal caso, aunque la estimación será menos precisa, los estimadores apropiados que deben emplearse son los sintéticos, porque s_d, el tamaño

de la muestra en el dominio es aleatoria y mejorará significativamente la estimación al aumentar el tamaño de la submuestra. Los estimadores sintéticos que podrían emplearse son los de razón, de regresión y diferencia, es decir $\widehat{P}^{(3)}_{rA_d}$ (4.26), $\widehat{P}^{(3)}_{regA_d}$ (4.108)y $\widehat{P}^{(3)}_{difA_d}$ (4.103).

Capítulo 6

Apéndice

A: Estimación del sesgo y el MSE

Dada una proporción $\widehat{p}_{rA_d}$, de un estimador de razón, el sesgo y su error cuadrático medio estará dado por

$$B(\widehat{p}_{rA_d}) = P_{A_d}\left[\frac{V(\widehat{p}_{B_d})}{P_{B_d}^2} - \frac{cov(\widehat{p}_{A_d},\widehat{p}_{B_d})}{P_{A_d}P_{B_d}}\right]$$

y

$$MSE(\widehat{p}_{rA_d}) = P_{A_d}^2\left[\frac{V(\widehat{p}_{A_d})}{P_{A_d}^2} + \frac{V(\widehat{p}_{B_d})}{P_{B_d}^2} - \frac{2cov(\widehat{p}_{A_d},\widehat{p}_{B_d})}{P_{A_d}P_{B_d}}\right]$$

Demostración:

Sea $e_0 = \frac{\widehat{p}_{A_d}-P_{A_d}}{P_{A_d}}$ y $e_1 = \frac{\widehat{p}_{B_d}-P_{B_d}}{P_{B_d}}$, puede notarse que

1. $E(e_0) = E\left[\frac{\widehat{p}_{A_d}-P_{A_d}}{P_{A_d}}\right] = 0$
2. $E(e_1) = E\left[\frac{\widehat{p}_{B_d}-P_{B_d}}{P_{B_d}}\right] = 0$
3. $E(e_0^2) = E\left[\frac{\widehat{p}_{A_d}-P_{A_d}}{P_{A_d}}\right]^2 = \frac{V(\widehat{p}_{A_d})}{P_{A_d}^2}$
4. $E(e_1^2) = E\left[\frac{\widehat{p}_{B_d}-P_{B_d}}{P_{B_d}}\right]^2 = \frac{V(\widehat{p}_{B_d})}{P_{B_d}^2}$
5. $E(e_0,e_1) = E\left[\frac{(\widehat{p}_{A_d}-P_{A_d})(\widehat{p}_{B_d}-P_{B_d})}{P_{A_d}P_{B_d}}\right] = \frac{cov(\widehat{p}_{A_d},\widehat{p}_{B_d})}{P_{A_d}P_{B_d}}$

Asumimos que el tamaño de muestra es suficientemente grande, así que $|e_0| < 1$ y $|e_1| < 1$. Esto es equivalente a asumir que para todas las muestras posibles $0 < \widehat{p}_{A_d} < 2P_{A_d}$ y $0 < \widehat{p}_{B_d} < 2P_{B_d}$.

Dado que $\widehat{p}_{A_d} = P_{A_d}(1 + e_0)$ y $\widehat{p}_{B_d} = P_{B_d}(1 + e_1)$, el estimador $\widehat{p}_{rA_d}$ puede escribirse como

$$\widehat{p}_{rA_d} = \frac{P_{A_d}(1 + e_0)}{P_{B_d}(1 + e_1)} P_{B_d} = P_{A_d}(1 + e_0)(1 + e_1)^{-1}$$

que mediante un desarrollo en series de potencias da

$$\widehat{p}_{rA_d} = P_{A_d}(1 + e_0)(1 - e_1 + e_1^2 - e_1^3 + \cdots)$$

$$\widehat{p}_{rA_d} = P_{A_d}(1 - e_1 + e_1^2 - e_1^3 + \cdots + e_0 - e_0e_1 + e_0e_1^2 - e_0e_1^3 + \cdots)$$

Usando 1 y 2 e ignorando los términos de grado mayor que dos conseguimos

$$E[\widehat{p}_{rA_d} - P_{A_d}] = P_{A_d}E[e_1^2 - e_0e_1] = P_{A_d}\left[\frac{V(\widehat{p}_{B_d})}{P_{B_d}^2} - \frac{cov(\widehat{p}_{A_d}, \widehat{p}_{B_d})}{P_{A_d}P_{B_d}}\right]$$

que es la expesión para el sesgo.

Ahora, procediendo de manera análoga, obtenemos la expresión para el error cuadrático medio

$$E[\widehat{p}_{rA_d} - P_{A_d}]^2 = P_{A_d}^2[e_0^2 + e_1^2 - 2e_0e_1]$$

$$= P_{A_d}^2\left[\frac{V(\widehat{p}_{A_d})}{P_{A_d}^2} + \frac{V(\widehat{p}_{B_d})}{P_{B_d}^2} - \frac{2cov(\widehat{p}_{A_d}, \widehat{p}_{B_d})}{P_{A_d}P_{B_d}}\right]$$

$$MSE(\widehat{p}_{rA_d}) = V(\widehat{p}_{A_d}) + R_d^2V(\widehat{p}_{B_d}) - 2R_dcov(\widehat{p}_{A_d}, \widehat{p}_{B_d}).$$

Para un dominio particular, consideremos los datos dispuestos en una tabla de doble entrada como se muestra en cuadro (6.1), donde $N_{1\cdot} = \sum_{i=1}^{N_d} A_d$ es el

Tabla 6.1: Clasificación a nivel de dominio.

	B	B^c	Totales
A	N_{11}	N_{12}	$N_{1\cdot}$
A^c	N_{21}	N_{22}	$N_{2\cdot}$
Totales	$N_{\cdot 1}$	$N_{\cdot 2}$	N_d

número de unidades en la población que posee el atributo A, $N_{2\cdot}$ es el número de unidades en la población que no posee el atributo A, $N_{\cdot 1}$ es el número de

unidades de la población que posee el atributo B y $N_{\cdot 2}$ es el número de unidades de la población que no posee el atributo B. Análogamente, N_{11} es el número de unidades en la población que poseen simultáneamente los atributos A y B, N_{12} es el número de unidades de la población que poseen simultáneamente los atributos A y B^c, etc. Esta clasificación puede definirse a nivel de la muestra como se observa en el cuadro (6.2).

Tabla 6.2: Clasificación a nivel de la muestra.

	B	B^c	Totales
A	n_{11}	n_{12}	$n_{1\cdot}$
A^c	n_{21}	n_{22}	$n_{2\cdot}$
Totales	$n_{\cdot 1}$	$n_{\cdot 2}$	n_d

Toda vez que suponemos A y B relacionado, el coeficiente V de Cramer estará dado por

$$\phi = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}}$$

y

$$cov(N_d\widehat{p}_{A_d}, N_d\widehat{p}_{B_d}) = cov(n_{1\cdot}, n_{\cdot 1}) = cov(n_{11} + n_{12}, n_{11} + n_{21})$$
$$= V(n_{11}) + cov(n_{11}, n_{12}) + cov(n_{11}, n_{21}) + cov(n_{12}, n_{21})$$

Las variables $(n_{11}, n_{12}, n_{21}, n_{22}) \simeq HG(N_d, n_d, N_{11}, N_{12}, N_{21})$, por lo que

- $V(n_{11}) = \frac{N_d - n_d}{N_d - 1} n_d \frac{N_{11}}{N_d}(1 - \frac{N_{11}}{N_d})$,
- $cov(n_{11}, n_{12}) = -\frac{N_d - n_d}{N_d - 1} n_d \frac{N_{11}N_{12}}{N_d^2}$
- $cov(n_{11}, n_{21}) = -\frac{N_d - n_d}{N_d - 1} n_d \frac{N_{11}N_{21}}{N_d^2}$
- $cov(n_{12}, n_{21}) = \frac{N_d - n_d}{N_d - 1} n_d \frac{N_{12}N_{21}}{N_d^2}$

y al expresión para la covarianza se escribe como

$$cov(N_d\widehat{p}_{A_d}, N_d\widehat{p}_{B_d}) = \frac{N}{N_d - 1}\frac{1-f}{n}\left[\frac{N_{11}}{N_d}(1 - \frac{N_{11}}{N_d}) - \frac{N_{11}N_{12}}{N_d^2} - \frac{N_{11}N_{21}}{N_d^2} - \frac{N_{12}N_{21}}{N_d^2}\right]$$

$$= \frac{N}{N_d - 1}\frac{1-f}{n}\left[\frac{N_{11}N_d - N_{11}^2 - N_{11}N_{12} - N_{11}N_{21} - N_{12}N_{21}}{N_d^2}\right]$$

pero como $N_d = N_{11} + N_{12} + N_{21} + N_{22}$, sustituyendo se llega a la expresión

$$cov(N_d\widehat{p}_{A_d}, N_d\widehat{p}_{B_d}) = \frac{N}{N_d - 1}\frac{1-f}{n}\left[\frac{N_{11}N_{22} - N_{12}N_{21}}{N_d^2}\right].$$

Del coeficiente V de Cramer se obtiene,

$$N_{11}N_{22} - N_{12}N_{21} = \phi\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}$$

por lo que si se sustituye en la expresión anterior obtenemos una última expresión para la covarianza

$$cov(N_d\widehat{p}_{A_d}, N_d\widehat{p}_{B_d}) = \frac{N}{N_d - 1}\frac{1-f}{n}\left[\frac{\phi\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}}{N_d^2}\right]$$

$$= \frac{N}{N_d - 1}\frac{1-f}{n}\phi\sqrt{\frac{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}{N_d^4}} = \frac{N}{N_d - 1}\frac{1-f}{n}\phi\sqrt{\frac{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}{N_dN_dN_dN_d}}$$

$$cov(N_d\widehat{p}_{A_d}, N_d\widehat{p}_{B_d}) = \frac{N}{N_d - 1}\frac{1-f}{n}\phi\sqrt{P_{A_d}Q_{A_d}P_{B_d}Q_{B_d}}$$

Selección de la matriz de pesos w

El criterio para la optimalidad del vector de pesos $\mathbf{w} = (w_1, \cdots, w_p)$ con $\sum w_i = 1$ es minimizar $V(\widetilde{y})$. Para obtener el extremo, hacemos uso de la desigualdad generalizada de Cauchy-Schwarz.

LEMMA.

$$(xy')^2 \leq (xMx')(yM^{-1}y'),$$

donde M es una matriz simétrica definida positiva. La igualdad se mantiene si y solo si $xM = \theta y$, donde $\theta \neq 0$ es un escalar.

Para aplicar el lemma, fijemos $e = (1, \cdots, 1)$ y hagamos la correspondencia $x = w, y = e, M = A$.

Así

$$1 = (we')^2 \leq (wAw')(eA^{-1}e')$$

y la igualdad se consigue si y solo si $wA = \theta e$ o $w = \theta eA^{-1}$. Por la restricción $we' = 1$, se sigue que $\theta = 1/(eA^{-1}e')$, y entonces el peso óptimo está dado por

$$\widehat{w} = \frac{eA^{-1}}{eA^{-1}e'}.$$

B: Programa en R para los cálculos de las simulaciones

```
    #Población simulada
nsim <- 1000
pop.size1<-c(rep(500,6))
pop.size2<-c(rep(750,6))
pop.size3<-c(rep(1000,6))
pop.size4<-c(rep(1250,6))
pop.size5<-c(rep(1500,6))
pop.size<-c(pop.size1,pop.size2,pop.size3,pop.size4,pop.size5)
(10:39)/41->prop
library(car)
library(MASS)
library(nnet)
library(splines)
library(survival)

x1<-rbinom(pop.size[1],1,prop[1]); y1s<-rbinom(pop.size[1],1,prop[1])
recode(x1, "0=2")->x1r; recode(y1s, "0=2")->y1r; recode(x1*y1s, "2=1; 4=0")->y1
x2<-rbinom(pop.size[2],1,prop[2]); y2s<-rbinom(pop.size[2],1,prop[2])
recode(x2, "0=2")->x2r; recode(y2s, "0=2")->y2r; recode(x2*y2s, "2=1; 4=0")->y2
x3<-rbinom(pop.size[3],1,prop[3]); y3s<-rbinom(pop.size[3],1,prop[3])
recode(x3, "0=2")->x3r; recode(y3s, "0=2")->y3r; recode(x3*y3s, "2=1; 4=0")->y3
x4<-rbinom(pop.size[4],1,prop[4]); y4s<-rbinom(pop.size[4],1,prop[4])
recode(x4, "0=2")->x4r; recode(y4s, "0=2")->y4r; recode(x4*y4s, "2=1; 4=0")->y4
x5<-rbinom(pop.size[5],1,prop[5]); y5s<-rbinom(pop.size[5],1,prop[5])
recode(x5, "0=2")->x5r; recode(y5s, "0=2")->y5r; recode(x5*y5s, "2=1; 4=0")->y5
x6<-rbinom(pop.size[6],1,prop[6]); y6s<-rbinom(pop.size[6],1,prop[6])
recode(x6, "0=2")->x6r; recode(y6s, "0=2")->y6r; recode(x6*y6s, "2=1; 4=0")->y6
x7<-rbinom(pop.size[7],1,prop[7]); y7s<-rbinom(pop.size[7],1,prop[7])
recode(x7, "0=2")->x7r; recode(y7s, "0=2")->y7r; recode(x7*y7s, "2=1; 4=0")->y7
x8<-rbinom(pop.size[8],1,prop[8]); y8s<-rbinom(pop.size[8],1,prop[8])
recode(x8, "0=2")->x8r; recode(y8s, "0=2")->y8r; recode(x8*y8s, "2=1; 4=0")->y8
x9<-rbinom(pop.size[9],1,prop[9]); y9s<-rbinom(pop.size[9],1,prop[9])
recode(x9, "0=2")->x9r; recode(y9s, "0=2")->y9r; recode(x9*y9s, "2=1; 4=0")->y9
x10<-rbinom(pop.size[10],1,prop[10]); y10s<-rbinom(pop.size[10],1,prop[10])
recode(x10, "0=2")->x10r; recode(y10s, "0=2")->y10r; recode(x10*y10s, "2=1; 4=0")->y10
x11<-rbinom(pop.size[11],1,prop[11]); y11s<-rbinom(pop.size[11],1,prop[11])
recode(x11, "0=2")->x11r; recode(y11s, "0=2")->y11r; recode(x11*y11s, "2=1; 4=0")->y11
x12<-rbinom(pop.size[12],1,prop[12]); y12s<-rbinom(pop.size[12],1,prop[12])
recode(x12, "0=2")->x12r; recode(y12s, "0=2")->y12r; recode(x12*y12s, "2=1; 4=0")->y1
x13<-rbinom(pop.size[13],1,prop[13]); y13s<-rbinom(pop.size[13],1,prop[13])
recode(x13, "0=2")->x13r; recode(y13s, "0=2")->y13r; recode(x13*y13s, "2=1; 4=0")->y13
x14<-rbinom(pop.size[14],1,prop[14]); y14s<-rbinom(pop.size[14],1,prop[14])
recode(x14, "0=2")->x14r; recode(y14s, "0=2")->y14r; recode(x14*y14s, "2=1; 4=0")->y14
x15<-rbinom(pop.size[15],1,prop[15]); y15s<-rbinom(pop.size[15],1,prop[15])
recode(x15, "0=2")->x15r; recode(y15s, "0=2")->y15r; recode(x15*y15s, "2=1; 4=0")->y15
x16<-rbinom(pop.size[16],1,prop[16]); y16s<-rbinom(pop.size[16],1,prop[16])
recode(x16, "0=2")->x16r; recode(y16s, "0=2")->y16r; recode(x16*y16s, "2=1; 4=0")->y16
x17<-rbinom(pop.size[17],1,prop[17]); y17s<-rbinom(pop.size[17],1,prop[17])
recode(x17, "0=2")->x17r; recode(y17s, "0=2")->y17r; recode(x17*y17s, "2=1; 4=0")->y17
x18<-rbinom(pop.size[18],1,prop[18]); y18s<-rbinom(pop.size[18],1,prop[18])
recode(x18, "0=2")->x18r; recode(y18s, "0=2")->y18r; recode(x18*y18s, "2=1; 4=0")->y18
x19<-rbinom(pop.size[19],1,prop[19]); y19s<-rbinom(pop.size[19],1,prop[19])
recode(x19, "0=2")->x19r; recode(y19s, "0=2")->y19r; recode(x19*y19s, "2=1; 4=0")->y19
x20<-rbinom(pop.size[20],1,prop[20]); y20s<-rbinom(pop.size[20],1,prop[20])
recode(x20, "0=2")->x20r; recode(y20s, "0=2")->y20r; recode(x20*y20s, "2=1; 4=0")->y20
x21<-rbinom(pop.size[21],1,prop[21]); y21s<-rbinom(pop.size[21],1,prop[21])
recode(x21, "0=2")->x21r; recode(y21s, "0=2")->y21r; recode(x21*y21s, "2=1; 4=0")->y21
```

```
x22<-rbinom(pop.size[22],1,prop[22]); y22s<-rbinom(pop.size[22],1,prop[22])
recode(x22, "0=2")->x22r; recode(y22s, "0=2")->y22r; recode(x22*y22s, "2=1; 4=0")->y22
x23<-rbinom(pop.size[23],1,prop[23]); y23s<-rbinom(pop.size[23],1,prop[23])
recode(x23, "0=2")->x23r; recode(y23s, "0=2")->y23r; recode(x23*y23s, "2=1; 4=0")->y23
x24<-rbinom(pop.size[24],1,prop[24]); y24s<-rbinom(pop.size[24],1,prop[24])
recode(x24, "0=2")->x24r; recode(y24s, "0=2")->y24r; recode(x24*y24s, "2=1; 4=0")->y24
x25<-rbinom(pop.size[25],1,prop[25]); y25s<-rbinom(pop.size[25],1,prop[25])
recode(x25, "0=2")->x25r; recode(y25s, "0=2")->y25r; recode(x25*y25s, "2=1; 4=0")->y25
x26<-rbinom(pop.size[26],1,prop[26]); y26s<-rbinom(pop.size[26],1,prop[26])
recode(x26, "0=2")->x26r; recode(y26s, "0=2")->y26r; recode(x26*y26s, "2=1; 4=0")->y26
x27<-rbinom(pop.size[27],1,prop[27]); y27s<-rbinom(pop.size[27],1,prop[27])
recode(x27, "0=2")->x27r; recode(y27s, "0=2")->y27r; recode(x27*y27s, "2=1; 4=0")->y27
x28<-rbinom(pop.size[28],1,prop[28]); y28s<-rbinom(pop.size[28],1,prop[28])
recode(x28, "0=2")->x28r; recode(y28s, "0=2")->y28r; recode(x28*y28s, "2=1; 4=0")->y28
x29<-rbinom(pop.size[29],1,prop[29]); y29s<-rbinom(pop.size[29],1,prop[29])
recode(x29, "0=2")->x29r; recode(y29s, "0=2")->y29r; recode(x29*y29s, "2=1; 4=0")->y29
x30<-rbinom(pop.size[30],1,prop[30]); y30s<-rbinom(pop.size[30],1,prop[30])
recode(x30, "0=2")->x30r; recode(y30s, "0=2")->y30r; recode(x30*y30s, "2=1; 4=0")->y30

x<-c(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,x18,x19,x20,x21,x22,x23,
x24,x25,x26,x27,x28,x29,x30)

y<-c(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10,y11,y12,y13,y14,y15,y16,y17,y18,y19,y20,y21,y22,y23,
y24,y25,y26,y27,y28,y29,y30)

id<-seq(1:sum(pop.size))
regioncode<-rep(1:30,pop.size)
pop<-cbind(id,y,x,regioncode)

#Inicializar a cero todos los objetos que serán usados.
tamao.muestra <- array(c(0,0),dim=c(1,30,nsim))
true.mean<-array(c(0,3000),dim=c(1,30,nsim))
true.mean.x<-array(c(0,3000),dim=c(1,30,nsim))

# Estimador directo
direct.estimate<-array(c(0,0),dim=c(1,30,nsim))
direct.estimate.b<-array(c(0,0),dim=c(1,30,nsim))
MSE.direct<-array(c(0,3000),dim=c(1,30,nsim))
varhat.direct.con <- array(c(0,3000),dim=c(1,30,nsim))
varhat.direct.Ucon <- array(c(0,3000),dim=c(1,30,nsim))
bias.direct<-array(c(0,3000),dim=c(1,30,nsim))

# ratio
ratio.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.ratio<-array(c(0,3000),dim=c(1,30,nsim))
varhat.ratio <- array(c(0,3000),dim=c(1,30,nsim))
bias.ratio<-array(c(0,3000),dim=c(1,30,nsim))

# ratio1
ratio1.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.ratio1<-array(c(0,3000),dim=c(1,30,nsim))
varhat.ratio1 <- array(c(0,3000),dim=c(1,30,nsim))
bias.ratio1<-array(c(0,3000),dim=c(1,30,nsim))

# ratio2
ratio2.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.ratio2<-array(c(0,3000),dim=c(1,30,nsim))
varhat.ratio2 <- array(c(0,3000),dim=c(1,30,nsim))
bias.ratio2<-array(c(0,3000),dim=c(1,30,nsim))

# regresion
```

```
regresion.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.regresion<-array(c(0,3000),dim=c(1,30,nsim))
varhat.regresion <- array(c(0,3000),dim=c(1,30,nsim))
bias.regresion<-array(c(0,3000),dim=c(1,30,nsim))

# regresion1
regresion1.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.regresion1<-array(c(0,3000),dim=c(1,30,nsim))
varhat.regresion1 <- array(c(0,3000),dim=c(1,30,nsim))
bias.regresion1<-array(c(0,3000),dim=c(1,30,nsim))

# regresion2
regresion2.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.regresion2<-array(c(0,3000),dim=c(1,30,nsim))
varhat.regresion2 <- array(c(0,3000),dim=c(1,30,nsim))
bias.regresion2<-array(c(0,3000),dim=c(1,30,nsim))

# diferencia
diferencia.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.diferencia<-array(c(0,3000),dim=c(1,30,nsim))
varhat.diferencia <- array(c(0,3000),dim=c(1,30,nsim))
bias.diferencia<-array(c(0,3000),dim=c(1,30,nsim))

# diferencia1
diferencia1.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.diferencia1<-array(c(0,3000),dim=c(1,30,nsim))
varhat.diferencia1 <- array(c(0,3000),dim=c(1,30,nsim))
bias.diferencia1<-array(c(0,3000),dim=c(1,30,nsim))

# diferencia2
diferencia2.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.diferencia2<-array(c(0,3000),dim=c(1,30,nsim))
varhat.diferencia2 <- array(c(0,3000),dim=c(1,30,nsim))
bias.diferencia2<-array(c(0,3000),dim=c(1,30,nsim))

# cra1 combinado de razón con alfa (directo con dominio)
cra1.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.cra1<-array(c(0,3000),dim=c(1,30,nsim))
varhat.cra1 <- array(c(0,3000),dim=c(1,30,nsim))
bias.cra1<-array(c(0,3000),dim=c(1,30,nsim))

# cra2 combinado de razón con alfa (directo con población)
cra2.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.cra2<-array(c(0,3000),dim=c(1,30,nsim))
varhat.cra2 <- array(c(0,3000),dim=c(1,30,nsim))
bias.cra2<-array(c(0,3000),dim=c(1,30,nsim))

# cra3 combinado de razón con alfa (directo con sintético)
cra3.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.cra3<-array(c(0,3000),dim=c(1,30,nsim))
varhat.cra3 <- array(c(0,3000),dim=c(1,30,nsim))
bias.cra3<-array(c(0,3000),dim=c(1,30,nsim))

# cra4 combinado de razón con alfa (dominio con sintético)
cra4.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.cra4<-array(c(0,3000),dim=c(1,30,nsim))
varhat.cra4 <- array(c(0,3000),dim=c(1,30,nsim))
bias.cra4<-array(c(0,3000),dim=c(1,30,nsim))

# cra5 combinado de razón con alfa (poblaci
on con sintético)
cra5.estimate<-array(c(0,0),dim=c(1,30,nsim))
```

```
MSE.cra5<-array(c(0,3000),dim=c(1,30,nsim))
varhat.cra5 <- array(c(0,3000),dim=c(1,30,nsim))
bias.cra5<-array(c(0,3000),dim=c(1,30,nsim))

# crega1 combinado de razón con alfa (directo con dominio)
crega1.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.crega1<-array(c(0,3000),dim=c(1,30,nsim))
varhat.crega1 <- array(c(0,3000),dim=c(1,30,nsim))
bias.crega1<-array(c(0,3000),dim=c(1,30,nsim))

# crega2 combinado de razón con alfa (directo con población)
crega2.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.crega2<-array(c(0,3000),dim=c(1,30,nsim))
varhat.crega2 <- array(c(0,3000),dim=c(1,30,nsim))
bias.crega2<-array(c(0,3000),dim=c(1,30,nsim))

# crega3 combinado de razón con alfa (directo con sintético)
crega3.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.crega3<-array(c(0,3000),dim=c(1,30,nsim))
varhat.crega3 <- array(c(0,3000),dim=c(1,30,nsim))
bias.crega3<-array(c(0,3000),dim=c(1,30,nsim))

# crega4 combinado de razón con alfa (dominio con sint
etico)
crega4.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.crega4<-array(c(0,3000),dim=c(1,30,nsim))
varhat.crega4 <- array(c(0,3000),dim=c(1,30,nsim))
bias.crega4<-array(c(0,3000),dim=c(1,30,nsim))

# crega5 combinado de razón con alfa (población con sintético)
crega5.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.crega5<-array(c(0,3000),dim=c(1,30,nsim))
varhat.crega5 <- array(c(0,3000),dim=c(1,30,nsim))
bias.crega5<-array(c(0,3000),dim=c(1,30,nsim))

# logistic
TLgreg.estimate<-array(c(0,0),dim=c(1,30,nsim))
logistic.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.logistic<-array(c(0,3000),dim=c(1,30,nsim))
varhat.logistic <- array(c(0,3000),dim=c(1,30,nsim))
bias.logistic<-array(c(0,3000),dim=c(1,30,nsim))

# logistic1
TLgreg.estimate<-array(c(0,0),dim=c(1,30,nsim))
logistic1.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.logistic1<-array(c(0,3000),dim=c(1,30,nsim))
varhat.logistic1 <- array(c(0,3000),dim=c(1,30,nsim))
bias.logistic1<-array(c(0,3000),dim=c(1,30,nsim))

# logistic2
TLgreg.estimate<-array(c(0,0),dim=c(1,30,nsim))
logistic2.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.logistic2<-array(c(0,3000),dim=c(1,30,nsim))
varhat.logistic2 <- array(c(0,3000),dim=c(1,30,nsim))
bias.logistic2<-array(c(0,3000),dim=c(1,30,nsim))

# logistic3
TLgreg.estimate<-array(c(0,0),dim=c(1,30,nsim))
logistic3.estimate<-array(c(0,0),dim=c(1,30,nsim))
MSE.logistic3<-array(c(0,3000),dim=c(1,30,nsim))
varhat.logistic3 <- array(c(0,3000),dim=c(1,30,nsim))
bias.logistic3<-array(c(0,3000),dim=c(1,30,nsim))
```

```
#Generar muestras distintas en cada repetición
sed1<-c(30:32)
sed2<-c(34:38)
sed3<-c(44)
sed4<-c(64:73)
sed5<-c(80:86)
sed6<-c(101:120)
sed7<-c(201:217)
sed8<-c(220:256)
sed<-c(sed1,sed2,sed3,sed4,sed5,sed6,sed7,sed8)

#Designemos por h el índice de la simulación
n.size<-900
#for(h in 1:nsim){
h<-1
while (h <nsim+1 ) {

#set.seed(sed[h])

s<-sample(id,n.size)
y<-pop[s,2]
x<-pop[s,3]
regioncode<-pop[s,4]
datanew<-cbind(s,y,x,regioncode)
datanew<-datanew[order(datanew[,4]),1:4] #Muestra con la unidad, la y, la x y la región o área
samp.size<-c(table(datanew[,4]))
#Condiciones
prod(aggregate(datanew[,3],list(d2=datanew[,4]),mean)[,2])->propx
prod(aggregate(datanew[,2],list(d2=datanew[,4]),mean)[,2])->propy
prod(1-aggregate(datanew[,3],list(d2=datanew[,4]),mean)[,2])->proqx
prod(1-aggregate(datanew[,2],list(d2=datanew[,4]),mean)[,2])->proqy
# para que ninguna de las proporciones estimadas en las areas sea cero, "propx*propy distinto de
cero"
# para que ninguna de las proporciones estimadas en las areas sea uno, "proqx*proqy distinto de
cero"
if (propx*propy*proqx*proqy==0) {
next
} else
{

#Las verdaderas medias
true.mean[,,h]<-aggregate(pop[,2],list(d2=pop[,4]),mean)[,2]
true.mean.x[,,h]<-aggregate(pop[,3],list(d2=pop[,4]),mean)[,2]
tamao.muestra[,,h]<- as.vector(samp.size)

# Estimadores Propuestos
# 1.Estimador directo
direct.estimate[,,h]<- aggregate(datanew[,2],list(d2=datanew[,4]),mean)[,2] # Ec. (4.9)

# Varianza del estimador directo.
vhat.direct.con <- rep(0,30)
vhat.direct.Ucon <- rep(0,30)

for(i in 1:30){
nd <- samp.size[i]
Nd <- pop.size[i]
ad <- rep(0,nd)
pad <- mean(datanew[,2][datanew[,4]==i])
ad <- datanew[,2][datanew[,4]==i]
vhat.direct.con[i] <- (1/nd)*(1-n.size/sum(pop.size))*(n.size/(n.size-1))*pad*(1-pad)# Ec. (??)
vhat.direct.Ucon[i] <- (1/nd)*(1-n.size/sum(pop.size))*pad*(1-pad)# Ec. (??)
```

```
}

varhat.direct.con[,,h] <- vhat.direct.con
varhat.direct.Ucon[,,h] <- vhat.direct.Ucon
bias.direct[,,h]<-(direct.estimate[,,h]-true.mean[,,h])
MSE.direct[,,h]<-(direct.estimate[,,h]-true.mean[,,h])2̂

#Informacion auxiliar del atributo B
direct.estimate.b[,,h]<- aggregate(datanew[,3],list(d2=datanew[,4]),mean)[,2] #Medias del atributo
B en la muestra

#2.Estimador de razón (ratio)
ratio.estimate[,,h]<- direct.estimate[,,h]/direct.estimate.b[,,h]*true.mean.x[,,h]# Ec. (4.13)

# Varianza del estimador de razón
vhat.ratio <- rep(0,30)

for(i in 1:30){
nd <- samp.size[i]
Nd <- pop.size[i]
ad <- rep(0,nd)
pad <- mean(datanew[,2][datanew[,4]==i])
pbd <- mean(datanew[,3][datanew[,4]==i])
Pb <- mean(pop[,3][pop[,4]==i])
ad <- datanew[,2][datanew[,4]==i]
rd <- pad/pbd
phid <- cor(datanew[,2][datanew[,4]==i],datanew[,3][datanew[,4]==i])
vhat.ratio[i] <- (Pb/pbd)2̂*(1/nd)*(1-n.size/sum(pop.size))*(n.size/(n.size-1))*(pad*(1-pad)
+ (rd2̂)*pbd*(1-pbd)-2*rd*phid*sqrt(pad*(1-pad)*pbd*(1-pbd)))# Ec. (4.16)
}

varhat.ratio[,,h] <- vhat.ratio
bias.ratio[,,h]<-(ratio.estimate[,,h]-true.mean[,,h])
MSE.ratio[,,h]<-(ratio.estimate[,,h]-true.mean[,,h])2̂

#3.Estimador de razón1
ratio1.estimate[,,h]<- direct.estimate[,,h]/direct.estimate.b[,,h]*mean(pop[,3])# Ec. (4.18)

# Varianza del estimador de razón1
vhat.ratio1 <- rep(0,30)
vhat.ratio1 <- vhat.ratio
varhat.ratio1[,,h] <- vhat.ratio1
bias.ratio1[,,h]<-(ratio1.estimate[,,h]-true.mean[,,h])
MSE.ratio1[,,h]<-(ratio1.estimate[,,h]-true.mean[,,h])2̂

#4.Estimador de razón sintético
ratio2.estimate[,,h]<- mean(datanew[,2])/mean(datanew[,3])*true.mean.x[,,h]# Ec. (4.26)

# Varianza del estimador de razón sintético
vhat.ratio2 <- rep(0,30)
pa <- mean(datanew[,2])
pb <- mean(datanew[,3])
Pb <- mean(pop[,3][pop[,4]==i])
pbd <- mean(datanew[,3][datanew[,4]==i])
phi <- cor(datanew[,2],datanew[,3])
r <- pad/pbd
const <- (pbd/Pb)2̂*(1/(n.size-1))*(1-n.size/sum(pop.size))*(pa*(1-pa) + (r2̂)*pb*(1-pb)
- 2*r*phi*sqrt(pa*(1-pa)*pb*(1-pb)))# Ec. (4.33)

vhat.ratio2 <- rep(const, 30)
varhat.ratio2[,,h] <- vhat.ratio2
```

```
bias.ratio2[,,h]<-(ratio2.estimate[,,h]-true.mean[,,h])
MSE.ratio2[,,h]<-(ratio2.estimate[,,h]-true.mean[,,h])2̂

#5.Estimador de regresión con información de dominio
#calcular el coeficiente de regresión primero y las demás cantidades
vhat.regresion <- rep(0,30)
for(i in 1:30){
nd <- samp.size[i]
Nd <- pop.size[i]
ad <- rep(0,nd)
pad <- mean(datanew[,2][datanew[,4]==i])
pbd <- mean(datanew[,3][datanew[,4]==i])
bd <- phid*sqrt(pad*(1-pad)/(pbd*(1-pbd)))
phid <- cor(datanew[,2][datanew[,4]==i],datanew[,3][datanew[,4]==i])
vhat.regresion[i] <- (1/nd)*(1-n.size/sum(pop.size))*pad*(1-pad)*(1-phid2̂)# Ec. (4.93)
}

regresion.estimate[,,h]<- direct.estimate[,,h] + bd*(true.mean.x[,,h] - direct.estimate.b[,,h])#
Ec. (4.89)

# Varianza del estimador de regresión
varhat.regresion[,,h] <- vhat.regresion
bias.regresion[,,h]<-(regresion.estimate[,,h]-true.mean[,,h])
MSE.regresion[,,h]<-(regresion.estimate[,,h]-true.mean[,,h])2̂

#6.Estimador de regresión1 con información poblacional
regresion1.estimate[,,h]<- direct.estimate[,,h] + bd*(mean(pop[,3]) - direct.estimate.b[,,h])#
Ec. (4.99)

# Varianza del estimador de regresión1
vhat.regresion1 <- rep(0,30)
vhat.regresion1 <- vhat.regresion
varhat.regresion1[,,h] <- vhat.regresion1
bias.regresion1[,,h]<-(regresion1.estimate[,,h]-true.mean[,,h])
MSE.regresion1[,,h]<-(regresion1.estimate[,,h]-true.mean[,,h])2̂

#7.Estimador de regresión2 sintético
pa <- mean(datanew[,2])
pb <- mean(datanew[,3])
phi <- cor(datanew[,2],datanew[,3])
b <- phi*sqrt(pa*(1-pa)/(pb*(1-pb)))
regresion2.estimate[,,h]<- pa + b*(true.mean.x[,,h] - pb)# Ec. (4.108)

# Varianza del estimador de regresión2
vhat.regresion2 <- rep(0,30)
const2 <- (1/(n.size-1))*(1-n.size/sum(pop.size))*pa*(1-pa)*(1-phi2̂)# Ec. (4.110)
vhat.regresion2 <- rep(const2, 30)
varhat.regresion2[,,h] <- vhat.regresion2
bias.regresion2[,,h]<-(regresion2.estimate[,,h]-true.mean[,,h])
MSE.regresion2[,,h]<-(regresion2.estimate[,,h]-true.mean[,,h])2̂

# 8.Estimador de diferencia con información de dominio
# calcular el coeficiente de diferencia primero y las demás cantidades
vhat.diferencia <- rep(0,30)
for(i in 1:30){
nd <- samp.size[i]
Nd <- pop.size[i]
ad <- rep(0,nd)
pad <- mean(datanew[,2][datanew[,4]==i])
pbd <- mean(datanew[,3][datanew[,4]==i])
phid <- cor(datanew[,2][datanew[,4]==i],datanew[,3][datanew[,4]==i])
vhat.diferencia[i] <- (1/nd)*(1-n.size/sum(pop.size))*(pad*(1-pad) + pbd*(1-pbd)
```

```
-2*phid*sqrt(pad*(1-pad)*pbd*(1-pbd)))# Ec. (4.105)
}
diferencia.estimate[,,h]<- direct.estimate[,,h] + (true.mean.x[,,h] - direct.estimate.b[,,h])#
Ec. (4.101)

# Varianza del estimador de diferencia
varhat.diferencia[,,h] <- vhat.diferencia
bias.diferencia[,,h]<-(diferencia.estimate[,,h]-true.mean[,,h])
MSE.diferencia[,,h]<-(diferencia.estimate[,,h]-true.mean[,,h])2̂


    #9. Estimador de diferencia1 con información poblacional
diferencia1.estimate[,,h]<- direct.estimate[,,h] + (mean(pop[,3]) - direct.estimate.b[,,h])# Ec.
(4.102)


# Varianza del estimador de diferencia1
vhat.diferencia1 <- rep(0,30)
vhat.diferencia1 <- vhat.diferencia
varhat.diferencia1[,,h] <- vhat.diferencia1
bias.diferencia1[,,h]<-(diferencia1.estimate[,,h]-true.mean[,,h])
MSE.diferencia1[,,h]<-(diferencia1.estimate[,,h]-true.mean[,,h])2̂


#10. Estimador de diferencia2 sintético
pa <- mean(datanew[,2])
pb <- mean(datanew[,3])
phi <- cor(datanew[,2],datanew[,3])
diferencia2.estimate[,,h]<- pa + (true.mean.x[,,h] - pb)# Ec. (4.103)


# Varianza del estimador de diferencia2 sintético
vhat.diferencia2 <- rep(0,30)
const2 <- (1/nd)*(1-n.size/sum(pop.size))*(pa*(1-pa) + pb*(1-pb) -2*phi*sqrt(pa*(1-pa)*pb*(1-pb)))#
Ec. (4.107)
vhat.diferencia2 <- rep(const2, 30)
varhat.diferencia2[,,h] <- vhat.diferencia2 #este es el (22)
bias.diferencia2[,,h]<-(diferencia2.estimate[,,h]-true.mean[,,h])
MSE.diferencia2[,,h]<-(diferencia2.estimate[,,h]-true.mean[,,h])2̂


# 11. Estimador combinado de razón con alfa (directo con Razón dominio)
#calcular el coeficiente de cra1
alfa.cra1 <- rep(0,30)
alfa.cra1 <- vhat.ratio/(vhat.direct.Ucon+vhat.ratio)# Ec. (4.52)
vhat.cra1 <- rep(0,30)
vhat.cra1 <- (vhat.ratio*vhat.direct.Ucon)/(vhat.direct.Ucon+vhat.ratio)


cra1.estimate [,,h]<- alfa.cra1*direct.estimate[,,h] + (1-alfa.cra1)*ratio.estimate[,,h] # Ec.
```

```
(4.50)

# Varianza del estimador - cra1
varhat.cra1[,,h] <- vhat.cra1
bias.cra1[,,h]<-(cra1.estimate[,,h]-true.mean[,,h])
MSE.cra1[,,h]<-(cra1.estimate[,,h]-true.mean[,,h])2̂

# 12. Estimador combinado de razón con alfa (directo con razón población)
# calcular el coeficiente de cra2
alfa.cra2 <- rep(0,30)
alfa.cra2 <- vhat.ratio1/(vhat.direct.Ucon+vhat.ratio1)# Ec. (4.56)
vhat.cra2 <- rep(0,30)
vhat.cra2 <- (vhat.ratio1*vhat.direct.Ucon)/(vhat.direct.Ucon+vhat.ratio1)
cra2.estimate [,,h]<- alfa.cra2*direct.estimate[,,h] + (1-alfa.cra2)*ratio1.estimate[,,h] # Ec.
(4.54)

# Varianza del estimador - cra2
varhat.cra2[,,h] <- vhat.cra2
bias.cra2[,,h]<-(cra2.estimate[,,h]-true.mean[,,h])
MSE.cra2[,,h]<-(cra2.estimate[,,h]-true.mean[,,h])2̂

#13. Estimador combinado de razón con alfa (directo con razón sintético)
#calculo el coeficiente de cra3 primero y ya que estamos en el for todo lo demás
alfa.cra3 <- rep(0,30)
alfa.cra3 <- vhat.ratio2/(vhat.direct.Ucon+vhat.ratio2)# Ec. (4.60)
vhat.cra3 <- rep(0,30)
vhat.cra3 <- (vhat.ratio2*vhat.direct.Ucon)/(vhat.direct.Ucon+vhat.ratio2)

cra3.estimate [,,h]<- alfa.cra3*direct.estimate[,,h] + (1-alfa.cra3)*ratio2.estimate[,,h] # Ec.
(4.58)

# Varianza del estimador - cra3
varhat.cra3[,,h] <- vhat.cra3
bias.cra3[,,h]<-(cra3.estimate[,,h]-true.mean[,,h])
MSE.cra3[,,h]<-(cra3.estimate[,,h]-true.mean[,,h])2̂

#14. Estimador combinado de razón con alfa (razón de dominio con razón sintético)
#calcular el coeficiente de cra4
alfa.cra4 <- rep(0,30)
alfa.cra4 <- vhat.ratio2/(vhat.ratio+vhat.ratio2)# Ec. (4.64)
vhat.cra4 <- rep(0,30)
```

```
vhat.cra4 <- (vhat.ratio2*vhat.ratio)/(vhat.ratio+vhat.ratio2)
cra4.estimate [,,h]<- alfa.cra4*ratio.estimate[,,h]+ (1-alfa.cra4)*ratio2.estimate[,,h] # Ec. (4.62)

# Varianza del estimador - cra4
varhat.cra4[,,h] <- vhat.cra4
bias.cra4[,,h]<-(cra4.estimate[,,h]-true.mean[,,h])
MSE.cra4[,,h]<-(cra4.estimate[,,h]-true.mean[,,h])2̂

#15. Estimador combinado de razón con alfa (razón población con razón sintético)
#calcular el coeficiente de cra5
alfa.cra5 <- rep(0,30)
alfa.cra5 <- vhat.ratio2/(vhat.ratio1+vhat.ratio2)# Ec. (4.68)
vhat.cra5 <- rep(0,30)
vhat.cra5 <- (vhat.ratio2*vhat.ratio1)/(vhat.ratio1+vhat.ratio2)
cra5.estimate [,,h]<- alfa.cra5*ratio1.estimate[,,h]+ (1-alfa.cra5)*ratio2.estimate[,,h] # Ec.
(4.66)

# Varianza del estimador - cra5
varhat.cra5[,,h] <- vhat.cra5
bias.cra5[,,h]<-(cra5.estimate[,,h]-true.mean[,,h])
MSE.cra5[,,h]<-(cra5.estimate[,,h]-true.mean[,,h])2̂

#16. Estimador combinado de regresión con alfa (directo con regresión dominio)
#calcular el coeficiente de crega1
alfa.crega1 <- rep(0,30)
alfa.crega1 <- vhat.regresión/(vhat.direct.Ucon+vhat.regresión)# Ec. (4.118)
vhat.crega1 <- rep(0,30)
vhat.crega1 <- (vhat.regresión*vhat.direct.Ucon)/(vhat.direct.Ucon+vhat.regresión)
crega1.estimate [,,h]<- alfa.crega1*direct.estimate[,,h] + (1-alfa.crega1)*regresión.estimate[,,h]
# Ec. (4.117)

# Varianza del estimador - crega1
varhat.crega1[,,h] <- vhat.crega1
bias.crega1[,,h]<-(crega1.estimate[,,h]-true.mean[,,h])
MSE.crega1[,,h]<-(crega1.estimate[,,h]-true.mean[,,h])2̂

#17. Estimador combinado de regresión con alfa (directo con regresión población)
#calcular el coeficiente de crega2
alfa.crega2 <- rep(0,30)
alfa.crega2 <- vhat.regresión1/(vhat.direct.Ucon+vhat.regresión1)# Ec. (4.120)
vhat.crega2 <- rep(0,30)
```

```
vhat.crega2 <- (vhat.regresión1*vhat.direct.Ucon)/(vhat.direct.Ucon+vhat.regresión1)
crega2.estimate [,,h]<- alfa.crega2*direct.estimate[,,h] + (1-alfa.crega2)*regresión1.estimate[,,h]
# Ec. (4.119)

# Varianza del estimador - crega2
varhat.crega2[,,h] <- vhat.crega2
bias.crega2[,,h]<-(crega2.estimate[,,h]-true.mean[,,h])
MSE.crega2[,,h]<-(crega2.estimate[,,h]-true.mean[,,h])2̂

#18. Estimador combinado de regresión con alfa (directo con regresión sintético)
#calcular el coeficiente de crega3
alfa.crega3 <- rep(0,30)
alfa.crega3 <- vhat.regresión2/(vhat.direct.Ucon+vhat.regresión2)# Ec. (4.122)
vhat.crega3 <- rep(0,30)
vhat.crega3 <- (vhat.regresión2*vhat.direct.Ucon)/(vhat.direct.Ucon+vhat.regresión2)
crega3.estimate [,,h]<- alfa.crega3*direct.estimate[,,h] + (1-alfa.crega3)*regresión2.estimate[,,h]
# Ec. (4.121)

# Varianza del estimador - crega3
varhat.crega3[,,h] <- vhat.crega3
bias.crega3[,,h]<-(crega3.estimate[,,h]-true.mean[,,h])
MSE.crega3[,,h]<-(crega3.estimate[,,h]-true.mean[,,h])2̂

#19. Estimador combinado de regresión con alfa (regresión dominio con regresión sintético)
#calcular el coeficiente de crega4
alfa.crega4 <- rep(0,30)
alfa.crega4 <- vhat.regresión2/(vhat.regresión+vhat.regresión2)# Ec. (4.124)
vhat.crega4 <- rep(0,30)
vhat.crega4 <- (vhat.regresión2*vhat.regresión)/(vhat.regresión+vhat.regresión2)
crega4.estimate [,,h]<- alfa.crega4*regresión.estimate[,,h]+ (1-alfa.crega4)*regresión2.estimate[,,h]
# Ec. (4.123)

# Varianza del estimador - crega4
varhat.crega4[,,h] <- vhat.crega4
bias.crega4[,,h]<-(crega4.estimate[,,h]-true.mean[,,h])
MSE.crega4[,,h]<-(crega4.estimate[,,h]-true.mean[,,h])2̂

#20. Estimador combinado de regresión con alfa (regresión población con regresión sintético)
#calcular el coeficiente de crega1
alfa.crega5 <- rep(0,30)
alfa.crega5 <- vhat.regresión2/(vhat.regresión1+vhat.regresión2)# Ec. (4.126)
```

```
vhat.crega5 <- rep(0,30)
vhat.crega5 <- (vhat.regresión2*vhat.regresión1)/(vhat.regresión1+vhat.regresión2)
crega5.estimate [,,h]<- alfa.crega5*regresión1.estimate[,,h]+ (1-alfa.crega5)*regresión2.estimate[,,h]
# Ec. (4.125)

# Varianza del estimador - crega5
varhat.crega5[,,h] <- vhat.crega5
bias.crega5[,,h]<-(crega5.estimate[,,h]-true.mean[,,h])
MSE.crega5[,,h]<-(crega5.estimate[,,h]-true.mean[,,h])2̂

#21. Estimador logistic con información poblacional completa para estimar β
reg<-lm(pop[s,2]∼-1+pop[s,3]) #regresión con la muestra completa por el origen
Betahat<-reg$coef

# Varianza del estimador - logistic
logistic <- rep(0,30)
vhat.logistic <- rep(0,30)
for(i in 1:30){
nd <- samp.size[i]
ysd <- datanew[,2][datanew[,4]==i]
xsd <- datanew[,3][datanew[,4]==i]
xUd <- pop[,3][datanew[,4]==i]
etahat.Ud <- t(xUd)*Betahat
yhat.Ud <- exp(etahat.Ud)/(1+exp(etahat.Ud))
pud <- mean(yhat.Ud)
etahat.sd <- t(xsd)*Betahat
yhat.sd <- exp(etahat.sd)/(1+exp(etahat.sd))
psd <- mean(yhat.sd)
logistic[i] <- sum(yhat.Ud) + (sum(pop.size)/n.size)*(sum(ysd)- sum(yhat.sd))
logistic[i] <- (1/nd)*(n.size/sum(pop.size))*logistic[i]
vhat.logistic[i] <- (1/nd)*(1-n.size/sum(pop.size))*(n.size/(n.size-1))*pad*(1-pad)
}
logistic.estimate[,,h]<- logistic
varhat.logistic[,,h] <- vhat.logistic
bias.logistic[,,h]<-(logistic.estimate[,,h]-true.mean[,,h])
MSE.logistic[,,h]<-(logistic.estimate[,,h]-true.mean[,,h])2̂

#22. Estimador logistic1
#reg<-lm(pop[s,2]∼-1+pop[s,3]) #regresión con la muestra completa por el origen

# Varianza del estimador - logistic1
```

```
reg.d<-rep(0,30)
logistic1 <- rep(0,30)
vhat.logistic1 <- rep(0,30)

for(i in 1:30){
nd <- samp.size[i]
ysd <- datanew[,2][datanew[,4]==i]
xsd <- datanew[,3][datanew[,4]==i]
reg.d <- lm(ysd∼-1+xsd) #regresión con la muestra en el dominio por el origen
Betahat.d <- reg.d$coef
xUd <- pop[,3][datanew[,4]==i]
etahat.Ud <- t(xUd)*Betahat.d
yhat.Ud <- exp(etahat.Ud)/(1+exp(etahat.Ud))
pud <- mean(yhat.Ud)
etahat.sd <- t(xsd)*Betahat.d
yhat.sd <- exp(etahat.sd)/(1+exp(etahat.sd))
psd <- mean(yhat.sd)
logistic1[i] <- sum(yhat.Ud) + (sum(pop.size)/n.size)*(sum(ysd)- sum(yhat.sd))
logistic1[i] <- (1/nd)*(n.size/sum(pop.size))*logistic1[i]
vhat.logistic1[i] <- (1/nd)*(1-n.size/sum(pop.size))*(n.size/(n.size-1))*pad*(1-pad)
}
logistic1.estimate[,,h]<- logistic1
varhat.logistic1[,,h] <- vhat.logistic1
bias.logistic1[,,h]<-(logistic1.estimate[,,h]-true.mean[,,h])
MSE.logistic1[,,h]<-(logistic1.estimate[,,h]-true.mean[,,h])^2

#23. Estimador logistic2
#reg<-lm(pop[s,2]∼-1+pop[s,3]) #regresión con la muestra completa por el origen

# Varianza del estimador - logistic2
reg.d<-rep(0,30)
logistic2 <- rep(0,30)
vhat.logistic2 <- rep(0,30)

for(i in 1:30){
nd <- samp.size[i]
ysd <- datanew[,2][datanew[,4]==i]
xsd <- datanew[,3][datanew[,4]==i]
reg.d <- glm(ysd∼-1+xsd, family=binomial("logit")) #regresión con la muestra en el dominio por
el origen
Betahat.d <- reg.d$coef
```

```
xUd <- pop[,3][datanew[,4]==i]
etahat.Ud <- t(xUd)*Betahat.d
yhat.Ud <- exp(etahat.Ud)/(1+exp(etahat.Ud))
pud <- mean(yhat.Ud)
etahat.sd <- t(xsd)*Betahat.d
yhat.sd <- exp(etahat.sd)/(1+exp(etahat.sd))
psd <- mean(yhat.sd)
logistic2[i] <- sum(yhat.Ud) + (sum(pop.size)/n.size)*(sum(ysd)- sum(yhat.sd))
logistic2[i] <- (1/nd)*(n.size/sum(pop.size))*logistic2[i]
vhat.logistic2[i] <- (1/nd)*(1-n.size/sum(pop.size))*(n.size/(n.size-1))*pad*(1-pad)
}
logistic2.estimate[,,h]<- logistic2
varhat.logistic2[,,h] <- vhat.logistic2
bias.logistic2[,,h]<-(logistic2.estimate[,,h]-true.mean[,,h])
MSE.logistic2[,,h]<-(logistic2.estimate[,,h]-true.mean[,,h])2̂

#24. Estimador logistic3
reg<-glm(pop[s,2]∼-1+pop[s,3],family=binomial("logit")) #regresión con la muestra completa por
el origen
Betahat<-reg$coef

# Varianza del estimador - logistic3
logistic3 <- rep(0,30)
vhat.logistic3 <- rep(0,30)

for(i in 1:30){
nd <- samp.size[i]
ysd <- datanew[,2][datanew[,4]==i]
xsd <- datanew[,3][datanew[,4]==i]
xUd <- pop[,3][datanew[,4]==i]
etahat.Ud <- t(xUd)*Betahat
yhat.Ud <- exp(etahat.Ud)/(1+exp(etahat.Ud))
pud <- mean(yhat.Ud)
etahat.sd <- t(xsd)*Betahat
yhat.sd <- exp(etahat.sd)/(1+exp(etahat.sd))
psd <- mean(yhat.sd)
logistic3[i] <- sum(yhat.Ud) + (sum(pop.size)/n.size)*(sum(ysd)- sum(yhat.sd))
logistic3[i] <- (1/nd)*(n.size/sum(pop.size))*logistic3[i]
vhat.logistic3[i] <- (1/nd)*(1-n.size/sum(pop.size))*(n.size/(n.size-1))*pad*(1-pad)
}
logistic3.estimate[,,h]<- logistic3
```

```
varhat.logistic3[,,h] <- vhat.logistic3
bias.logistic3[,,h]<-(logistic3.estimate[,,h]-true.mean[,,h])
MSE.logistic3[,,h]<-(logistic3.estimate[,,h]-true.mean[,,h])2̂

#Para ver en pantalla el número de simulaciones
print(h)
h<- h+1
} #Fin del else de las condiciones
} #Fin del for de las simulaciones en el índice h

#Cálculos
m.tamao<- apply(tamao.muestra[,,1:nsim],1,mean)#Tamaño muestral medio
# Estimador directo
m.direct<-apply(direct.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.direct<-apply(bias.direct[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.direct<-sqrt(apply(MSE.direct[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.direct <- sqrt(apply(varhat.direct.con[,,1:nsim],1,mean))
rmean.MSE.direct.U <- sqrt(apply(varhat.direct.Ucon[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

# ratio
m.ratio<-apply(ratio.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.ratio<-apply(bias.ratio[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.ratio<-sqrt(apply(MSE.ratio[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.ratio <- sqrt(apply(varhat.ratio.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#ratio1
m.ratio1<-apply(ratio1.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.ratio1<-apply(bias.ratio1[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.ratio1<-sqrt(apply(MSE.ratio1[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.ratio1 <- sqrt(apply(varhat.ratio1.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#ratio2
m.ratio2<-apply(ratio2.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.ratio2<-apply(bias.ratio2[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.ratio2<-sqrt(apply(MSE.ratio2[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
```

```
las simulaciones(empíricos)
rmean.MSE.ratio2 <- sqrt(apply(varhat.ratio2.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#regresion
m.regresion<-apply(regresion.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.regresion<-apply(bias.regresion[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.regresion<-sqrt(apply(MSE.regresion[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del
MSE sobre las simulaciones(empíricos)
rmean.MSE.regresion <- sqrt(apply(varhat.regresion.con[,,1:nsim],1,mean))# Raź cuadrada de los
promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#regresion1
m.regresion1<-apply(regresion1.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.regresion1<-apply(bias.regresion1[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.regresion1<-sqrt(apply(MSE.regresion1[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del
MSE sobre las simulaciones(empíricos)
rmean.MSE.regresion1 <- sqrt(apply(varhat.regresion1.con[,,1:nsim],1,mean))# Raź cuadrada de los
promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#regresion2
m.regresion2<-apply(regresion2.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.regresion2<-apply(bias.regresion2[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.regresion2<-sqrt(apply(MSE.regresion2[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del
MSE sobre las simulaciones(empíricos)
rmean.MSE.regresion2 <- sqrt(apply(varhat.regresion2.con[,,1:nsim],1,mean))# Raź cuadrada de los
promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#diferencia
m.diferencia<-apply(diferencia.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.diferencia<-apply(bias.diferencia[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.diferencia<-sqrt(apply(MSE.diferencia[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del
MSE sobre las simulaciones(empíricos)
rmean.MSE.diferencia <- sqrt(apply(varhat.diferencia.con[,,1:nsim],1,mean))# Raź cuadrada de los
promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#diferencia1
m.diferencia1<-apply(diferencia1.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.diferencia1<-apply(bias.diferencia1[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.diferencia1<-sqrt(apply(MSE.diferencia1[,,1:nsim],1,mean)) # Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos)
```

```
rmean.MSE.diferencia1 <- sqrt(apply(varhat.diferencia1.con[,,1:nsim],1,mean))# Raź cuadrada de
los promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#diferencia2
m.diferencia2<-apply(diferencia2.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.diferencia2<-apply(bias.diferencia2[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.diferencia2<-sqrt(apply(MSE.diferencia2[,,1:nsim],1,mean)) # Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos)
rmean.MSE.diferencia2 <- sqrt(apply(varhat.diferencia2.con[,,1:nsim],1,mean))# Raź cuadrada de
los promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#cra1
m.cra1<-apply(cra1.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.cra1<-apply(bias.cra1[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.cra1<-sqrt(apply(MSE.cra1[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.cra1 <- sqrt(apply(varhat.cra1.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#cra2
m.cra2<-apply(cra2.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.cra2<-apply(bias.cra2[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.cra2<-sqrt(apply(MSE.cra2[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.cra2 <- sqrt(apply(varhat.cra2.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#cra3
m.cra3<-apply(cra3.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.cra3<-apply(bias.cra3[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.cra3<-sqrt(apply(MSE.cra3[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.cra3 <- sqrt(apply(varhat.cra3.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#cra4
m.cra4<-apply(cra4.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.cra4<-apply(bias.cra4[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.cra4<-sqrt(apply(MSE.cra4[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.cra4 <- sqrt(apply(varhat.cra4.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
```

```
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#cra5
m.cra5<-apply(cra5.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.cra5<-apply(bias.cra5[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.cra5<-sqrt(apply(MSE.cra5[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.cra5 <- sqrt(apply(varhat.cra5.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#crega1
m.crega1<-apply(crega1.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.crega1<-apply(bias.crega1[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.crega1<-sqrt(apply(MSE.crega1[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.crega1 <- sqrt(apply(varhat.crega1.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#crega2
m.crega2<-apply(crega2.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.crega2<-apply(bias.crega2[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.crega2<-sqrt(apply(MSE.crega2[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.crega2 <- sqrt(apply(varhat.crega2.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#crega3
m.crega3<-apply(crega3.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.crega3<-apply(bias.crega3[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.crega3<-sqrt(apply(MSE.crega3[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.crega3 <- sqrt(apply(varhat.crega3.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#crega4
m.crega4<-apply(crega4.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.crega4<-apply(bias.crega4[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.crega4<-sqrt(apply(MSE.crega4[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.crega4 <- sqrt(apply(varhat.crega4.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.
```

```
#crega5
m.crega5<-apply(crega5.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.crega5<-apply(bias.crega5[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.crega5<-sqrt(apply(MSE.crega5[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE sobre
las simulaciones(empíricos)
rmean.MSE.crega5 <- sqrt(apply(varhat.crega5.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#logistic
m.logistic<-apply(logistic.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.logistic<-apply(bias.logistic[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.logistic<-sqrt(apply(MSE.logistic[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del MSE
sobre las simulaciones(empíricos)
rmean.MSE.logistic <- sqrt(apply(varhat.logistic.con[,,1:nsim],1,mean))# Raź cuadrada de los promedios
del MSE sobre las simulaciones(empíricos)
# usando estimadores de la varianza.

#logistic1
m.logistic1<-apply(logistic1.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.logistic1<-apply(bias.logistic1[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.logistic1<-sqrt(apply(MSE.logistic1[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del
MSE sobre las simulaciones(empíricos)
rmean.MSE.logistic1 <- sqrt(apply(varhat.logistic1.con[,,1:nsim],1,mean))# Raź cuadrada de los
promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#logistic2
m.logistic2<-apply(logistic2.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.logistic2<-apply(bias.logistic2[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.logistic2<-sqrt(apply(MSE.logistic2[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del
MSE sobre las simulaciones(empíricos)
rmean.MSE.logistic2 <- sqrt(apply(varhat.logistic2.con[,,1:nsim],1,mean))# Raź cuadrada de los
promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.

#logistic3
m.logistic3<-apply(logistic3.estimate[,,1:nsim],1,mean)# Proporciones medias sobre las simulaciones
b.logistic3<-apply(bias.logistic3[,,1:nsim],1,mean)# Sesgo medio sobre las simulaciones
mse.logistic3<-sqrt(apply(MSE.logistic3[,,1:nsim],1,mean)) # Raź cuadrada de los promedios del
MSE sobre las simulaciones(empíricos)
rmean.MSE.logistic3 <- sqrt(apply(varhat.logistic3.con[,,1:nsim],1,mean))# Raź cuadrada de los
promedios del MSE sobre las simulaciones(empíricos) usando estimadores de la varianza.
```

```
# Generación de tablas.
area.ID <- c(1:30)
corre<- rep(0,30)
corre<- c(cor(x1,y1),cor(x2,y2),cor(x3,y3),cor(x4,y4),cor(x5,y5),cor(x6,y6),cor(x7,y7),
cor(x8,y8),cor(x9,y9),cor(x10,y10), cor(x11,y11), cor(x12,y12), cor(x13,y13),cor(x14,y14),
cor(x15,y15),cor(x16,y16),cor(x17,y17),cor(x18,y18), cor(x19,y19),cor(x20,y20),cor(x21,y21),
cor(x22,y22),cor(x23,y23), cor(x24,y24),cor(x25,y25),cor(x26,y26),cor(x27,y27), cor(x28,y28),
cor(x29,y29),cor(x30,y30))
Pa<- true.mean[,,1]
Pb<- true.mean.x[,,1]

#De información general de la población simulada.
table.infor <- cbind(pop.size, area.ID, m.tamao)
table.infor <- table.infor[order(table.infor[,1]),1:3]
medias <- c(mean(pop.size), NA, mean(m.tamao), mean(pop[,2]),mean(pop[,3]),cor(pop[,2],pop[,3]))
table.infor <- cbind(table.infor,Pa,Pb,corre)
table.infor <- rbind(table.infor,medias)
table.infor

#direct
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.direct,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.direct,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.direct,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.direct,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.direct <- cbind(pop.size, area.ID, m.direct, b.direct, mse.direct, rmean.MSE.direct)
table.direct <- table.direct[order(table.direct[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.direct), mean(b.direct), mean(mse.direct), mean(rmean.MSE.direct))
table.direct <- rbind(table.direct,medias)
table.direct <- rbind(table.direct,medias.grupo)

#ratio
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
```

```
c(0,0,0)->medias.grupo[,2]
aggregate(m.ratio,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.ratio,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.ratio,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.ratio,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.ratio <- cbind(pop.size, area.ID, m.ratio, b.ratio, mse.ratio, rmean.MSE.ratio)
table.ratio <- table.ratio[order(table.ratio[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.ratio), mean(b.ratio), mean(mse.ratio), mean(rmean.MSE.ratio))
table.ratio <- rbind(table.ratio,medias)
table.ratio <- rbind(table.ratio,medias.grupo)

#ratio1
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.ratio1,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.ratio1,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.ratio1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.ratio1,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.ratio1 <- cbind(pop.size, area.ID, m.ratio1, b.ratio1, mse.ratio1, rmean.MSE.ratio1)
table.ratio1 <- table.ratio1[order(table.ratio1[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.ratio1), mean(b.ratio1), mean(mse.ratio1), mean(rmean.MSE.ratio1))
table.ratio1 <- rbind(table.ratio1,medias)
table.ratio1 <- rbind(table.ratio1,medias.grupo)

#ratio2
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.ratio2,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.ratio2,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.ratio2,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.ratio2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.ratio2 <- cbind(pop.size, area.ID, m.ratio2, b.ratio2, mse.ratio2, rmean.MSE.ratio2)
table.ratio2 <- table.ratio2[order(table.ratio2[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.ratio2), mean(b.ratio2), mean(mse.ratio2), mean(rmean.MSE.ratio2))
table.ratio2 <- rbind(table.ratio2,medias)
```

```
table.ratio2 <- rbind(table.ratio2,medias.grupo)

#regresion
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.regresion,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.regresion,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.regresion,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.regresion,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.regresion <- cbind(pop.size, area.ID, m.regresion, b.regresion, mse.regresion, rmean.MSE.regresion)
table.regresion <- table.regresion[order(table.regresion[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.regresion), mean(b.regresion), mean(mse.regresion),
mean(rmean.MSE.regresion))
table.regresion <- rbind(table.regresion,medias)
table.regresion <- rbind(table.regresion,medias.grupo)

#regresion1
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.regresion1,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.regresion1,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.regresion1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.regresion1,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.regresion1 <- cbind(pop.size, area.ID, m.regresion1, b.regresion1, mse.regresion1,
rmean.MSE.regresion1)
table.regresion1 <- table.regresion1[order(table.regresion1[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.regresion1), mean(b.regresion1),
mean(mse.regresion1), mean(rmean.MSE.regresion1))
table.regresion1 <- rbind(table.regresion1,medias)
table.regresion1 <- rbind(table.regresion1,medias.grupo)

#regresion2
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
```

```
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.regresion2,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.regresion2,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.regresion2,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.regresion2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.regresion2 <- cbind(pop.size, area.ID, m.regresion2, b.regresion2, mse.regresion2,
rmean.MSE.regresion2)
table.regresion2 <- table.regresion2[order(table.regresion2[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.regresion2), mean(b.regresion2), mean(mse.regresion2),
mean(rmean.MSE.regresion2))
table.regresion2 <- rbind(table.regresion2,medias)
table.regresion2 <- rbind(table.regresion2,medias.grupo)

#diferencia
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.diferencia,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.diferencia,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.diferencia,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.diferencia,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.diferencia <- cbind(pop.size, area.ID, m.diferencia, b.diferencia, mse.diferencia,
rmean.MSE.diferencia)
table.diferencia <- table.diferencia[order(table.diferencia[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.diferencia), mean(b.diferencia), mean(mse.diferencia),
mean(rmean.MSE.diferencia))
table.diferencia <- rbind(table.diferencia,medias)
table.diferencia <- rbind(table.diferencia,medias.grupo)

#diferencia1
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.diferencia1,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.diferencia1,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.diferencia1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
```

```
aggregate(rmean.MSE.diferencia1,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.diferencia1 <- cbind(pop.size, area.ID, m.diferencia1, b.diferencia1, mse.diferencia1,
rmean.MSE.diferencia1)
table.diferencia1 <- table.diferencia1[order(table.diferencia1[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.diferencia1), mean(b.diferencia1), mean(mse.diferencia1),
mean(rmean.MSE.diferencia1))
table.diferencia1 <- rbind(table.diferencia1,medias)
table.diferencia1 <- rbind(table.diferencia1,medias.grupo)

#diferencia2
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.diferencia2,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.diferencia2,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.diferencia2,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.diferencia2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.diferencia2 <- cbind(pop.size, area.ID, m.diferencia2, b.diferencia2, mse.diferencia2,
rmean.MSE.diferencia2)
table.diferencia2 <- table.diferencia2[order(table.diferencia2[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.diferencia2), mean(b.diferencia2), mean(mse.diferencia2),
mean(rmean.MSE.diferencia2))
table.diferencia2 <- rbind(table.diferencia2,medias)
table.diferencia2 <- rbind(table.diferencia2,medias.grupo)

#cra1
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.cra1,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.cra1,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.cra1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.cra1,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.cra1 <- cbind(pop.size, area.ID, m.cra1, b.cra1, mse.cra1, rmean.MSE.cra1)
table.cra1 <- table.cra1[order(table.cra1[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.cra1), mean(b.cra1), mean(mse.cra1), mean(rmean.MSE.cra1))
table.cra1 <- rbind(table.cra1,medias)
```

```
table.cra1 <- rbind(table.cra1,medias.grupo)

#cra2
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.cra2,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.cra2,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.cra2,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.cra2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.cra2 <- cbind(pop.size, area.ID, m.cra2, b.cra2, mse.cra2, rmean.MSE.cra2)
table.cra2 <- table.cra2[order(table.cra2[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.cra2), mean(b.cra2), mean(mse.cra2), mean(rmean.MSE.cra2))
table.cra2 <- rbind(table.cra2,medias)
table.cra2 <- rbind(table.cra2,medias.grupo)

#cra3
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.cra3,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.cra3,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.cra3,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.cra3,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.cra3 <- cbind(pop.size, area.ID, m.cra3, b.cra3, mse.cra3, rmean.MSE.cra3)
table.cra3 <- table.cra3[order(table.cra3[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.cra3), mean(b.cra3), mean(mse.cra3), mean(rmean.MSE.cra3))
table.cra3 <- rbind(table.cra3,medias)
table.cra3 <- rbind(table.cra3,medias.grupo)

#cra4
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.cra4,list(d2=tamap),mean)[,2]->medias.grupo[,3]
```

```
aggregate(b.cra4,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.cra4,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.cra4,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.cra4 <- cbind(pop.size, area.ID, m.cra4, b.cra4, mse.cra4, rmean.MSE.cra4)
table.cra4 <- table.cra4[order(table.cra4[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.cra4), mean(b.cra4), mean(mse.cra4), mean(rmean.MSE.cra4))
table.cra4 <- rbind(table.cra4,medias)
table.cra4 <- rbind(table.cra4,medias.grupo)

#cra5
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.cra5,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.cra5,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.cra5,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.cra5,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.cra5 <- cbind(pop.size, area.ID, m.cra5, b.cra5, mse.cra5, rmean.MSE.cra5)
table.cra5 <- table.cra5[order(table.cra5[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.cra5), mean(b.cra5), mean(mse.cra5), mean(rmean.MSE.cra5))
table.cra5 <- rbind(table.cra5,medias)
table.cra5 <- rbind(table.cra5,medias.grupo)

#crega1
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.crega1,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.crega1,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.crega1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.crega1,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.crega1 <- cbind(pop.size, area.ID, m.crega1, b.crega1, mse.crega1, rmean.MSE.crega1)
table.crega1 <- table.crega1[order(table.crega1[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.crega1), mean(b.crega1), mean(mse.crega1),
mean(rmean.MSE.crega1))
table.crega1 <- rbind(table.crega1,medias)
table.crega1 <- rbind(table.crega1,medias.grupo)
```

```
#crega2
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.crega2,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.crega2,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.crega2,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.crega2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.crega2 <- cbind(pop.size, area.ID, m.crega2, b.crega2, mse.crega2, rmean.MSE.crega2)
table.crega2 <- table.crega2[order(table.crega2[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.crega2), mean(b.crega2), mean(mse.crega2),
mean(rmean.MSE.crega2))
table.crega2 <- rbind(table.crega2,medias)
table.crega2 <- rbind(table.crega2,medias.grupo)

#crega3
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.crega3,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.crega3,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.crega3,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.crega3,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.crega3 <- cbind(pop.size, area.ID, m.crega3, b.crega3, mse.crega3, rmean.MSE.crega3)
table.crega3 <- table.crega3[order(table.crega3[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.crega3), mean(b.crega3), mean(mse.crega3),
mean(rmean.MSE.crega3))
table.crega3 <- rbind(table.crega3,medias)
table.crega3 <- rbind(table.crega3,medias.grupo)

#crega4
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
```

```
aggregate(m.crega4,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.crega4,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.crega4,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.crega4,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.crega4 <- cbind(pop.size, area.ID, m.crega4, b.crega4, mse.crega4, rmean.MSE.crega4)
table.crega4 <- table.crega4[order(table.crega4[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.crega4), mean(b.crega4), mean(mse.crega4),
mean(rmean.MSE.crega4))
table.crega4 <- rbind(table.crega4,medias)
table.crega4 <- rbind(table.crega4,medias.grupo)

#crega5
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.crega5,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.crega5,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.crega5,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.crega5,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.crega5 <- cbind(pop.size, area.ID, m.crega5, b.crega5, mse.crega5, rmean.MSE.crega5)
table.crega5 <- table.crega5[order(table.crega5[,1]),1:6]
medias <- c(mean(pop.size), NA, mean(m.crega5), mean(b.crega5), mean(mse.crega5),
mean(rmean.MSE.crega5))
table.crega5 <- rbind(table.crega5,medias)
table.crega5 <- rbind(table.crega5,medias.grupo)

#logistic
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.logistic,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.logistic,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.logistic,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.logistic,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.logistic <- cbind(pop.size, area.ID, m.logistic, b.logistic, mse.logistic,
rmean.MSE.logistic)
table.logistic <- table.logistic[order(table.logistic[,1]),1:7]
```

```
medias <- c(mean(pop.size), NA, mean(m.logistic), mean(b.logistic), mean(mse.logistic),
mean(rmean.MSE.logistic))
table.logistic <- rbind(table.logistic,medias)
table.logistic <- rbind(table.logistic,medias.grupo)

#logistic1
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.logistic1,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.logistic1,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.logistic1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.logistic1,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.logistic1 <- cbind(pop.size, area.ID, m.logistic1, b.logistic1, mse.logistic1,
rmean.MSE.logistic1)
table.logistic1 <- table.logistic1[order(table.logistic1[,1]),1:7]
medias <- c(mean(pop.size), NA, mean(m.logistic1), mean(b.logistic1), mean(mse.logistic1),
mean(rmean.MSE.logistic1))
table.logistic1 <- rbind(table.logistic1,medias)
table.logistic1 <- rbind(table.logistic1,medias.grupo)

#logistic2
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.logistic2,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.logistic2,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.logistic2,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.logistic2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.logistic2 <- cbind(pop.size, area.ID, m.logistic2, b.logistic2, mse.logistic2,
rmean.MSE.logistic2)
table.logistic2 <- table.logistic2[order(table.logistic2[,1]),1:7]
medias <- c(mean(pop.size), NA, mean(m.logistic2), mean(b.logistic2), mean(mse.logistic2),
mean(rmean.MSE.logistic2))
table.logistic2 <- rbind(table.logistic2,medias)
table.logistic2 <- rbind(table.logistic2,medias.grupo)
```

```
#logistic3
medias.grupo<-array(0,dim=c(3,6))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(m.logistic3,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(b.logistic3,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(mse.logistic3,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(rmean.MSE.logistic3,list(d2=tamap),mean)[,2]->medias.grupo[,6]
table.logistic3 <- cbind(pop.size, area.ID, m.logistic3, b.logistic3, mse.logistic3,
rmean.MSE.logistic3)
table.logistic3 <- table.logistic3[order(table.logistic3[,1]),1:7]
medias <- c(mean(pop.size), NA, mean(m.logistic3), mean(b.logistic3), mean(mse.logistic3),
mean(rmean.MSE.logistic3))
table.logistic3 <- rbind(table.logistic3,medias)
table.logistic3 <- rbind(table.logistic3,medias.grupo)

#Sesgo relativo empírico de los estimadores.
RB.direct <- round((b.direct/Pa)*100,2)
RE.direct <- round((mse.direct/mse.direct)*100,2)
RB.ratio <- round((b.ratio/Pa)*100,2)
RE.ratio <- round((mse.direct/mse.ratio)*100,2)
RB.ratio1 <- round((b.ratio1/Pa)*100,2)
RE.ratio1 <- round((mse.direct/mse.ratio1)*100,2)
RB.ratio2 <- round((b.ratio2/Pa)*100,2)
RE.ratio2 <- round((mse.direct/mse.ratio2)*100,2)
RB.regresion <- round((b.regresion/Pa)*100,2)
RE.regresion <- round((mse.direct/mse.regresion)*100,2)
RB.regresion1 <- round((b.regresion1/Pa)*100,2)
RE.regresion1 <- round((mse.direct/mse.regresion1)*100,2)
RB.regresion2 <- round((b.regresion2/Pa)*100,2)
RE.regresion2 <- round((mse.direct/mse.regresion2)*100,2)
RB.diferencia <- round((b.diferencia/Pa)*100,2)
RE.diferencia <- round((mse.direct/mse.diferencia)*100,2)
RB.diferencia1 <- round((b.diferencia1/Pa)*100,2)
RE.diferencia1 <- round((mse.direct/mse.diferencia1)*100,2)
RB.diferencia2 <- round((b.diferencia2/Pa)*100,2)
RE.diferencia2 <- round((mse.direct/mse.diferencia2)*100,2)
RB.cra1 <- round((b.cra1/Pa)*100,2)
RE.cra1 <- round((mse.direct/mse.cra1)*100,2)
```

```
RB.cra2 <- round((b.cra2/Pa)*100,2)
RE.cra2 <- round((mse.direct/mse.cra2)*100,2)
RB.cra3 <- round((b.cra3/Pa)*100,2)
RE.cra3 <- round((mse.direct/mse.cra3)*100,2)
RB.cra4 <- round((b.cra4/Pa)*100,2)
RE.cra4 <- round((mse.direct/mse.cra4)*100,2)
RB.cra5 <- round((b.cra5/Pa)*100,2)
RE.cra5 <- round((mse.direct/mse.cra5)*100,2)
RB.crega1 <- round((b.crega1/Pa)*100,2)
RE.crega1 <- round((mse.direct/mse.crega1)*100,2)
RB.crega2 <- round((b.crega2/Pa)*100,2)
RE.crega2 <- round((mse.direct/mse.crega2)*100,2)
RB.crega3 <- round((b.crega3/Pa)*100,2)
RE.crega3 <- round((mse.direct/mse.crega3)*100,2)
RB.crega4 <- round((b.crega4/Pa)*100,2)
RE.crega4 <- round((mse.direct/mse.crega4)*100,2)
RB.crega5 <- round((b.crega5/Pa)*100,2)
RE.crega5 <- round((mse.direct/mse.crega5)*100,2)
RB.logistic <- round((b.logistic/Pa)*100,2)
RE.logistic <- round((mse.direct/mse.logistic)*100,2)
RB.logistic1 <- round((b.logistic1/Pa)*100,2)
RE.logistic1 <- round((mse.direct/mse.logistic1)*100,2)
RB.logistic2 <- round((b.logistic2/Pa)*100,2)
RE.logistic2 <- round((mse.direct/mse.logistic2)*100,2)
RB.logistic3 <- round((b.logistic3/Pa)*100,2)
RE.logistic3 <- round((mse.direct/mse.logistic3)*100,2)

#Tabla resúmen del sesgo emírico de los estimadores
table.RB <- cbind(pop.size, area.ID, RB.direct, RB.ratio, RB.ratio1,RB.ratio2, RB.regresion,
RB.regresion1,RB.regresion2,RB.diferencia,RB.diferencia1, RB.diferencia2,RB.cra1,RB.cra2,
RB.cra3,RB.cra4,RB.cra5,RB.crega1,RB.crega2,RB.crega3, RB.crega4,RB.crega5, RB.logistic,
RB.logistic1,RB.logistic2,RB.logistic3)
table.RB <- table.RB[order(table.RB[,1]),1:26]
medias <- c(mean(pop.size), NA, mean(RB.direct), mean(RB.ratio), mean(RB.ratio1),mean(RB.ratio2),
mean(RB.regresion),mean(RB.regresion1),mean(RB.regresion2),mean(RB.diferencia),mean(RB.diferencia1),
mean(RB.diferencia2),mean(RB.cra1),mean(RB.cra2),mean(RB.cra3), mean(RB.cra4), mean(RB.cra5),
mean(RB.crega1),mean(RB.crega2),mean(RB.crega3),mean(RB.crega4),mean(RB.crega5),mean(RB.logistic),
mean(RB.logistic1),mean(RB.logistic2),mean(RB.logistic3))
table.RB <- rbind(table.RB,medias)

medias.grupo<-array(0,dim=c(3,26))
```

```
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(RB.direct,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(RB.ratio,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(RB.ratio1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(RB.ratio2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
aggregate(RB.regresion,list(d2=tamap),mean)[,2]->medias.grupo[,7]
aggregate(RB.regresion1,list(d2=tamap),mean)[,2]->medias.grupo[,8]
aggregate(RB.regresion2,list(d2=tamap),mean)[,2]->medias.grupo[,9]
aggregate(RB.diferencia,list(d2=tamap),mean)[,2]->medias.grupo[,10]
aggregate(RB.diferencia1,list(d2=tamap),mean)[,2]->medias.grupo[,11]
aggregate(RB.diferencia2,list(d2=tamap),mean)[,2]->medias.grupo[,12]
aggregate(RB.cra1,list(d2=tamap),mean)[,2]->medias.grupo[,13]
aggregate(RB.cra2,list(d2=tamap),mean)[,2]->medias.grupo[,14]
aggregate(RB.cra3,list(d2=tamap),mean)[,2]->medias.grupo[,15]
aggregate(RB.cra4,list(d2=tamap),mean)[,2]->medias.grupo[,16]
aggregate(RB.cra5,list(d2=tamap),mean)[,2]->medias.grupo[,17]
aggregate(RB.crega1,list(d2=tamap),mean)[,2]->medias.grupo[,18]
aggregate(RB.crega2,list(d2=tamap),mean)[,2]->medias.grupo[,19]
aggregate(RB.crega3,list(d2=tamap),mean)[,2]->medias.grupo[,20]
aggregate(RB.crega4,list(d2=tamap),mean)[,2]->medias.grupo[,21]
aggregate(RB.crega5,list(d2=tamap),mean)[,2]->medias.grupo[,22]
aggregate(RB.logistic,list(d2=tamap),mean)[,2]->medias.grupo[,23]
aggregate(RB.logistic1,list(d2=tamap),mean)[,2]->medias.grupo[,24]
aggregate(RB.logistic2,list(d2=tamap),mean)[,2]->medias.grupo[,25]
aggregate(RB.logistic3,list(d2=tamap),mean)[,2]->medias.grupo[,26]
table.RB <- rbind(table.RB,medias.grupo)

#Tabla de eficiencia empírica de los estimadores

table.RE <- cbind(pop.size, area.ID, RE.direct, RE.ratio, RE.ratio1,RE.ratio2, RE.regresion,
RE.regresion1,RE.regresion2,RE.diferencia,RE.diferencia1,RE.diferencia2,RE.cra1,RE.cra2,RE.cra3,
RE.cra4,RE.cra5,RE.crega1,RE.crega2,RE.crega3,RE.crega4,RE.crega5,RE.logistic,RE.logistic1,
RE.logistic2,RE.logistic3)
table.RE <- table.RE[order(table.RE[,1]),1:26]
medias <- c(mean(pop.size), NA, mean(RE.direct), mean(RE.ratio), mean(RE.ratio1),mean(RE.ratio2),
mean(RE.regresion),mean(RE.regresion1),mean(RE.regresion2),mean(RE.diferencia),mean(RE.diferencia1),
mean(RE.diferencia2),mean(RE.cra1),mean(RE.cra2),mean(RE.cra3),mean(RE.cra4),mean(RE.cra5),
mean(RE.crega1),mean(RE.crega2),mean(RE.crega3),mean(RE.crega4),mean(RE.crega5),mean(RE.logistic),
```

```
mean(RE.logistic1), mean(RE.logistic2),mean(RE.logistic3))
table.RE <- rbind(table.RE,medias)

medias.grupo<-array(0,dim=c(3,26))
levels(as.factor(pop.size))->tipo
recode(pop.size, "500=1; 750=2; 1000=2; 1250=2; 1500=3")->tamap
aggregate(pop.size,list(d2=tamap),mean)[,2]->medias.grupo[,1]
c(0,0,0)->medias.grupo[,2]
aggregate(RE.direct,list(d2=tamap),mean)[,2]->medias.grupo[,3]
aggregate(RE.ratio,list(d2=tamap),mean)[,2]->medias.grupo[,4]
aggregate(RE.ratio1,list(d2=tamap),mean)[,2]->medias.grupo[,5]
aggregate(RE.ratio2,list(d2=tamap),mean)[,2]->medias.grupo[,6]
aggregate(RE.regresion,list(d2=tamap),mean)[,2]->medias.grupo[,7]
aggregate(RE.regresion1,list(d2=tamap),mean)[,2]->medias.grupo[,8]
aggregate(RE.regresion2,list(d2=tamap),mean)[,2]->medias.grupo[,9]
aggregate(RE.diferencia,list(d2=tamap),mean)[,2]->medias.grupo[,10]
aggregate(RE.diferencia1,list(d2=tamap),mean)[,2]->medias.grupo[,11]
aggregate(RE.diferencia2,list(d2=tamap),mean)[,2]->medias.grupo[,12]
aggregate(RE.cra1,list(d2=tamap),mean)[,2]->medias.grupo[,13]
aggregate(RE.cra2,list(d2=tamap),mean)[,2]->medias.grupo[,14]
aggregate(RE.cra3,list(d2=tamap),mean)[,2]->medias.grupo[,15]
aggregate(RE.cra4,list(d2=tamap),mean)[,2]->medias.grupo[,16]
aggregate(RE.cra5,list(d2=tamap),mean)[,2]->medias.grupo[,17]
aggregate(RE.crega1,list(d2=tamap),mean)[,2]->medias.grupo[,18]
aggregate(RE.crega2,list(d2=tamap),mean)[,2]->medias.grupo[,19]
aggregate(RE.crega3,list(d2=tamap),mean)[,2]->medias.grupo[,20]
aggregate(RE.crega4,list(d2=tamap),mean)[,2]->medias.grupo[,21]
aggregate(RE.crega5,list(d2=tamap),mean)[,2]->medias.grupo[,22]
aggregate(RE.logistic,list(d2=tamap),mean)[,2]->medias.grupo[,23]
aggregate(RE.logistic1,list(d2=tamap),mean)[,2]->medias.grupo[,24]
aggregate(RE.logistic2,list(d2=tamap),mean)[,2]->medias.grupo[,25]
aggregate(RE.logistic3,list(d2=tamap),mean)[,2]->medias.grupo[,26]
table.RE <- rbind(table.RE,medias.grupo)

# Sumario de tablas.
table.infor
round(table.direct,3)
round(table.ratio,3)
round(table.ratio1,3)
round(table.ratio2,3)
round(table.regresion,3)
```

```
round(table.regresion1,3)
round(table.regresion2,3)
round(table.diferencia,3)
round(table.diferencia1,3)
round(table.diferencia2,3)
round(table.cra1,3)
round(table.cra2,3)
round(table.cra3,3)
round(table.cra4,3)
round(table.cra5,3)
round(table.crega1,3)
round(table.crega2,3)
round(table.crega3,3)
round(table.crega4,3)
round(table.crega5,3)
round(table.logistic,3)
round(table.logistic1,3)
round(table.logistic2,3)
round(table.logistic3,3)
round(table.RB,2)
round(table.RE,2)

#La segunda población se simula usando la proporción
(39:10)/41->prop
```

Índice de Tablas

Bibliografía

[1] Anderson, T.W., and Hsiao, C. (1981), Asymptotically Efficient Estimation of Covariance Matrices with Linear Covariance Structure, *Annals of Statistics*, 1, 135-141.

[2] Ansley, C.F., and Kohn, R.(1986), Prediction Mean Squared Error for State Space Models with Estimated Parameters, *Biometrika*, 73, 467-473.

[3] Aragon, Y. (1984), Random Variance Linear Models: Estimation, *Computational Statistics Quarterly*, 1, 295-309.

[4] Arora, V., and Lahiri, P. (1997), On the Superiority of the bayes Method over the BLUP in Small Area Estimation Problems, *Statistica Sinica*, 7, 1053-1063.

[5] Banerjee, M., and Fress, E.W. (1977), Influence Diagnostics for Linear Longitudinal Models, *Journal of the American Statistical Association*, 92, 999-1005.

[6] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988), An Error Component Model for prediction of County Crop Areas Using Survey and Satelite Data, *Journal of the American Statistical Association*, 83, 28-36

[7] Bayarri, M.J. and Berger, J. (2000), P Values for Composite Null Models, *Journal of the American Statistical Association*, 95, 1127-1142.

[8] Beckman, R.J., Nachtsheim, C.J., and Cook, R.D. (1987), Diagnostic for Mixed-Model Analysis of Variance, *Technometrics*, 1987, 413-426.

[9] Bell, W.R. (1999), Accounting for Uncertainty About Variances in Small Area Estimation, *Bulletin of the international Statistical Institute*.

[10] Berger, J.O., and Pericchi, L.R. (2001), Objective bayesian Methods for Model Selection: Introduction and Comparison, in P. Lahiri(ed.), *Model Selection*, Lecture Notes-Monograph Series, Volume 38, Beachwood, OH: Institute of Mathematical Statistics.

[11] Bilodeau, M., and Srivastava, M.S. (1988), Estimation of the MSE Matrix of the Stein Estimator, *Canadian Journal of Statistics*, 16, 153-159.

[12] Brandwein, A.C., and Strawderman, W.E. (1990), Stein-Estimation: The Spherically Symmetric Case, *Statistical Science*, 5, 356-369.

[13] Breslow, N., and Clayton, D. (1993), Approximate Inference in Generalized Linear Mixed Models, *Journal of the American Statistical Association*, 88, 9-25.

[14] Brooks, S.P., Catchpole, E.A. and Morgan, B.J. (2000), bayesian Animal Survival Estimation, *Statistical Science*, 15, 357-376.

[15] Browne, W.J., and Draper, D. (2001), A Comparison of bayesian and Likelihood-based Methods for Fitting Multilevel Models, Technixal Report, Institute for Education, London.

[16] Butar, F.B., and Lahiri, P. (2001), On Measures of Uncertainty of Empirical bayes Small-Area Estimators, Technical Report, Department of Statistics, University of Nabraska, Lincoln.

[17] Casady, R.J., and Valliant, R. (1993), Conditional Properties of Poststratified Estimators Under Normal Theory, *Survey Methodology*, 19, 183-192.

[18] Casella, G., and Berger, R.L. (1990), *Statistical Inference*, Belmonte, CA:Wadsworth & Brooks/Cole.

[19] Casella, G., Lavine, M., and Robert, C.P. (2001), Explaining the Perfect Sampler, *American Statistician*, 55, 299-305.

[20] Chambers, R.L and Dunstan, R., 1986. Estimating distribution functions from survey data. *Biometrika*, **73**, 597-604

[21] Chambers, R.L and Tzavidis, N., M-quantile models for small area estimation, *Biometrika*(2006), **93**, 2, pp. 255-268.

[22] Chandra, H., Chambers, R. and Salvati, N., Small area estimation of proportions in Business Survey. *Center for Statistical & Survey Methodology*. Workin paper 15(2009).

[23] Christiansen, C.L., and Morris, C.N. (1997), Hierarchical Poisson Regression Modeling, *Journal of the American Statistical Association*, 92, 618-632.

[24] Christiansen, C.L., Pearson, L.M., and Johnson, W. (1992), Case-Deletion Diagnostics for Mixed Models, *Technometrics*,34, 38-45.

[25] Clayton, D., and Bernardinelli, L. (1992), bayesian for Mapping Disease Risk, in P. Elliot, J. Cuzick, D. English and R. Stern (Eds.), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, London: Oxford University Press.

[26] Clayton, D., and Kaldor, J. (1987), Empirical bayes Estimates of Age Standardized Relative Risks for Use in Disease Mapping, *Biometrics*, 43, 671-681.

[27] Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed., Mew York: Wiley.

[28] Cook, R.D. (1977), Detection of Influential Observations in Linear Regression, *Technometrics*, 19, 15-18.

[29] Cook, R.D. (1986), Assessment of Local Influence, *Journal of the Royal Statistical Society, Series B*,48, 135-155.

[30] Cowles, M.K. and Karlin, B.P. (1996), Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, *Journal of the American Statistical Association*, 91, 883-904.

[31] Cressie, N. (1991), Small-Area Prediction of Undercount Using the General Linear Model, *Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada, pp. 93-105.

[32] Cressie, N. (1992), REML Estimation in Empirical bayes Smoothing of Census Undercount, *Survey Methodology*, 18, 75-94.

[33] Cressie, N., and Chan, N.H. (1989), Spatial Modelling of Reginal Variables, *Journal of the American Statistical Association*, 84, 393-401.

[34] Daniels, M.J., and Kass, R.E. (1999), Nonconjugate bayesian Estimation of Covariance Matrices and Its Use in Hierarchical Models, *Journal of the American Statistical Association*, 94, 1254-1263.

[35] Das, K., Jiang, J., and Rao, J.N.K. (2001), Mean Squared Error of Empirical Predictor, Technical Report, School of Mathematics and Statistics, Carleton University, Ottawa, Canada.

[36] Datta, G.S., Day, B., and Basawa, I. (1999), Empirical Best Linear Unbiased and Empirical bayes Prediction in Multivariate Small Area Estimation, *Journal of Statistical Planning and Inference*, 75, 169-179.

[37] Datta, G.S., Ghosh, M., and Waller, L.A. (2000), Hierarchical and Empirical bayes Methods for Environmental Risk Assessment, in P.K. Sen and C.R. Rao (Eds.), *Handbook of Statistics*, Vlume 18, Amsterdam: Elsevier Science B.V., pp. 223-245.

[38] Datta, G.S., and Lahiri, P. (2000), A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problemas, *Statistica Sinica*, 10, 613-627.

[39] Datta, G.S., Lahiri, P., and Maiti, T. (2002), Empirical bayes Estimation of Median Income of Four-Person Families by State Using Time Series and Cross-Sectional Data, *Journal of Statistical Planning and Inference*, 102, 83-97.

[40] Datta, G.S., and Ghosh, M. (1991), bayesian Prediction in Linear Models: Applications to Small Area Estimation, *Annals of Statistics*, 19, 1748-1770.

[41] Datta, G.S., Ghosh, M., Smith, D.D., and Lahiri, P. (2002), On the Asymptotic Theory of Conditional and Unconditional Coverage Probabilities of Empirical bayes Confidence Intervals, *Scandinavian Journal of Statistics*, 29, 139-152.

[42] Datta, G.S., Kubakawa, K., and Rao, J.N.K. (2002), Estimation of MSE in Small Area Estimation, Technical Report, Departament of Statistics, University of Georgia, Athens.

[43] Datta, G.S., Rao, J.N.K., and Smith, D.D. (2002), On Measures of Uncertainty of Small Area Estimation in the Fay-Herriot Model, Technical Report, University of Georgia, Athens.

[44] Dawid, A.P. (1985), Calibration-Based Empirical Probability, *Annals of Statistics*, 13, 1251-1274.

[45] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

[46] DeSouza, C.M. (1992), An Appropriate Bivariate bayesian Method for Analysing Small Frequencies, *Biometrics*, 48, 1113-1130.

[47] Deville, J.C., and Särndal, C.E. (1992), Calibration Estimation in Survey Sampling, *Journal of the American Statistical Association*, 87, 376-382.

[48] Duchesne, P., Estimation of a proportion with survey data (2003). *Journal of Statistics Education* volume II, 3(2003).

[49] Efron, B., and Morris, C.E. (1973), Stein's Estimation Rule and Its Competitors - An Empirical bayes Approach, *Journal of the American Statistical Association*, 68, 117-130.

[50] Efron, B., and Morris, C.E. (1975), Data Analysis Using Stein's Estimate and Its Generalizations, *Journal of the American Statistical Association*, 70, 311-319.

[51] Ericksen, E.P. (1974), A Regression Method for Estimating Population Changes of Local Areas, *Journal of the American Statistical Association*, 69, 867-875.

[52] Estevao, V., Hidiroglou, M.A., and Särndal, C.E. (1995), Methodological Principles for a Generalized Estimations Systems at Statistics Canada, *Journal of Official Statistics*, 11, 181-204.

[53] Estevao, V. and Särndal, C., A Functional form Approach to Calibration. *Journal of Official Statistics*, vol. 16, No. 4 (2000), pp.379-399.

[54] Estevao, V. and Särndal, C., Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review (2006)*, 74, 2, 127-147.

[55] Eustat, (2002), Técnicas de Estimación en Áreas Pequeñas, Technical report.

[56] Falorsi, P. D., Falorsi, S. and Russo, A. (1994). Empirical comparison of small area estimation methods for the Italian Labour Force Survey. Survey Methodology 20, 171-176.

[57] Farrell, P.J., MacGibbon, B. and Tomberlin, T.J. (1997), Empirical bayes Estimators of Small area proportions in multistage designs; *Statistica Sinica*, **7**(1997), 1065-1083.

[58] Farrell, P.J. (2000), bayesian Inference for Small Area Proportions, *Sankhyâ, Series B*, 62, 402-416.

[59] Farrell, P.J., MacGibbon, B., and Thomberlin, T.J. (1977a), Empirical bayes Estimators of Small Area Proportions in Multistage Designs, *Statistica Sinica*, 7, 1065-1083.

[60] Fay, R.E., and Herriot, R.A. (1979), Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data, *Journal of the American Statistical Association*, 74, 269-277.

[61] Fuller, W. A. (1975), Regression Analysis for Sample Surveys, *Sankhyâ, Series C*, 37, 117-132.

[62] Fuller, W. A. (1999), Environmental Survey over Time, *Journal of Agricultural, Biological and environmental Statistics*, 4, 331-345.

[63] Fuller, W. A., and Battese, G.E. (1973), Transformations for Estimation of Linear Models with Nested-Error Structure, *Journal of the American Statistical Association*, 68, 626-632.

[64] Fuller, W. A., and Harter, R.M. (1987), The Multivariate Components of Variance Model for Small Area Estimation, in R. Platek, J.N.K. Rao, C.E. Särndal, and M.P. Singh (Eds.), *Small Area Statistics*, New York: Wiley, pp. 103-123.

[65] Gelfand, A.E. (1996), Model Determination Using Sample-Based Methods, in W.R. Gilks, S. Richardson, and Spiegelhalter (Eds.), *Monte Carlo Chain in Practice*, London: Champman and Hall, pp. 145-161.

[66] Gelfand, A.E., and Smith, A.F.M. (1990), Sample-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 972-985.

[67] Gelfand, A.E., and Smith, A.F.M. (1991), Gibbs Sampling for Marginal Posterior Expectations, *Communications in Statistics - Theory and Methods*, 20, 1747-1766.

[68] Gelfand, A.E., and Meng, S.L. (1996), Model Checking and Model Improvement, in W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall, pp. 189-201.

[69] Gelfand, A.E., and Rubin, D.B. (1992), Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, 7, 457-472.

[70] Ghosh, M. (1992b), Constrained bayes Estimation with Applications, *Journal of the American Statistical Association*, 87, 533-540.

[71] Ghosh, M., and Lahiri, P. (1987), Robust Empirical bayes Estimation of Means from Stratified Samples, *Journal of the American Statistical Association*, 82, 1153-1162.

[72] Ghosh, M., and Maiti, T. (1999), Adjusted bayes Estimators with Applications to Small Area Estimations, *Sankhyā, Series B*, 61, 71-90.

[73] Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B.P. (1998), Generalized Linear Models for Small Area Estimation, *Journal of the American Statistical Association*, 93, 273-282.

[74] Ghosh, M., Natarajan, K., Waller, L.A., and Kim, D.H. (1999), Hierarchical bayes GLMs for the Analysis of Spatial Data: An Aplication to Disease Mapping, *Journal of Statistical Planing and Inference*, 75, 305-318.

[75] Gilks, W.R., and Wild, P. (1992), Adaptive Rejection Sampling for Gibbs Sampling, *Applied Statistics*, 41, 337-348.

[76] Goldstein, H. (1989), Unbiased Interative Generalized Least Squares Estimation, *Biometrika*, 76, 622-623.

[77] Gonzalez, M.E. (1973), Use and Evaluation of Synthetic Estimates, *Proceedings of the Social Statistics Section*, American Statistical Asociation, pp. 33-36.

[78] Gonzalez, M.E., and Wakesberg, J. (1973), Estimation of the Error of Synthetic Estimates, Paper Presented at the First meeting of the International Association of Survey Statisticians, Vienna, Austria.

[79] Gonzalez-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. and Santamaria, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics & Data Analisis* **51**(2007) 2720-2733.

[80] Griffin, B., and Krutchkoff, R. (1971), Optimal Linear Estimation: An Empirical bayes Version with Application to the Binomial Distribution, *Biometrika*, 58, 195-203.

[81] Guggemos, F. and Tille, Y., Penalized Calibration in Survey Sampling: Design-based estimation assisted by mixedmodels. *Journal of Statistical Planning and Inference* 140(2010)31993212

[82] Harville, D.A., and Jeske, D.R. (1992), Mean Squared Error of Estimation or Prediction Under General Linear Models, *Journal of the American Statistical Association*, 87, 724-731.

[83] Hastings, W,K., (1970), Monte Carlo Sampling Methods Using Markov Chains and Their Aplications, *Biometrika*, 57, 97-109.

[84] Henderson, C.R. (1975), Best Linear Unbiased Estimation and Prediction Under a Selection Model, *Biometrics*, 31, 423-447.

[85] Hobert, J.P., and Casella, G. (1996), The Effect of Improper Prior on Gibbs Sampling in Hierarchical Linear Mixed Models, *Journal of the American Statistical Association*, 91, 1473.

[86] Holt, D., and Smith, T.M.F. (1979), Post-Stratification, *Journal of the Royal Statistical Society, Series A*, 142, 33-46.

[87] IASS, Small Area Estimation, Publicación de las ponencias de la conferencia satélite celebrada en 1999 en Riga, Letonia.

[88] Jiang, J., Lahiri, P., and Wan, S-M. (2002), A unified Jackknife Theory, *Annals of Statistics*, 30, in press.

[89] Jiang, J., and Lahiri, P. (2001), Empirical Best Prediction for Small Area Inference with Binari Data, *Annals of the Institute of Statistical Mathematics*, 53, 217-243.

[90] Jiang, J., Lahiri, P., 2006. Mixed model prediction and small area estimation. TEST 15 (1), 196. URL: http://dx.doi.org/10.1007/BF02595419.

[91] Kackar, R.N., and Harville, D.A. (1981), Unbiasedness of Two-stage Estiamtion and Prediction Procedures for Mixed Linear Models, *Communications in Statistics, Series A*, 10, 1249-1261.

[92] Kackar, R.N., and Harville, D.A. (1984), Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models, *Journal of the American Statistical Association*, 79, 853-862.

[93] Kass, R.E., and Raftery, A. (1995), bayes Factors, *Journal of the American Statistical Association*, 90, 773-795.

[94] Kim, H., Sun, D., and Tsutakawa, R.K. (2001), A Bivariate bayes Method for Improving the Estimates of Mortality Rates with a Twofold Conditional Autorregressive Model, *Journal of the American Statistical Association*, 96, 1506-1521.

[95] Kim, J-M, Sungur, E.A. and Heo, T-Y, Calibration Approach Estimator in Stratified Sampling. *Statistics & Probability Letters*, 77(2007), 99-103.

[96] Kim, K. and Park, M., Calibration Estimation in Survey Sampling. *International Statistical Review (2010)*, 78, 1, 21-39.

[97] Kleffe, J., and Rao, J.N.K. (1992), Estimation of mean Square Error of Empirical Best Linear Unbiased Predictors Under a Random Error Variance Linear Model, *Journal of Multivariate Analysis*, 43, 1-15.

[98] Kott, P.S., A Practical Use for Instrumental-variable Calibration. *Joint Statistical Meetings-Section on Survey Research Methods*, (2002).

[99] Kott, P.S., Using Calibration Weighting to Ajust for Non response and coverage errors. *Survey Methodology (2006)*, vol. 32, No. 2, pp.133-142.

[100] Krapavickaitė, D. and Plicusas, A., Estimation of a Ratio in the Finite Population. *Informatica (2005)*, vol.16, No.3, 343-364.

[101] Lahiri, P., and Maiti, T. (199), Empirical bayes Estimation of Relative Risks in Disease Mapping. Technical Report, Department of Statistics, University of Nebraska, Lincoln.

[102] Laird, N.M., and Louis, T.A. (1987), Empirical bayes Confidence Intevals Based on Bootstrap Samples, *Journal of the American Statistical Association*, 82, 739-750.

[103] Larsen, M.D., Estimation of Small-area proportion using covariates and survey data. *Journal of Statistical Planing and Inference* 112(2003) 89-98.

[104] Lehtonen, R., Särndal, C.E. and Veijanen, A., Model calibration and generalized regression estimation for domains and small areas.

[105] Lehtonen, R., Särndal, C.E. and Veijanen, A., Generalized Regression and model-calibration estimation for domains: Accuracy Comparizon.

[106] Liu, B. (2009), Adaptive Hierarchical bayes Estimation of small-area proportions. *Social Statistics Section*-JSM 2009.

[107] Longford, N. T. (2007). On standard errors of model-based small-area estimators, *Survey Methodology*, 33,69-79.

[108] Lohr, S.L., and Prasad, N.G.N. (2001), Small Area Estimation with Auxiliary Survey Data, Technical Report, Department of Mathematics, Arizona State University, Tempe.

[109] MacGibbon, B., and Tomberlin, T.J. (1989), Samall Area Estiamtion of Proportions Via Empirical bayes Techniques, *Survey Methodology*, 15, 237-252.

[110] Maiti, T. (1998), Hierarchical bayes Estimation of Motality Rates for Disease Mapping, *Journal of Statistical Planning and Inference*, 69, 339-348.

[111] Malec, D. Sedransk, J., Moriarity, C.L., and LeClere, F.B. (1997). Samall Area Inference for Binary Variables in National Health Interview Survey, *Journal of the American Statistical Association*, 92, 815-826.

[112] Menéndez, E. y Ferrales, J., El estimador de razón generalizado. *Trabajos de estadística*, vol.4, No.1, 1989. pp.3-11.

[113] McCulloch, C.E., and Searle, S.R. (2001), *Generalized, Linear, and Mixed Models*, New York: Wiley.

[114] Montanari, G.E. and Ranalli, G., On calibration methods for design based finite population inferences. *Departament of Satatistical Sciences* University of Perugia, (2003).

[115] Moura, F.A.S., and Holt, D. (1999), Small Area Estimation Using Multilevel Models, *Survey Methodology*, 25, 73-80.

[116] Moura, F.A.S., and Migon, H.S. (2002), bayesian Spatial Model for Small Area Estimation of Proportions. *Statistical Modelling* (2002),**2**: 183-201.

[117] Murgui, J.S. y Ayvar, C., Estimadores de regresión y razón para proporciones. *Estadística Española*, vol.37, Núm. 138, 1995, pp.5-13.

[118] Natarajan, R., and Kass, R.E. (2000), Reference bayesian Methods for Generalized Linear Mixed Models, *Journal of the American Statistical Association*, 95, 227-237.

[119] Olkin, I., "Multivariate ratio estimation for finite populations," *Biometrika*, 45 (1958), 154-65.

[120] Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., Breidt, F.J., 2008. Nonparametric small area estimation using penalized spline regression. Journal of the Royal Statistical Society. Series B 70, 265-286.

[121] Prasad, N.G.N., and Rao, J.N.K. (1990), The Estimation of the Mean Squared Error of Small-Area Extimators, *Journal of the American Statistical Association*, 85, 163-171.

[122] Pratesi, M., Ranalli, G. y Salvati, N., Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US, *Environmetrics* **19** (2008), pp. 687-701.

[123] Pratesi, M., Ranalli, G. y Salvati, N. (2009), Nonparametric M-quantile regression using penalised splines, *Journal of Nonparametric Statistics*, **21**:3, 287-304.

[124] Proyecto EURAREA (Enhancing Small Area Estimation Techniques to meet European needs), (2005), Technical report.

[125] Rao, C.R. (1971), Estimation of Variance and Covariance Components MINQUE Theory, *Journal of Multivariate Analysis*, 1, 257-275.

[126] Rao, J.N.K. (2003), Small Area Estimation, New York: Wiley.

[127] Rao, J.N.K., and Yu, M. (1992), Small Area Estimation by Combining Time Series and Cross-Sectional Data, *Proceedings of the Section on Survey Research Method*, American Statistical Association, pp. 1-9.

[128] Rivest, L-P., and Belmonte, E. (2000), A Conditional Mean Squared Error of Small Area Estimators, *Survey Methodology*, 26, 67-78.

[129] Robinson, G.K., (1991), That BLUP Is a Good Thing: The Estimation of Random Effects, *Statistical Science*, 6, 15-31.

[130] Rueda, M. y Arcos, A., Sobre un estimador de razón múltiple. *Estadística Española*, vol.36, Núm. 137, 1994, pp. 459-471.

[131] Saei, A. and Chambers, R., Small Area Estimation: A review of methods based on the application of mixed models. *SRI Methology Working paper* M03/16 (2003).

[132] Salvati, N., Chandrab, H., Ranalli, G. and Chambers, R. Small area estimation using a nonparametric model-based direct estimator. *Computational Statistics and Data Analysis* 54 (2010) 2159-2171.

[133] Särndal, C. E. (1984). Design-consistent versus model-dependent estimation for small domains. Journal of the American Statistical Association 79, 624-631.

[134] Särdnal, C.E., and Hidiroglou, M.A. (1989), Small Domain Estimation: A Conditional Analysis, *Journal of the American Statistical Association*, 84, 266-275.

[135] Särndal, C.E., The Calibration Approach in Survey Theory and Practice. *Survey Methology (2007)*, vol.33, No.2, pp.99-119.

[136] Singh, A.C., Mantel, J.H., and Thomas, B.W. (1994), Time Series EBLUPs for Small Areas Using Survey Data, *Survey Methodology*, 20, 33-43.

[137] Singh, S. and Arnab, R., On Calibration of design weights.*Statistics - Methodology*(2009).

[138] Spiegelman, D., Carrol, R. and Kipnis, V., Efficient Regression Calibration for Logistic Regression. *Statistics in Medicine*, **20**(2001), 139-160.

[139] Spjøtvoll, E., and Thomsen, I. (1987), Application of Some Empirical bayes Methods to Samll Area Statistics, *Bulletin of the International statistical Institute (Vol. 2)*, pp.435-449.

[140] Srivastava, M.S., and Bilodeau, M. (1989), Stein Estimation Under Elliptical Distributions, *Journal of Multivariate Analysis*, 28, 247-259.

[141] Stukel, D.M., and Rao, J.N.K. (1999), Small-Area Estimation Under Twofold Nested Errors Regression Models, *Journal of Statistical Planning and Inference*, 78, 131-147.

[142] Theberge, A., Extensions of Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, vol. 94, No. 446(1999) pp. 635-644.

[143] Trevisani, M. and Torelli, N., Hierarchical bayesian models for small area estimation with count data. *Working paper* no. 115, Università degli Studi di Triste, Dipartamento di Scienze Economiche e Statistiche (2007).

[144] Wolter, K. (1985), Introduction to Variance Estimation, New York: Springer-Verlag.

[145] Wright, R.L., Finite population sampling with multivariate auxiliary onformation. *Journal of the American Statistical Association*, vol. 78, No. 384(1983) pp. 879-884.

[146] Wu, Ch. and Sitter, R.R, A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data. *Journal of the American Statistical Association*, vol. 96, No. 453(2001) pp. 185-193.

[147] Wu, Ch, Optimal Calibration Estimators in Survey Sampling. *Biometrika (2003)*, 90, 4, pp. 937-951.

[148] You, Y., and Rao, J.N.K. (2002b), Samall Area Estimation Using Unmatched Sampling and Linking Models, *Canadian Journal of statistics*, 30, 3-15.

[149] Xie, D., Raghunatan, T.E. and Lepkowski, J.M., Estimation of the proportion of overweight individuals in small areas –a robust extension of the Fay-Herriot model. *Statist. Med.* (2007); **26**: 2699-2715.

[150] Research Online is the open access institutional repository for the University of Wollongong. For further information contact Manager Repository Services: morgan@uow.edu.au.