UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

COMPARACIÓN EMPÍRICA DE ESTIMADORES DEL ERROR CUADRÁTICO MEDIO PARA COEFICIENTES DE REGRESIÓN LINEAL Y LOGÍSTICA EN DISEÑOS COMPLEJOS

TESIS

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS MATEMÁTICAS

P R E S E N T A :

ALBERTO MANUEL PADILLA TERÁN

DIRECTOR DE TESIS: DR. IGNACIO MÉNDEZ RAMÍREZ

MÉXICO, D.F. **ENERO 2008**

DIVISION DE ESTUDIOS
DE POSGRADO

Quisiera agradecer,

a la Universidad Nacional Autónoma de México, por la formación recibida durante los estudios de maestria,

al Dr. Ignacio Méndez, por la oportunidad de colaborar en el tema de la tesis, el cual permitió que me adentrara en diversos aspectos relacionados, así como por su disposición para tratar puntos en cualquier momento,

a Catalina Palmer, Silvia Ruiz-Velasco y Patricia Romero, por los comentarios y observaciones hechos al trabajo,

a Ramsés Mena por sus comentarios y apoyo, que mejoraron tanto el contenido como la presentación de la tesis,

a Paco por su ayuda en Tex y R,

a mi madre y mis hermanos, por el cariño que siempre me han dado.

Con amor para Xóchitl

Índice

Índice de figuras

1. Introducción

Actualmente, la mayor parte de las encuestas realizadas por agencias gubernamentales o empresas de investigación de mercado, entre otras, son levantadas por medio de diseños muestrales complejos que incluyen estratificación, varias etapas de selección, así como probabilidades desiguales de selección de los elementos de la población, ver *Wolter*[31]. Las agencias o empresas liberan al público o a sus clientes los datos resultado de las mediciones de interés, y en particular los denominados factores de expansión, que son el inverso de la probabilidad de selección de los elementos de la muestra. A esta probabilidad se le conoce como probabilidad de inclusión de primer orden. Los usuarios de la información están interesados en estimar totales de la población objeto de la encuesta, así como la varianza o error cuadrático medio de dichos estimadores.

Por otra parte, cada vez es más frecuente la realización de análisis de regresión lineal o de datos categóricos con los resultados de dichas encuestas. La estimación puntual de los coeficientes de regresión lineal o de una regresión logística, se calculan usando los factores de expansión como pesos o ponderadores; sin embargo, la estimación del error cuadrático medio de los coeficientes considerando las características del diseño complejo requiere del conocimiento de las probabilidades de inclusión de segundo orden. Infortunadamente, estas probabilidades no se entregan con los resultados.

En caso de que no se conozcan las probabilidades de inclusión de segundo orden se emplean métodos que estiman la varianza o el error cuadrático medio:

- i) en ciertos casos es posible aproximar las probabilidades de segundo orden con base en las probabilidades de primer orden, véase *Tillè* [28];
- ii) dependiendo del tipo de estadística, pueden usarse métodos de remuestreo como el jacknife, bootstrap, grupos aleatorios, medias muestras balanceadas, entre otros;
- iii) se emplea el supuesto de asociar a las unidades primarias de muestreo, variables aleatorias independientes e idénticamente distribuidas al interior de cada estrato.

Este último método es empleado por algunos paquetes estadísticos comerciales para estimar totales y estimaciones de varianza con datos provenientes de encuestas complejas. Cabe aclarar que en algunos de estos paquetes puede calcularse la varianza estimada del estimador linealizado por

Taylor considerando todos los aspectos del diseño complejo; empero, es necesario proporcionar información detallada de tamaños de conglomerados para cada etapa de submuestreo, los cuales en muchas ocasiones no se tienen.

Así, en el presente trabajo se comparan empíricamente los errores cudráticos medios estimados por el método (iii) arriba mencionado con las varianzas aproximadas estimadas del estimador linealizado por Taylor que incluye todos los aspectos del diseño complejo, para los coeficientes en regresión lineal y logística en una población finita. También se hace lo propio con el estimador del total usando un estimador de regresión generalizado, *Särndal et al.*[24]. Para ello el trabajo se encuentra organizado de la siguiente manera. En la Sección 2 se hace una breve mención de los enfoques empleados en el muestreo con el fin de situar en perspectiva el uso de modelos de regresión en poblaciones finitas. Sobre esta base, en la Sección 3 se construyen las expresiones para los estimadores puntuales y sus varianzas estimadas. En la sección 4 se describen las poblaciones y variables por analizar, así como la forma en que se efectuaron las simulaciones[1]. Por último, en la Sección 5 se muestran y analizan los resultados.

[1]Es necesario efectuar simulaciones, ya que en el presente caso, la comparación analítica de los errores cuadráticos medios estimados para diversos diseños muestrales complejos es difícil de analizar.

2. Enfoques empleados en el muestreo de poblaciones finitas

Con el fin de situar en perspectiva el análisis de regresión en encuestas complejas, es necesario describir primero dos esquemas generales de selección de muestras complejas empleados comúnmente, esquemas A y B, así como una aproxmación a la varianza empleada por paquetes estadísticos, esquema C. Dichos esquemas son los siguientes, véase *Des Raj*[6] y *Särndal et al*[24]:

- **Esquema A**: se tienen H estratos y varias etapas de selección dentro de estratos. Las unidades primarias de muestreo, *UPM*, se seleccionan con muestreo sin reemplazo. Dentro de cada *UPM* se emplea cualquier forma de submuestreo, incluso muestreo sistemático. Un caso particular, consiste en seleccionar tanto las *UPM*, como las *USM*, unidades secundarias de muestreo, con muestreo aleatorio simple, *MAS*. Este último diseño muestral será referido en lo sucesivo como Esquema A y se denotará como *(MAS,MAS)* para el caso del muestreo por conglomerados en dos etapas o *(H,MAS,MAS)* para el muestreo aleatorio estratificado bietápico.

- **Esquema B**: dentro de estratos las *UPM* se seleccionan con reemplazo. Cada vez que se extrae una *UPM* se realiza el muestreo dentro de ella con cualquier forma de submuestreo en forma independiente[2]. Un caso particular, consiste en seleccionar las *UPM* con reemplazo y con probabilidad proporcional a alguna medida de tamaño y las *USM* con muestreo aleatorio simple, *MAS*. Este último diseño muestral será referido en lo sucesivo como Esquema B y se denotará como: *(ppt,MAS)* para el caso del muestreo con probabilidad proporcional a alguna medida de tamaño para las *UPM* y selección de *USM* por *MAS* y, *(H,ppt,MAS)* para el caso del muestreo con probabilidad proporcional a alguna medida de tamaño para las *UPM* por estrato y selección de *USM* por *MAS*.

- **Esquema C**: se usará este nombre para una *forma de cálculo, no de selección de muestras, empleada en la práctica* y que supone lo siguiente. Se tienen H estratos y varias etapas de selección dentro de estratos. Se supone que las unidades primarias de muestreo se seleccionan con muestreo aleatorio simple y que hay independencia e idéntica distribución entre las variables asociadas a las *UPM*. También se supone que, dentro de cada *UPM* se emplea cualquier forma de submuestreo. Este esquema es el que usan por omisión los

[2]En la práctica se emplea el muestreo con probabilidad proporcional al tamaño sin reemplazo, denotado como πpt, pero se usa la varianza estimada del *ppt*. Se usa esta aproximación porque no existe la estimación de la varianza al emplear muestreo sistemático, ya que algunas de las probabilidades de inclusión doble son cero, véase *Särndal et al.*[24], *Wolter*[31] y más recientemente *Tillè*[28]

paquetes, cuando no se cuenta con las probabilidades de inclusión de segundo orden. Obsérvese que en este tipo de cálculo se desprecia la contribución del submuestreo a las varianzas.

Así, con el esquema A ó B se extrae una muestra de tamaño fijo[3] con algún diseño muestral que incluye estratos, conglomerados, probabilidades proporcionales a alguna medida de tamaño y submuestreo con muestreo aleatorio simple y/o sistemático. Posteriormente, se lleva a cabo el análisis de regresión usando alguno de los diseños arriba mencionados, como se muestra en el siguiente diagrama.

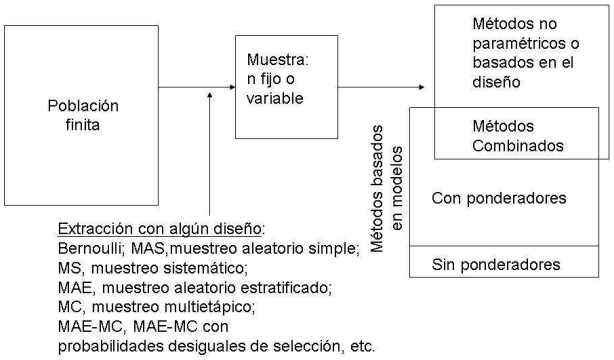

FIGURA 1. Esquemas usados en el muestreo

2.1. Regresión lineal. El análisis de regresión en una población finita con datos provenientes de encuestas complejas, puede abordarse en general de tres formas distintas:

- Modelo: En este enfoque se supone que las variables son una muestra de una superpoblación infinita y los diferentes aspectos del diseño muestral se modelan; por ejemplo, la estratificación puede incluirse como una variable explicativa. Esta forma de abordar el problema, fue propuesta por *Royall*, véase por ejemplo, *Valliant et al.*[30].
- Combinado: Se supone que los datos son una realización de una superpoblación infinita, pero se reconoce el mecanismo aleatorio de selección y se emplean las densidades del modelo de generación de los datos, así como la inducida por el diseño muestral.

[3]No se emplean esquemas con tamaño de muestra variable, ya que dificultaría el trabajo de campo; además podrían tenerse tamaños de muestra cero o uno en submuestras de conglomerados pequeños. Por otra parte, en diseños con tamaños de muestra variable se incrementa la varianza del estimador en cuestión.

- Diseño: los datos en la población, constituyen un conjunto finito de constantes y el elemento aleatorio es el mecanismo de selección probabilística. Se emplea un modelo de trabajo como un mecanismo para ayudar a estimar los parámetros del modelo de regresión en la población finita. Este es el enfoque que se empleará en el presente trabajo.

El uso de modelos de regresión lineal desde el punto de vista clásico, supone que el término de error del modelo sigue una distribución de probabilidades, siendo la distribución normal la más usada. También supone que la esperanza de los errores es cero, que dichos errores no están autocorrelacionados en serie y que tienen varianza constante, véase *Lohr* [17].

Por otra parte, en una población finita, en la que la inferencia se realiza bajo el enfoque del diseño asistido por el modelo, este último se trata solo como una vía para estimar los parámetros de interés, como los coeficientes de regresión que se obtendrían si se ajustare un modelo a todos los datos de la población finita. No se está suponiendo que los datos de la población finita fueron generados por el modelo en cuestión, sino que el modelo es empleado como una forma de describir la idea que se tenga acerca de la relación de la variable respuesta con las variables explicativas para la población finita. En este sentido se dice que se trata de un modelo de trabajo. Por ejemplo, para el caso de los coeficientes de regresión, una vez que se tiene un modelo pueden calcularse los coeficientes para la población completa y, en este sentido, son características de una población finita por estimar al extraer una muestra compleja. Cabe señalar que no se estiman los coeficientes de regresión con la estructura de error arriba mencionada. De esta manera, se tiene para cada elemento de la población la siguiente información: $(y_k, \mathbf{x}_k)$, donde y_k y $\mathbf{x}_k$ son un escalar y el vector de información auxiliar de tamaño $p \times 1, p \in N$ respectivamente.

Observación 1. *El concepto de modelo asistido se refiere a un vehículo para encontrar estimadores eficientes y constituye una ayuda para que el analista utilice el conocimiento y experiencia que tiene acerca de la relación, en la población finita, entre una variable de estudio y las variables auxiliares. En caso de que se tenga una relación fuerte en la población finita entre la variable de estudio y las auxiliares, el enfoque del modelo asistido producirá un estimador eficiente del total de una población finita.*

2.2. Regresión logística. Por otra parte, para la estimación de los coeficientes de un modelo de regresión logística en una población finita, bajo el enfoque del diseño asistido por un modelo, se postula una verosimilitud para el total de datos de la población y los coeficientes de regresión que resultan de maximizar dicha verosimilitud, son las características por estimar con los

datos de una muestra compleja [4].

- Existen varias formas de estimar los coeficientes, siendo las más usadas las siguientes:

    1. En el caso de datos extraídos con una muestra compleja y bajo el supuesto de independencia para la variable respuesta, se postula una verosimilitud que incluya los pesos muestrales y se maximiza. En *Hosmer y Lemeshow*[11] se encuentra descrito este caso, el cual es empleado por los paquetes comerciales en el evento de contar solo con los factores de expansión, véase *Stata*[25] y *Sudaan* [26]. La estimación del error cuadrático medio se obtiene de la aproximación de Taylor de segundo orden a la verosimilitud, una vez que se satisfaga algún criterio de convergencia como: cambios en la verosimilitud o en los coeficientes de regresión estimados. La estimación se efectúa dependiendo del esquema muestral empleado y la información que se tenga. Así, se empleará el *esquema A* si se conocen las probabilidades de inclusión de primer y segundo orden. Si se usó el esquema πpt y se conocen dichas probabilidades para la selección de las *UPM* se emplea el *esquema B*. En otro caso, el error cuadrático medio se estima con el *esquema C*, es decir, asociando a las *UPM* variables aleatorias independendientes e idénticamente distribuidas al interior de cada estrato.

    2. Otra forma de diseño asistido por un modelo es el de emplear un modelo de trabajo con correlación distinta de cero dentro de conglomerados y cero entre conglomerados, ver *McCulloch*[19] y *Sudaan*[26]. En este enfoque, no se postula una forma funcional para la verosimilitud y se emplean las ecuaciones de estimación no lineales en los coeficientes de regresión, que tienen los mismos segundos momentos que la verosimilitud. Esto es una versión adaptada a poblaciones finitas y diseños complejos de lo que se conoce como modelo de trabajo con correlación intercambiable, que es un caso particular de ecuaciones generalizadas de estimación, *Liang y Zeger*[15]. Esto conduce a estimaciones puntualespara los coeficientes de regresión y sus varianzas estimadas distintas que las obtenidas bajo el supuesto de independencia.

[4]Véase *Agresti*[1] para el caso en el que se suponga independencia entre valores de la variable respuesta y se ignore el diseño muestral.

3. Estimadores puntuales y sus varianzas aproximadas estimadas

3.1. Estrategia para la estimación de características en una población finita. Antes de mostrar las expresiones correspondientes a los estimadores, es necesario mencionar algunos puntos relativos a la forma de efectuar estimaciones de cantidades en una población finita bajo el esquema del muestreo probabilístico. En una población finita con N elementos denotada como U, los elementos que la conforman no son variables aleatorias sino números reales y se denota como: $U = \{u_1, u_2, ..., u_N\}$. Cada unidad se identifica sin ambigüedad con una etiqueta y es común denotar a la población con las etiquetas: $U = \{1, ..., k, ...N\}$. La variable de interés se denota como y, que adquiere el valor y_k para la unidad k en la población. El objetivo es estimar una función t que depende de las y_k, $t = t(y_1, ..., y_k, ..., y_N)$.

Las cantidades que se desean estimar comúnmente con base en una muestra son: totales, promedios, proporciones o total de elementos que pertenecen a una clase de interés. Por ejemplo, considérese a las personas en el estado de Chiapas como la población U, entonces algunas características por estimar son: el total de ingreso, el promedio de años que asistió una persona a la escuela y el número de personas que no hablan español.

Un total poblacional se escribe de la siguiente manera: $y_U = \sum_{k=1}^{N} y_k$, en tanto que un promedio se expresa como: $\overline{y}_U = \sum_{k=1}^{N} y_k/N$. La estimación se lleva a cabo con base en una muestra probabilística M de tamaño n menor que N, en la que se conoce la probabilidad de inclusión en muestra de cada elemento k de la población, la cual se denota como π_k. La función de densidad inducida por el diseño muestral es discreta y multivariada y es el elemento aleatorio en el muestreo probabilístico. La muestra se denota por un vector columna, $m = (I_1, ..., I_k, ..., I_N)' \in \{0, 1\}^N$. En este caso I_k adquiere el valor 1 si el k-ésimo elemento pertenece a la muestra y 0 en otro caso.

Así, el estimador denominado Horvitz-Thompson, denotado como HT, ver *Horvitz y Thompson*[12], para estimar el total de alguna característica de la población finita es: $\hat{y}_U = \sum_{k=1}^{n} y_k/\pi_k = \sum_{k=1}^{N} I_k y_k/\pi_k$.

Obsérvese que el lado derecho del estimador de HT para un total, depende de la variable indicadora de elementos en muestra y corre sobre todos los elementos de la población. La expresión de un estimador como un total permite aplicar las propiedades de linealidad de los operadores de esperanza y varianza, junto con los valores de las probabilidades de inclusión. El resultado empleado, es el corolario del siguiente teorema, ver capítulo 5 de *Mood et al.*[21].

Teorema 1. *Sean $X_1, ..., X_n$ y $Y_1, ..., Y_n$ variables aleatorias y $a_1, ..., a_n$ y $b_1, ..., b_n$ constantes, entonces:*

$$Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j Cov\left(X_i, Y_j\right) \tag{3.1}$$

Corolario 1. *Sean $X_1, ..., X_n$ variables aleatorias y $a_1, ..., a_n$ constantes , entonces:*

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov\left(X_i, X_j\right) \tag{3.2}$$

$$= \sum_{i=1}^{n} a_i^2 V\left(X_i\right) + 2\sum_{i=1}^{n-1}\sum_{j>i}^{n} a_i a_j Cov\left(X_i, Xj\right) \tag{3.3}$$

En la práctica se trabaja con estimadores para totales y en caso de que se tenga un estimador no lineal, como una razón, se linealiza con una aproximación de Taylor de primer orden y se expresa como un total. Para encontrar la expresión de la varianza se aplica el corolario anterior. A continuación se ejemplifican algunos casos.

Ejemplo 1. *En el muestreo aleatorio simple, $\pi_k = n/N$, por lo que el estimador del total adquiere la forma $\hat{y} = \sum_{k=1}^{n} N y_k/n$. A N/n se le conoce como el factor de expansión y es la cantidad que se entrega en las bases de datos para cada elemento en la muestra. Estos factores tienen la propiedad de que $\sum_{k=1}^{n} N/n = N$. En la práctica, los factores de expansión no son iguales para todos los elementos de la muestra debido al empleo de esquemas de selección complejos.*

Por otra parte, para obtener la expresión de la varianza del estimador se aplica la varianza al estimador $\hat{y}_k$:

$$V\left(\hat{y}\right) = V\left(\sum_{k=1}^{N} \frac{I_k y_k}{\pi_k}\right) \tag{3.4}$$

$$= \sum_{k=1}^{N}\sum_{l=1}^{N} Cov\left(I_k, I_l\right)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{3.5}$$

$$= N^2\left(1-f\right)\frac{S_U^2}{n} \tag{3.6}$$

En la expresión anterior: $Cov\left(I_k, I_l\right) = \pi_{kl} - \pi_k\pi_l = -f(1-f)/(N-1)$. A $\pi_{kl} = -f(n-1)/(N-1)$ se le denomina probabilidad de inclusión de k y l en muestra, conocida como probabilidad de inclusión de segundo orden, en tanto que a $f = n/N$ se le denomina la corrección por población finita y

$S_U^2 = \sum_{k=1}^{N}(y_k - \overline{y}_U)^2/(N-1)$ *es la varianza poblacional entre elementos.*

Para el estimador de la varianza se utiliza lo siguiente:

$$\widehat{V(\hat{y})} = \hat{V}\left(\sum_{k=1}^{N}\frac{I_k y_k}{\pi_k}\right) \tag{3.7}$$

$$= \sum_{k=1}^{N}\sum_{l=1}^{N}\frac{Cov\left(I_k, I_l\right)}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{3.8}$$

$$= N^2\left(1-f\right)\frac{S^2}{n} \tag{3.9}$$

A $S^2 = \sum_{k=1}^{n}(y_k - \hat{\overline{y}}_m)^2/(n-1)$ *se le conoce como la varianza muestral entre elementos, con* $\hat{\overline{y}}_m = \sum_{k=1}^{n} y_k/n$.

En este ejemplo se observa que la construcción de estimadores para totales facilita la obtención de fórmulas usando la linealidad y aditividad de la esperanza.

Ejemplo 2. *En el caso de que se extraiga una muestra aleatoria simple de tamaño n de una población con N elementos y se tenga una variable auxiliar positiva* x_k *con una alta correlación positiva con la variable de interés* y_k *y además la relación de* y_k *con* x_k *sea una recta que pasa por el origen, entonces es adecuado el uso de un estimador de razón* $\hat{r} = \hat{y}/\hat{x}$, *con* $\hat{\overline{x}} > 0$. *Este es un estimador no lineal, ya que es el cociente de dos estimadores, por lo que se usa una aproximación de Taylor de primer orden, se calcula el error cuadrático medio y su estimación. En el capítulo 6 de Cochran* [5], *se encuentra detallado el procedimiento de aproximación, la determinación del error cuadrático medio, su estimación, así como las condiciones bajo las cuales el estimador de razón de un total es preferible a la expansión simple bajo muestreo aleatorio simple y las condiciones en las que el sesgo del estimador es despreciable. La aproximación devuelve la estimación del error cuadrático medio (varianza si el sesgo es despreciable):*

$$\widehat{ECM(\hat{r})} \approx \hat{V}(\hat{r}) = \hat{V}\left(\hat{y} - \hat{r}\hat{x}\right)/\hat{x}^2 = \hat{V}(\hat{g})/\hat{x}^2.$$

En esta expresión, $g_k = y_k - \hat{r}x_k$, *es una variable que estima un total* g_U *en la población finita y su varianza estimada es fácil de calcular según lo visto en el ejemplo anterior. Nótese que dicha variable tiene media muestral cero. Debido a que la muestra se extrajo con MAS, se aplica lo visto en el ejemplo anterior a la variable g para la varianza y su estimación obteniéndose*

lo siguiente[5]*:*

$$\hat{V}(\hat{r}) = N^2 \frac{(1-f)}{\hat{x}^2 n} \frac{\sum_{k=1}^{n} (y_k - \hat{r}x_k)^2}{n-1} \tag{3.10}$$

3.2. Estimadores de varianza para los esquemas A, B o C. En la sección 2 se describieron dos esquemas de muestreo y un esquema de cálculo, usados comúnmente en la práctica. Para cada uno de ellos se tiene una expresión de varianza estimada que refleja el esquema en cuestión. A continuación se introduce la notación que se empleará, así como las expresiones empleadas para la varianza y varianza estimada del estimador *HT* de un total.

Supóngase que se tiene una población finita U dividida en H estratos, en cada estrato se tienen N_h conglomerados o *UPM* y cada conglomerado tiene M_{hi} elementos o *USM*. También suponga que se tiene una medida de tamaño ν_h para cada *UPM* en la población. Se usará la siguiente notación:

- H=número de estratos en la población,
- N_h=número de *UPM* en población, en el $h-$ésimo estrato,
- n_h=número de *UPM* en muestra, en el $h-$ésimo estrato,
- M_{hi}=número de *USM*, unidad secundaria de muestreo, en cada *UPM* en población, en el $h-$ésimo estrato,
- m_{hi}=número de USM en cada *UPM* en muestra, del $h-$ésimo estrato,
- π_{hi}=probabilidad de inclusión en muestra de la $i-$ésima *UPM* en muestra, en el $h-$ésimo estrato,
- π_{hij}=probabilidad de inclusión en muestra de la $i-$ésima *UPM* y la $j-$ésima *UPM* en muestra en el $h-$ésimo estrato, con $i \neq j$,
- $\pi_{hk|i}$=probabilidad de inclusión en muestra de la $k-$ésima *USM* en muestra, dado que la $i-$ésima *UPM* fue seleccionada en muestra, en el $h-$ésimo estrato,

[5]En *Méndez y Romero*[20], así como en *Thompson*[27], se encuentra detallado el procedimiento para obtener expresiones de varianzas estimadas de estimadores de totales para los principales diseños complejos empleados en la práctica.

- $\pi_{hkl|i}$=probabilidad de inclusión en muestra de la $k-$ésima *USM* y la $l-$ésima *USM*, dado que la $i-$ésima *UPM* fue seleccionada en muestra en el $h-$ésimo estrato, con $i \neq j$,

- I_{hi}=variable indicadora que adquiere el valor 1 si la i-ésima *UPM* del $h-$ésimo estrato pertenece a la muestra y 0 en otro caso,

- $Cov\left(I_{hi}, I_{hj}\right) = \pi_{hij} - \pi_{hi}\pi_{hj}$=covarianza entre la $i-$ésima y j-ésima *UPM*, en el $h-$ésimo estrato,

- $Cov\left(I_{hk}, I_{hl}|i\right) = \pi_{hkl|i} - \pi_{hk|i}\pi_{hl|i}$=covarianza entre los $k-$ésimo y l-ésimo elementos de la $i-$ésima *UPM* en muestra, en el $h-$ésimo estrato,

- p_{hi}=medida relativa de tamaño que se calcula como $\nu_{hi} / \sum_{i=1}^{N_h} \nu_{hi}$,

- ν_{hi}=medida de tamaño de la $i-$ésima *UPM* en el $h-$ésimo estrato. Esta última variable puede ser, por ejemplo, el número de personas en la $i-$ésima *UPM*,

- $w_{hij} = 1/\pi_{hij}$= factor de expansión del $j-$ésimo elemento en la $i-$ésima *UPM*, en el $h-$ésimo estrato,

- y_{hij}=valor de la variable respuesta para el $j-$ésim elemento en la $i-$ésima *UPM*, en el $h-$ésimo estrato,

- $\mathbf{x}_{hij} = \left(x_{1,hij}, \ldots, x_{p,hij}\right)'$ es el vector de información de auxiliar de tamaño $p \times 1$, en el cual $x_{1,hij} = 1$.

- **Esquema A**: la varianza del estimador de *Horvitz-Thompson* para un total y_U, se escribe como la suma de dos componentes,

$$Var\left(\hat{y}\right) = V_{UPM} + V_{USM} \tag{3.11}$$

$$V_{UPM} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} Cov\left(I_{hi}, I_{hj}\right) \frac{y_{hi}}{\pi_{hi}} \frac{y_{hj}}{\pi_{hj}} \tag{3.12}$$

$$V_{USM} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{V_{hi}}{\pi_{hi}} \tag{3.13}$$

$$V_{hi} = \sum_{k=1}^{M_{hi}} \sum_{l=1}^{M_{hi}} Cov\left(I_{hk}, I_{hl}|i\right) \frac{y_{hk}}{\pi_{hk|i}} \frac{y_{hl}}{\pi_{hl|i}} \tag{3.14}$$

El estimador insesgado del primer componente es:

$$\begin{aligned}(3.15)\quad \hat{V}_{UPM} &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_h}\frac{Cov\left(I_{hi},I_{hj}\right)}{\pi_{hij}}\sum_{k=1}^{m_{hi}}\frac{y_{hik}}{\pi_{hi}\pi_{hk|i}}\sum_{l=1}^{m_{hi}}\frac{y_{hil}}{\pi_{hi}\pi_{hl|i}} \\ &- \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hi}}\left(\frac{1}{\pi_{hi}}-1\right)\hat{V}_{hi}\end{aligned}$$

con:

$$(3.16)\qquad \hat{V}_{hi}=\sum_{k=1}^{m_{hi}}\sum_{l=1}^{m_{hi}}\frac{Cov\left(I_{hk},I_{hl}|i\right)}{\pi_{hkl|i}}\frac{y_{hk}}{\pi_{hk|i}}\frac{y_{hl}}{\pi_{hl|i}}$$

- **Esquema B**: La varianza estimada del estimador de un total y_U es,

(3.17)

$$\widehat{Var}\left(\hat{y}_{ppt}\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left\{\sum_{i=1}^{n_h}\frac{1}{\left(n_h p_{hi}\right)^2}\left(\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{\pi_{hj|i}}\right)^2-\frac{1}{n_h}\left(\sum_{i=1}^{n_h}\frac{1}{n_h p_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{\pi_{hj|i}}\right)^2\right\}$$

- **Esquema C**: La varianza estimada del estimador de un total y_U, suponiendo independencia e idéntica distribución de las variables y_{hij} es,

$$(3.18)\qquad \widehat{Var}\left(\hat{y}_{iid}\right)=\hat{V}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}y_{hi}\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\overline{y}_h\right)^2$$

En la ecuación anterior, $\overline{y}_h=\sum_{h=1}^{H}y_h/H$ y $y_{hi}=\sum_{j=1}^{m_{hi}}w_{hij}y_{hij}$.

3.3. Regresión lineal.

3.3.1. Estimación del vector poblacional de coeficientes beta, esquemas A, B y C. En este caso, y_{hij} es la variable dependiente, $\mathbf{x}_{hij}=\left(x_{1,hij},\ldots,x_{p,hij}\right)'$ es el vector de información, de tamaño $p\times 1$, para cada elemento de la población finita U. Además, se supone que las y_{hij} no están correlacionadas entre sí y que el modelo lineal de trabajo es:

$$E_\xi\left(y_{hij}\right) = \mathbf{x}_{hij}^{'}\mathbf{B} \tag{3.19}$$

$$V_\xi\left(y_{hij}\right) = \sigma^2, \quad \sigma^2 > 0 \tag{3.20}$$

En estas ecuaciones, ξ denota el cálculo de esperanzas bajo la función de densidad del modelo lineal de trabajo, aunque en el enfoque del diseño asistido no es necesario conocer la forma funcional, ya que no se emplea. Aquí $\mathbf{B}$ es de orden $p \times 1$. El valor poblacional por estimar del vector de coeficientes de regresión es, siempre que la inversa exista:

$$\mathbf{B} = \left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} \mathbf{x}_{hij}\,\mathbf{x}'_{hij}/\sigma^2\right)^{-1} \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} \mathbf{x}_{hij}\,y_{hij}/\sigma^2 \tag{3.21}$$

Para cualesquiera de los esquemas A, B y C, el vector de estimación puntual de los coeficientes beta poblacionales tiene la forma, si la inversa existe:

$$\hat{\mathbf{B}} = \left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \mathbf{x}_{hij}\,\mathbf{x}'_{hij}/\pi_{hij}\sigma^2\right)^{-1} \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \mathbf{x}_{hij}\,y_{hij}/\pi_{hij}\sigma^2 \tag{3.22}$$

Observación 2. *En la ecuación (3.22), σ^2 se cancela, esto se debe al modelo de trabajo postulado en (3.20), en el que se supone varianza constante para todos los elementos de la población. En el evento de postular un modelo de trabajo con varianza distinta por conglomerado, ésta tendría que estimarse. Es necesario mencionar que el estimador de los coeficientes de regresión es sesgado, a diferencia del caso clásico en el que las variables independientes son constantes. Esto sucede porque el elemento aleatorio I_{hij} está asociado a cada elemento en la muestra como se aprecia a continuación:*

$$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \mathbf{x}_{hij}\,\mathbf{x}'_{hij}/\pi_{hij}\sigma^2 = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} I_{hij}\mathbf{x}_{hij}\,\mathbf{x}'_{hij}/\pi_{hij}\sigma^2 \tag{3.23}$$

Por lo anterior, en general se tiene que:

(3.24)

$$E\left[\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\mathbf{x}_{hij}\,\mathbf{x}'_{hij}/\pi_{hij}\sigma^2\right)^{-1}\right] \neq \left[E\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\mathbf{x}_{hij}\,\mathbf{x}'_{hij}/\pi_{hij}\sigma^2\right)\right]^{-1}$$

La estimación de la varianza aproximada linealmente por Taylor se calcula para cualesquiera de los esquemas como, *Särndal et al*[24], pág. 194:

(3.25)

$$\widehat{ECM}(\hat{\mathbf{B}}) = \left[\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\frac{\mathbf{x}_{hij}\mathbf{x}'_{hij}}{\pi_{hij}}\right]^{-1}\begin{bmatrix}\hat{v}_{1,1} & \hat{v}_{1,2} & .. & \hat{v}_{1,p}\\ \hat{v}_{2,1} & \hat{v}_{2,2} & .. & \hat{v}_{2,p}\\ .. & .. & .. & ..\\ \hat{v}_{p,1} & \hat{v}_{p,2} & .. & \hat{v}_{p,p}\end{bmatrix}\left[\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\frac{\mathbf{x}_{hij}\mathbf{x}'_{hij}}{\pi_{hij}}\right]^{-1}$$

La diferencia entre los esquemas A, B y C, se tiene en los estimadores $\hat{v}_{l,k}$, con $l,k = 1,...,p$, del interior de la matriz en (3.25), ya que son los que tienen la expresión de la varianza aproximada según el esquema que se esté empleando. Obsérvese que los subíndices l y k se refieren a las variables independientes del modelo (3.20) etiquetadas con dichos índices. Para efectuar los cálculos correspondientes a cada esquema, a la $i-$ésima *UPM* muestral en el $h-$ésimo estrato, se le asocia una nueva variable $d_{hi} = \sum_{j=1}^{m_{hi}} w_{hij}\, d_{hij}$, donde $d_{hij} = y_{hij} - \mathbf{x_{hij}}'\hat{\mathbf{B}}$. De esta manera, se está estimando la varianza de un total, d_{hi} usando: (3.31) para el esquema A, (3.40) para el esquema B y (3.42) para el esquema C.

- **Esquema A**: cálculo de $\hat{v}_{l,k}$ para $l,k = 1,...,p$ considerando el diseño complejo y el hecho de que la variable $d_{hij} = y_{hij} - \mathbf{x_{hij}}'\hat{\mathbf{B}}$ es un estimador del tipo Horvitz-Thompson de un total:

(3.26)
$$\hat{v}_{l,k} = \hat{v}_{l,k\ UPM} + \hat{v}_{l,k\ USM}$$

donde:

$$
\begin{aligned}
\hat{v}_{l,k\,UPM} &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_h}\frac{Cov\left(I_{hi},I_{hj}\right)}{\pi_{hij}}\frac{\hat{d}_{hi\pi}}{\pi_{hi}}\frac{\hat{d}_{hj\pi}}{\pi_{hj}} \\
&- \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hi}}\left(\frac{1}{\pi_{hi}}-1\right)\hat{V}_{hi} \\
\hat{V}_{hi} &= \sum_{r=1}^{m_{hi}}\sum_{s=1}^{m_{hi}}\frac{Cov\left(I_{hr},I_{hs}|i\right)}{\pi_{hrs|i}}\frac{d_{hr|i}}{\pi_{hr|i}}\frac{d_{hs|i}}{\pi_{hs|i}} \\
\hat{v}_{l,k\,USM} &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{\hat{V}_{hi}}{\pi_i^2} \\
\frac{Cov\left(I_{hr},I_{hs}|i\right)}{\pi_{hrs|i}} &= \frac{\pi_{hrs|i}-\pi_{hr|i}\pi_{hs|i}}{\pi_{hrs|i}}
\end{aligned}
$$

La última expresión es la probabilidad de inclusión de los elementos r y s en muestra en el $h-$ésimo estrato, dado que la $i-$ésima UPM fue seleccionada en muestra. En las ecuaciones (3.10) y (3.11): $\pi_{hr|i}$ es la probabilidad de inclusión del $r-$ésimo elemento en muestra del $h-$ésimo estrato, dado que la $i-$ésima UPM fue seleccionada en muestra, $\pi_{hrs|i}$ es la probabilidad de inclusión del $r-$ésimo y $s-$ésimo elemento en muestra en el $h-$ésimo estrato, dado que la $i-$ésima UPM fue seleccionada en muestra y π_{hi} es probabilidad de inclusión del $i-$ésimo UPM en muestra del $h-$-esimo estrato.

- **Esquema B:** estimación de $\hat{v}_{l,k}$ para $l,k=1,...,p+1$ usando ppt.

(3.27)
$$
\hat{v}_{l,k}=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left\{\sum_{i=1}^{n_h}\frac{1}{(n_h p_{hi})^2}\left(\sum_{j=1}^{m_{hi}}\frac{d_{hij}}{\pi_{hj|i}}\right)^2-\frac{1}{n_h}\left(\sum_{i=1}^{n_h}\frac{1}{n_h p_{hi}}\sum_{j=1}^{m_{hi}}\frac{d_{hij}}{\pi_{hj|i}}\right)^2\right\}
$$

En esta expresión, la variable ν_{hi} puede ser, por ejemplo, el número de personas en la $i-$ésima UPM, del $h-$ésimo estrato.

- **Esquema C:** estimación de $\hat{v}_{l,k}$ para $i,j=1,...,p$ suponiendo independencia e idéntica distribución de variables y_{hij}:

(3.28)
$$
\hat{v}_{l,k}=\hat{V}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}d_{hi}\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(d_{hi}-\overline{d}_h\right)^2
$$

En la ecuación anterior, $\overline{d}_h = \sum_{h=1}^{H} d_h/H$

3.3.2. Estimadores esquema A. Estimador puntual del total usando el estimador de regresión generalizado.

$$\hat{y}_{regr} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hi}}\sum_{j=1}^{m_{hi}}\mathbf{x}_{hij}^{'}\hat{\mathbf{B}} + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hi}}\sum_{j=1}^{m_{hi}}\frac{e_{hij}}{\pi_{hj|i}} \tag{3.29}$$

En la ecuación anterior:

- $\hat{B}$ =véase ecuación (3.22),
- $e_{hij} = y_{hij} - \mathbf{x}_{hij}^{'}\hat{\mathbf{B}}$.

Al usar MAS en la primera y en la segunda etapa de selección se tiene que $\pi_{hi} = n_h/N_h$ y $\pi_{hj|i} = (n_h/N_h)\,(m_{hi}/M_{hi})$. Estos valores se insertan en (3.29), obteniéndose la expresión siguiente:

$$\hat{y}_{regr} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{N_h}{n_h}\sum_{j=1}^{m_{hi}}\mathbf{x}_{hij}^{'}\hat{\mathbf{B}} + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{N_h}{n_h}\sum_{j=1}^{m_{hi}}\frac{N_h M_{hi}}{n_h m_{hi}}e_{hij} \tag{3.30}$$

Observación 3. *Es necesario mencionar que la estimación puntual de un total usando el estimador de regresión generalizado en (3.29), requiere de los totales poblacionales de las variables auxiliares para cada UPM, así como de las probabilidades de inclusión de unidades de primera y segunda etapa.*

Observación 4. *El estimador en (3.29) es el adecuado cuando se cuenta con totales poblacionales de las variables auxiliares para cada UPM en muestra; sin embargo, en el caso de que se tuviere información de las variables auxiliares para toda la población, se emplearía otra expresión del estimador de regresión generalizado que difiere en el primer término de la derecha en (3.29). El primer término correría sobre todos los conglomerados de la población en lugar de efectuar estimaciones de totales de conglomerados. Por otra parte, cuando se desea estimar un total usando el estimador de regresión con datos a nivel elemento de la población, solo se requiere conocer el*

valor de las variables auxiliares en la muestra, así como el total poblacional de cada variable auxiliar.

Estimador puntual del error cuadrático medio del total usando el estimador de regresión generalizado.

$$\widehat{ECM}(\hat{y}_{regr}) \approx \hat{V}_{UPM} + \hat{V}_{USM} \tag{3.31}$$

Observación 5. *La expresión anterior es una aproximación al error cuadrático medio del estimador, ya que la segunda componente depende de las cantidades g_{hik} definidas en (3.34) conocidas como calibradores. Dichos calibradores dependen de las $\mathbf{x}_{hij}$ muestrales y proporcionan estimaciones sin error del total de cada variable auxiliar en la población. Los calibradores adquieren valores cercanos al 1 siempre que las variables auxiliares $\mathbf{x}_{hij}$ estimen con un error pequeño el total de cada UPM en muestra. Para obtener la expresión (3.31), se emplea $g_{hik} \approx 1$, con lo cual ya puede tratarse el estimador con los resultados del muestreo polietápico, véase* Cochran [5] .

En la ecuación (3.31), $\hat{V}_{UPM}$ y $\hat{V}_{USM}$ se refieren a componentes de estimación del error cuadrático medio por etapa de selección de conglomerados y tienen las expresiones siguientes, *Särndal et al*[24], pág. 326:

$$\begin{aligned}\hat{V}_{UPM} &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_h}\frac{Cov\left(I_{hi},I_{hj}\right)}{\pi_{hij}}\frac{\hat{y}_{hi\pi}}{\pi_{hi}}\frac{\hat{y}_{hj\pi}}{\pi_{hj}} \\ &- \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hj}}\left(\frac{1}{\pi_{hj}}-1\right)\sum_{j=1}^{m_{hi}}\sum_{k=1}^{m_{hi}}\frac{Cov\left(I_{hj},I_{hk}|i\right)}{\pi_{hjk|i}}\frac{y_{hj\pi}}{\pi_{hj|i}}\frac{y_{hk\pi}}{\pi_{hk|i}}\end{aligned} \tag{3.32}$$

$$\hat{V}_{USM} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hi}^2}\sum_{j=1}^{m_{hi}}\sum_{k=1}^{m_{hi}}\frac{Cov\left(I_{hj},I_{hk}|i\right)}{\pi_{hjk|i}}\frac{g_{hj\pi}e_{hj}}{\pi_{hj|i}}\frac{g_{hk\pi}e_{hk}}{\pi_{hk|i}} \tag{3.33}$$

En la última expresión, g_{hij} se refiere a los calibradores para el vector de información auxiliar $\mathbf{x}_{hi}$ y está dado por:

$$g_{hik} = 1 + \left[\sum_{i=1}^{n_h} \frac{(\mathbf{x}_{hUi} - \hat{\mathbf{x}}_{hi\pi})}{\pi_{hi}}\right]' \hat{\mathbf{T}}^{-1} \frac{\mathbf{x}_{hik}}{\sigma^2} \tag{3.34}$$

donde: $\hat{\mathbf{T}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \mathbf{x}_{hij}\, \mathbf{x}_{hij}^t / \sigma^2 \pi_{hij}$. Esto por supuesto, siempre que la inversa exista, en caso contrario puede usarse una inversa generalizada.

De manera similar a lo realizado en (3.29) para obtener una expresión con el diseño MAS, se insertan los valores de las probabilidades de selección π_{hi} y π_{hij} en las ecuaciones (3.32) y (3.33). En este caso, la probabilidad de inclusión de segundo orden es $\pi_{hij} = (n_h\,(n_h - 1))\,/\,(N_h\,(N_h - 1))$.

Así, se tiene que:

$$\begin{aligned}
\frac{Cov\,(I_{hi}, I_{hj})}{\pi_{hij}} &= -\frac{f_h\,(1 - f_h)}{(N_h - 1)} \\
f_h &= \frac{n_h}{N_h} \\
\pi_{hj|i} &= \frac{m_{hi}}{M_{hi}} \\
\pi_{hjk|i} &= \frac{m_{hi}\,(m_{hi} - 1)}{M_{hi}\,(M_{hi} - 1)} \\
\frac{Cov\,(I_{hj}, I_{hk}|i)}{\pi_{hjk|i}} &= -\frac{f_{hi}\,(1 - f_{hi})}{(M_{hi} - 1)} \\
f_{hi} &= \frac{m_{hi}}{M_{hi}}
\end{aligned} \tag{3.35}$$

Con estos valores para las probabilidades de inclusión, la ecuación (3.32) queda como:

$$\begin{aligned}
\hat{V}_{UPM} &= \sum_{h=1}^{H} N_h^2\,(1 - f_h)\,\frac{\hat{s_1^2}}{n_{I_h}} \\
&- \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{N_h}{n_h}\left(\frac{N_h}{n_h} - 1\right) \sum_{j=1}^{m_{hi}} \sum_{k=1}^{m_{hi}} \frac{Cov\,(I_{hj}, I_{hk}|i)}{\pi_{hjk|i}} \frac{y_{hj\pi}}{\pi_{hj|i}} \frac{y_{hk\pi}}{\pi_{hk|i}}
\end{aligned} \tag{3.36}$$

(3.37)

donde,

$$\sum_{j=1}^{m_{hi}}\sum_{k=1}^{m_{hi}}\frac{Cov\left(I_{hj},I_{hk}|i\right)}{\pi_{hjk|i}}\frac{y_{hj\pi}}{\pi_{hj|i}}\frac{y_{hk\pi}}{\pi_{hk|i}}=M_{hi}^{2}\left(1-f_{hi}\right)\frac{\hat{s_2}^2}{m_{hi}} \tag{3.38}$$

con,

$$\hat{s_2}^2=\frac{\sum_{k=1}^{m_{hi}}y_{hik}^2-m_{hi}\left(\sum_{k=1}^{m_{hi}}y_{hik}/m_{hi}\right)^2}{m_{hi}-1}$$

Observación 6. *Existe una aproximación a la varianza del esquema A que solo afecta a la componente de las unidades secundarias de muestreo y se lleva a cabo haciendo $g_{hik}=1$ en (3.33). La idea detrás de esto es que si las variables auxiliares en el calibrador estiman adecuadamente el total correspondiente, entonces el calibrador adquirirá valores cercanos a uno. Esta fórmula se aplicará para compararla con los otros esquemas en las simulaciones.*

3.3.3. Estimadores esquema B. Esquema ppt: estimador puntual del total usando el estimador ppt.

$$\hat{y}_{ppt}=\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{y}_{hi\pi}}{p_{hi}}=\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{p_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{\pi_{hj|i}} \tag{3.39}$$

Esquema ppt: estimador puntual del error cuadrático medio del total con esquema de selección ppt.

(3.40)

$$\widehat{ECM}_{ppt}=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left\{\sum_{i=1}^{n_h}\frac{1}{(n_hp_{hi})^2}\left(\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{\pi_{hj|i}}\right)^2-\frac{1}{n_h}\left(\sum_{i=1}^{n_h}\frac{1}{n_hp_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{\pi_{hj|i}}\right)^2\right\}$$

3.3.4. Estimadores esquema C. Este estimador no puede emplearse en caso de no contar con la información auxiliar y las probabilidades mencionadas, por lo cual, en el esquema C se emplea la estimación del total usando los factores de expansión. Estimador puntual del total usando el estimador de expansión del total, con $\omega_{hij}=1/\pi_{hij}$.

$$\hat{y}_{paq} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \frac{y_{hij}}{\pi_{hij}} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \omega_{hij} y_{hij} = \sum_{h=1}^{H}\sum_{i=1}^{n_h} d_{hi} \tag{3.41}$$

En esta expresión, $d_{hi} = \sum_{j=1}^{m_{hi}} \omega_{hij} y_{hij}$.

Estimador puntual del error cuadrático medio del total.

$$\begin{aligned}
\widehat{ECM}\left(\hat{y}_{paq}\right) &= \hat{V}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} d_{hi}\right) \\
&= \sum_{h=1}^{H} \hat{V}\left(\sum_{i=1}^{n_h} d_{hi}\right) \\
&= \sum_{h=1}^{H} n_h \hat{V}\left(d_{hi}\right) \\
&= \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(d_{hi} - \overline{d}_h\right)^2 \qquad (3.42)
\end{aligned}$$

En la ecuación anterior, la varianza se intercambia con la suma de estratos porque la selección es independiente entre estratos, con lo que se induce por diseño una covarianza cero. Por otra parte, dentro de cada estrato y por el supuesto de independencia entre unidades e igual distribución, la covarianza es cero y se tiene la misma varianza.

3.4. Regresión logística. Esquemas A, B y C, suponiendo independencia para la variable y_{hij}.

La función de logverosimilitud para la totalidad de los elementos en la población finita es:

(3.43)
$$L(b_{hij}; x_{hij}) = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} \{y_{hij} ln(P(\mathbf{x}_{hij})) + (1 - y_{hij}) ln(1 - P(\mathbf{x}_{hij}))\}$$

En la ecuación anterior $P(\mathbf{x}_{hij}) = \exp(\mathbf{x}'_{hij}\mathbf{B})/(1 + \exp(\mathbf{x}'_{hij}\mathbf{B}))$, donde $\mathbf{x}_{hij}$ es un vector de variables independientes categóricas y/o continuas de

orden $p \times 1$ y $\mathbf{B} = (b_1, b_2, ..., b_p)'$ es un vector columna de orden $p \times 1$. Los valores poblacionales por estimar son el vector de coeficientes $\mathbf{B}$ y son aquellos que maximizan la logverosimilitud. Para efectuar la maximización se emplea un método iterativo, como el método de Newton-Raphson usando una aproximación de Taylor de segundo orden de la logverosimilitud y resolviendo para el vector de coeficientes $\mathbf{B}$. El criterio de convergencia se basa en cambios pequeños en la logverosimilitud y/o en cambios pequeños en los coeficientes $\mathbf{B}$.

$$\partial L(b_{hij}; x_{hij})/\partial B_j = \mathbf{x}_{hij} P(\mathbf{x}_{hij})(1 - P(\mathbf{x}_{hij})) \tag{3.44}$$

En este caso se emplea una función que aproxime a la verosimilitud en la población finita muestreada con una función que incorpora la muestra extraída y los pesos muestrales conocidos. La función de logverosimilitud aproximada por la muestra es:

(3.45)

$$l(b_{hij}; x_{hij}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \{w_{hij} y_{hij} ln(P(\mathbf{x}_{hij})) + w_{hij}(1 - y_{hij}) ln(1 - P(\mathbf{x}_{hij}))\}$$

Esta última ecuación se iguala a cero y se resuelve para el vector $\mathbf{B}$. La forma común de solucionarlo es obteniendo el gradiente e igualando a cero. Al hacer esto se obtiene lo que se conoce como ecuaciones de estimación muestrales[6]. En términos matriciales el sistema de ecuaciones de estimación muestral se escribe como: $\mathbf{X'W}(\mathbf{y} - \mathbf{P}) = \mathbf{0}$, donde $\mathbf{X}$ es la matriz de covariables de orden $n \times p$, $\mathbf{W}$ es una matriz diagonal de pesos de orden $n \times n$, $\mathbf{y}$ es un vector de orden $n \times 1$ de valores observados de la variable respuesta y $\mathbf{P}$ es un vector de orden $n \times 1$ de probabilidades logísticas. Este sistema puede representarse en términos de los pesos muestrales y la función de puntaje, denominada score en inglés, como:

$$\hat{G}(\hat{b_j}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} s^{(l)}(b_j; y_{hij}, x_{hij}) = 0 \tag{3.46}$$

En la ecuación anterior, el índice l corre de 1 a p y $s^{(l)}(b_j; y_{hij}, x_{hij}) = x_{hij}(y_{hij} - P(\mathbf{x}_{hij}))$ es la función de puntaje para la $l-$ésima variable independiente. Las derivadas parciales son:

$$\partial l(b_{hij}; x_{hij})/\partial b_j = \mathbf{x}_{hij} w_{hij} P(\mathbf{x}_{hij})(1 - P(\mathbf{x}_{hij})) \tag{3.47}$$

La función $\hat{G}(\hat{b_j})$ puede escribirse como una suma ponderada de totales al hacer $d_{hij}^{(l)} = x_{hij}^{(l)} s^{(l)}(b_j; y_{hij}, x_{hij})$ y se tiene, para $l = 1, ..., p$:

[6]Véase *Binder* [4] para una exposición detallada de este tema.

$$\hat{G}(\hat{b_l}) = \sum_{h=1}^{H} \sum_{i=1}^{nh} \sum_{j=1}^{mhi} w_{hij} d_{hij}^{(l)} = 0 \tag{3.48}$$

Debido a que esta última cantidad es una estimación de un total, la matriz de varianzas y covarianzas aproximadas usando la aproximación de Taylor alrededor de $\mathbf{B} = \hat{\mathbf{B}}$ se calcula como:

$$\hat{Var}(\hat{B}) = \left[\left(\frac{\partial \hat{G}(\hat{B})}{\partial \mathbf{B}} \right)^{-1} \hat{Var}(\hat{G}(\mathbf{B})) \left(\frac{\partial \hat{G}(\hat{B})}{\partial \mathbf{B}} \right)^{T} \right] \tag{3.49}$$

En la ecuación anterior para el valor poblacional $\frac{\partial \hat{G}(\hat{B})}{\partial \mathbf{B}} = \mathbf{H}$ es el hessiano de la logverosimilitud con elemento típico (k, l):

$$\frac{\partial^2 \hat{G}(\hat{B})}{\partial \mathbf{b_k} \partial \mathbf{b_l}} = -\sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hij}} x_{hij}^{(k)} x_{hij}^{(l)} P(\mathbf{x}_{hij}) \left(1 - P(\mathbf{x}_{hij})\right) \tag{3.50}$$

El hessiano de la logverosimilitud para elementos de la muestra $\frac{\partial \hat{G}(\hat{B})}{\partial \mathbf{B}}$ tiene elemento típico (k, l):

$$\frac{\partial^2 \hat{G}(\hat{B})}{\partial b_k \partial b_l} = -\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hij}} w_{hij} x_{hij}^{(k)} x_{hij}^{(l)} P(\mathbf{x}_{hij}) \left(1 - P(\mathbf{x}_{hij})\right) \tag{3.51}$$

3.4.1. Estimación del error cuadrático medio, esquemas A, B y C:. Para cada uno de los esquemas de estimación, A,B y C, la varianza aproximada difiere solamente en la cantidad $\widehat{Var}(\hat{G}(\mathbf{B}))$, que es el interior del estimador en (3.49), tal como se hizo en la estimación de los coeficientes en regresión lineal. A continuación se presentan los estimadores de varianza aproximada solamente para el interior.

Sean $z_{dhi}^{(k)} = \sum_{j=1}^{m_{hi}} w_{hij} d_{hij}^{(k)}$ y $\overline{z}_{dh}^{(k)} = \sum_{i=1}^{n_h} z_{dhi}^{(k)} / n_h$ el total estimado por UPM, así como el promedio muestral por UPM para la $k-$ésima variable independiente.

Esquema A: en el caso de contar con los factores de expansión y las probabilidades de inclusión de segundo orden se emplea la expresión (3.26).

$$\hat{v}_{l,k} = \hat{v}_{l,k\ UPM} + \hat{v}_{l,k\ USM} \tag{3.52}$$

donde,

$$\hat{v}_{l,kUPM} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_h}\frac{Cov\left(I_{hi},I_{hj}\right)}{\pi_{hij}}\frac{\hat{d}_{hi\pi}}{\pi_{hi}}\frac{\hat{d}_{hj\pi}}{\pi_{hj}} - \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hi}}\left(\frac{1}{\pi_{hi}}-1\right)\hat{V}_{hi} \tag{3.53}$$

$$\hat{V}_{hi} = \sum_{r=1}^{m_{hi}}\sum_{s=1}^{m_{hi}}\frac{Cov\left(I_{hr},I_{hs}|i\right)}{\pi_{hrs|i}}\frac{d_{hr|i}}{\pi_{hr|i}}\frac{d_{hs|i}}{\pi_{hs|i}} \tag{3.54}$$

$$\hat{v}_{l,kUSM} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{\hat{V}_{hi}}{\pi_{hi}^2} \tag{3.55}$$

Como ejemplo, bajo el esquema (H, MAS, MAS), $\hat{V}_{hi}$ adquiere la forma:

$$\begin{aligned}
\hat{V}_{hi} &= (1-f_i)\left\{\sum_{j=1}^{m_{hi}}\frac{M_{hi}^2}{m_{hi}^2}d_{hij}^{(k)}d_{hij}^{(l)} - \frac{1}{m_{hi}-1}\sum_{j=1}^{m_{hi}}\sum_{r=1}^{m_{hi}}\frac{M_{hi}^2}{m_{hi}^2}d_{hij}^{(k)}d_{hir}^{(l)}\right\} \\
&= (1-f_i)\left\{\sum_{j=1}^{m_{hi}}\frac{M_{hi}^2}{m_{hi}^2}d_{hij}^{(k)}d_{hij}^{(l)} - \frac{1}{m_{hi}-1}\left(\sum_{j=1}^{m_{hi}}\frac{M_{hi}}{m_{hi}}d_{hij}^{(k)}\right)\left(\sum_{j=1}^{m_{hi}}\frac{M_{hi}}{m_{hi}}d_{hij}^{(l)}\right)\right\} \\
&\quad + (1-f_i)\left\{\frac{1}{m_{hi}-1}\sum_{j=1}^{m_{hi}}\frac{M_{hi}^2}{m_{hi}^2}d_{hij}^{(k)}d_{hij}^{(l)}\right\} \\
&= (1-f_i)\frac{m_{hi}}{m_{hi}-1}\left\{\sum_{j=1}^{m_{hi}}\frac{M_{hi}^2}{m_{hi}^2}d_{hij}^{(k)}d_{hij}^{(l)} - \frac{1}{m_{hi}}\left(\sum_{j=1}^{m_{hi}}\frac{M_{hi}}{m_{hi}}d_{hij}^{(k)}\right)\left(\sum_{j=1}^{m_{hi}}\frac{M_{hi}}{m_{hi}}d_{hij}^{(l)}\right)\right\}
\end{aligned} \tag{3.56}$$

Esquema B: en el caso de usar ppt en la primera etapa y contar con los factores de expansión se emplea, con $d_{hij}^{(k)} = x_{hij}(y_{hij} - P(\mathbf{x}_{hij}))$, para las variables independientes l y k:

$$\widehat{Cov}\left(\hat{G}(b_k),\hat{G}(b_l)\right) = \hat{G}(b_k b_l) - \hat{G}(b_k)\hat{G}(b_l) \tag{3.57}$$

donde,

$$\hat{G}(b_k b_l) = \sum_{h=1}^{H}\frac{n_h}{n_h-1}\left\{\sum_{i=1}^{n_h}\frac{1}{\left(n_h p_{hi}\right)^2}\left(\sum_{j=1}^{m_{hi}}\frac{d_{hij}^{(k)}}{\pi_{hj|i}}\right)\left(\sum_{j=1}^{m_{hi}}\frac{d_{hij}^{(l)}}{\pi_{hj|i}}\right)\right\} \tag{3.58}$$

y

(3.59)

$$\hat{G}(b_k)\hat{G}(b_l) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left\{ \frac{1}{n_h} \left[\sum_{i=1}^{n_h} \frac{1}{n_h p_{hi}} \left(\sum_{j=1}^{m_{hi}} \frac{d_{hij}^{(k)}}{\pi_{hj|i}} \right) \right] \left[\sum_{i=1}^{n_h} \frac{1}{n_h p_{hi}} \left(\sum_{j=1}^{m_{hi}} \frac{d_{hij}^{(l)}}{\pi_{hj|i}} \right) \right] \right\}$$

Esquema C: en el caso de contar solo con los factores de expansión se emplea la siguiente expresión:

$$\hat{V}(\hat{G}(\hat{b_k})) = \sum_{h=1}^{H} (1 - f_h) \frac{n_h}{n_h - 1} \left(\sum_{i=1}^{n_h} (z_{dhi}^{(k)})^2 - n_h \, (\overline{z}_{dh}^{(k)})^2 \right) \tag{3.60}$$

$$\widehat{Cov}\left(\hat{G}(b_k), \hat{G}(b_l)\right) = \sum_{h=1}^{H} (1 - f_h) \frac{n_h}{n_h - 1} \left(\sum_{i=1}^{n_h} z_{hi}^{(k)} \, z_{hi}^{(l)} - n_h \, \overline{z}_h^{(k)} \, \overline{z}_h^{(l)} \right) \tag{3.61}$$

4. Simulación

Con el fin de comparar las aproximaciones a las varianzas de estimadores de regresión lineal y regresión logística en diseños muestrales complejos, se efectuaron estudios de simulación con dos poblaciones distintas. En cada población se formaron estratos y dentro de cada estrato se conglomeraron los elementos de la población en *UPM*. Las extracciones de elementos de la población se efectuaron con los siguientes diseños muestrales:

- Caso (MAS, MAS): esta notación indica la forma en que se llevó a cabo la extracción de elementos. Las *UPM* se extraen con MAS y dentro de cada *UPM* seleccionada en la muestra se extraen las *USM* con MAS. En este estudio se ignora la estratificación. Nótese que este diseño es un caso del esquema A mencionado en la segunda sección del documento.

- Caso (ppt, MAS): Las UPM se seleccionan con πpt y dentro de cada *UPM* seleccionada en la muestra se extraen las *USM* con MAS, pero las estimaciones se efectúan como si la muestra se hubiese extraído con ppt. Aquí también se ignora la estratificación. Este diseño es un caso del esquema B.

- Caso (H, MAS_h, MAS_h): Las *UPM* en cada estrato se extraen con MAS y dentro de cada UPM seleccionada en la muestra se extraen las *USM* con MAS. Obsérvese que corresponde al esquema A.

- Caso (H, ppt_h, MAS_h): Las *UPM* en cada estrato se seleccionan con πps, pero las estimaciones se efectúan como si la muestra se hubiese extraído con ppt y dentro de cada *UPM* seleccionada en la muestra se extraen las *USM* con MAS. Este esquema corresponde al esquema B.

Es necesario mencionar que las estimaciones correspondientes al esquema C, no requieren la extracción de muestra alguna, ya que en cada simulación se emplean los datos obtenidos por los casos (MAS, MAS) y (H, MAS, MAS) de selección de muestras y con esta información se construyen dichas estimaciones. Asimismo, los cálculos que se efectuan con el esquema A haciendo el calibrador $g_{hij} = 1$, véase sección 3,2,1, hacen uso de la muestra extraída para el esquema en cuestión.

4.1. Regresión lineal.

4.1.1. Población, estratos y variables. La población empleada fue una parte de la base de datos de personas, BCSPER03.DBF, de la encuesta asociada

es el ingreso original más el valor simulado de Z. Esto modifica un poco el valor del ingreso por hogar y la desviación estándar decrece ligeramente dependiendo del estrato. El modelo de trabajo para estimar los coeficientes beta supone que σ^2 es igual para todos los estratos. Es necesario mencionar que los elementos de la población son aquellos para los que el ingreso fue menor que $\$100,000$ pesos al mes, aquellos hogares con un ingreso mayor que el indicado fueron desechados[10]. A continuación se muestra un resumen de las principales estadísticas de la población.

Tabla 2
Principales estadísticas poblacionales

Estrato	Hogares por estrato	Tamaño relativo del estrato	Desviación estándar del ingreso	Asimetría del ingreso	Curtosis del ingreso	Total de personas
1	617	0.12	3,505.19	2.07	11.34	2,436
2	2,872	0.55	3,333.21	1.44	6.46	11,396
3	1,776	0.34	3,176.40	1.69	7.95	6,921
Total	5,265		3,314.42	1.60	7.54	20,753

Tabla 3
Correlación de Pearson entre las variables

	escolaridad	edad	personas
ingreso	0.820	0.865	0.673
escolaridad		0.424	0.446
edad			0.676

4.1.2. Diseños y tamaños de muestra usados. Con el fin de evaluar las diferencias para los estimadores puntuales del ingreso y de los coeficientes de regresión con los valores poblacionales, así como sus varianzas aproximadas estimadas, para los métodos de estimación, A y C, se extrajeron 1,000 muestras independientes con los siguientes tamaños: en la primera etapa de selección se usaron valores de 25 y 35 para las UPM y en cada conglomerado seleccionado se submuestreo con tamaños 5 y 10. Para el método B también se extrajeron 1,000 muestras independientes, usando los mismos tamaños de UPM y USM. Esto es necesario, ya que la forma de

[10]Esto se hizo así, ya que el ingreso antes de eliminar registros era una variable con una asimetría y curtosis grande, lo cual habría provocado valores grandes en las estimaciones del ingreso y varianza estimadas, en caso de hubiesen sido seleccionados en muestra. En el sección de conclusiones se abunda un poco más en el tema.

extraer muestras con probabilidad proporcional al tamaño es diferente a la extracción (MAS, MAS) ó (H, MAS, MAS).

4.1.3. Estimadores y cantidades calculadas, para cada uno de los diseños.

- Estimadores puntuales del total de ingreso $\hat{y}_{Uj}$ y de los coeficientes de regresión, intercepto $\hat{b}_{0j}$, años de escolaridad $\hat{b}_{1j}$ y edad $\hat{b}_{2j}$; con $j = \{A, B, C\}$.
- Errores cuadráticos medios estimados para cada uno de los estimadores del ingreso o de los coeficientes de regresión, del inciso anterior, $\hat{V}_j$, con $j = \{A, B, C\}$.
- Intervalos de confianza al 95 % con base en el cociente estudentizado para cada estimador puntual, $\hat{y}_{Uj} \pm z_{\alpha/2,gl}\sqrt{\hat{V}_j}$ y $\hat{b}_{ij} \pm z_{\alpha/2,gl}\sqrt{\hat{V}_{ij}}$ con $j = \{A, B, C\}$, $i = \{0, 1, 2\}$ $\alpha = 0.05$ y gl son los grados de libertad de la distribución t de Student.

4.1.4. Resumen de estadísticas calculadas. En cada simulación se generaron las muestras produciendo los valores: $\hat{y}_l, \hat{b}_{l0}, \hat{b}_{l1}$ y $\hat{b}_{l2}$, con $l = \{1, 2, ..., 1,000\}$. Las principales estadísticas calculadas para las $1,000$ muestras fueron las siguientes.

- Promedio de los estimadores puntuales para las $1,000$ muestras independientes, es decir, $\overline{\hat{y}}_{sim} = \sum_{l=1}^{1000} \hat{y}_l/1,000$ y $\overline{\hat{b}}_{l,sim} = \sum_{l=1}^{1000} \hat{b}_{lk}/1,000$, con $l = \{0, 1, 2\}$ y sim se refiere al promedio de las simulaciones.
- Sesgo observado relativo de los $1,000$ valores del ingreso y los coeficientes de regresión, es decir, la diferencia relativa entre el promedio de las $1,000$ observaciones y el valor poblacional. El sesgo observado relativo se calcula como: $(\overline{\hat{y}}_{sim} - y_U)/y_U$ y $(\overline{\hat{b}}_{l,sim} - b_l)/b_l$, con $l = \{0, 1, 2\}$.
- Promedio de los errores cuadráticos medios estimados, para cada uno de los $1,000$ valores observados del ingreso y los coeficientes de regresión según el método A, B, C y el A con el calibrador igual a uno.
- Porcentaje de cobertura observado de cada una de las series de $1,000$ intervalos de confianza por estimador puntual, esto es, el número observado de los $1,000$ intervalos que contuvieron a la cantidad poblacional, el ingreso o los coeficientes de regresión.

4.2. Regresión logística. El objetivo era estimar los coeficientes de regresión logística en un modelo con la variable dependiente igual al nivel de ingreso, con las variables edad y años de escolaridad como explicativas, empleando datos extraídos con los esquemas A, B y C de una población finita para comparar los errores cuadráticos medios estimados. Asimismo, se compararon las estimaciones y los errores cuadráticos medios estimados de los coeficientes de regresión de la misma forma que en la sección anterior para el caso de regresión lineal, pero solo se generaron 500 simulaciones por diseño y tamaño de muestra. Es necesario mencionar que este estudio no

está vinculado al de la sección anterior, por lo que la población objetivo, características por estimar, etc. son distintas.

4.2.1. Población, estratos y variables. La población empleada fue la base de datos de personas, BCSPER03.DBF, de la encuesta asociada al Censo de Población y Vivienda del 2000, para el estado de Baja California Sur. Se incluyeron los cinco municipios de BCS: Mulege, la Paz, los Cabos, Comondú y Loreto, cada uno de ellos se consideró como un estrato. Los elementos de la población fueron los hogares y para formar los conglomerados de hogares, se usaron las UPM del archivo siempre que el número de hogares fuese mayor o igual a 10. En caso contrario se juntaron con la UPM siguiente en la lista, y así sucesivamente. De esta manera, se trabajó con una población de 36,017 personas, resultando un total de 9,070 hogares y 907 UPM distribuidas en 5 estratos de la siguiente forma: 149 en Comondú, 89 en Mulege, 341 en la Paz, 242 en los Cabos y 86 en Loreto.

Para efectuar la simulación se emplearon variables categóricas usando el método de codificación de variables indicadoras para eliminar la sobreparametrización. La variable respuesta fue el ingreso del jefe del hogar usando las variables sexo y edad del jefe del hogar como explicativas. A continuación se muestran la definición de variables y los totales de personas y hogares por estrato y tipo de variable.

Tabla 4
Regresión logística: variables de estudio

Variable	Descripción	Tipo	Codificación
y_k	nivel de ingreso mensual del k-ésimo hogar	dependiente	$y_i = 1$ si ingreso del i-ésimo hogar $\leq 2,000 pesos$ $y_i = 0$ en otro caso
sexo	sexo del jefe del hogar	independiente	$sexo_i = 0$ si jefe del i-ésimo hogar es hombre $sexo_i = 1$ es mujer
edad0	edad del jefe del hogar	independiente	$edad0_i = 0$ si edad del jefe i-ésimo hogar $\in [40, 60)$
edad1	edad del jefe del hogar	independiente	$edad1_i = 1$ si edad del jefe i-ésimo hogar $\in [12, 24)$ $edad1_i = 0$ en otro caso
edad2	edad del jefe del hogar	independiente	$edad2_i = 1$ si edad del jefe i-ésimo hogar $\in [24, 40)$ $edad2_i = 0$ en otro caso
edad3	edad del jefe del hogar	independiente	$edad3_i = 1$ si edad del jefe i-ésimo hogar $\in [60, 99]$ $edad3_i = 0$ en otro caso

El modelo logístico empleó al ingreso y_k como la variable respuesta y a las variables *sexo*, *edad*1, *edad*2 y *edad*3 como explicativas. La variable *edad*0 ya no se incluye en los cálculos porque es la categoría de referencia para la variable edad.

Tabla 5
Totales de personas y hogares por estrato y tipo de variable

Estrato	Personas por hogar	Sexo (mujeres)	Número de personas en $[40, 60)$	Número de personas en $[12, 24)$	Número de personas en $[24, 40)$	Número de personas en $[60, 99)$	Ingreso
1	5,969	220	609	80	626	177	1,200
2	3,525	88	366	69	344	112	606
3	13,576	567	1,379	196	1,461	377	2,140
4	9,493	332	778	211	1,198	241	1,218
5	3,454	107	353	60	360	96	615
Total	36,017	1,314	3,485	616	3,989	1,003	5,779

4.2.2. Diseños y tamaños de muestra usados. Con el fin de evaluar las diferencias para los estimadores puntuales del ingreso y de los coeficientes de regresión con los valores poblacionales, así como sus varianzas aproximadas estimadas, para los tres métodos de estimación, A, B y C, se extrajeron 500 muestras independientes con los tamaños siguientes: 48 UPM, con tamaños de submuestreo 5 para las USM.

4.3. Algunos puntos sobre los paquetes *pps* y *Survey* de *R* para el análisis de encuestas complejas.

4.3.1. Paquete pps. Este paquete no está diseñado para el análisis, sino para la extracción de muestras de acuerdo con los siguientes diseños muestrales:

pps1(pps): Selección de una unidad con probabilidad proporcional a alguna medida de tamaño,

ppss(pps): Selección de unidades con probabilidad proporcional a alguna medida de tamaño sin reemplazo,

ppssstrat(pps): Selección estratificada de unidades con probabilidad proporcional a alguna medida de tamaño sin reemplazo,

ppswr(pps): Selección de unidades con probabilidad proporcional a alguna medida de tamaño con reemplazo,

sampford(pps): Selección de unidades con probabilidad proporcional a alguna medida de tamaño usando el método de Sampford,

sampfordpi(pps): Probabilidad de inclusión conjunta de unidades para el método de Sampford probabilidad proporcional a alguna medida de tamaño.

En este trabajo, se empleó la función `ppss` con el número de personas por hogar como la medida de tamaño. Esta función requiere dos argumentos: el primero es un vector con las medidas de tamaño, en tanto que el segundo es el tamaño de muestra deseado. La función devuelve los números de elementos en muestra como se ilustra a continuación.

Ejemplo 3. *Supóngase que se tiene una población con 9 elementos y se desea extraer una muestra de tamaño 4 con probabilidad proporcional a una medida de tamaño, almacenada en el arreglo* `medida`*. En la siguiente instrucción, se guarda el resultado en el vector llamado* `indicesmuestra`*.*

```
> medida <- c(9,2,5,17,4,21,15,7,4)
> indicesmuestra <- ppss(medida,4)
> indicesmuestra
[1] 1 4 6 7
```

Los elementos seleccionados en muestra son el 1, 4, 6 y 7.

Esta función puede emplearse para la selección de muestras sistemáticas, al usar un vector con medidas de tamaño igual a uno para todos los elementos de la población. Por ejemplo, si se desea extraer una muestra de tamaño 4 de una población con 9 elementos, se ejecutan las instrucciones siguientes:

```
> sistematico <- c(1,1,1,1,1,1,1,1,1)
> indicesist <- ppss(sistematico,4)
> indicesist
[1] 1 3 6 8
```

Los elementos seleccionados son el 1, 3, 6 y 8.

4.3.2. Paquete Survey. Este es un paquete general para el análisis de encuestas complejas. El diseño muestral puede describirse explícitamente o usando pesos para réplicas. La aproximación de la varianza se efectúa por métodos de réplicas basados en el método jackknife o usando el método de linealización de primer orden en serie de Taylor.

Las principales funciones del paquete son:

- Cálculo de estimaciones puntuales de totales y medias, así como de estimaciones de razón.
- Cálculo de los errores estándar de las cantidades anteriores.
- Postestratificación en las variables indicadas de acuerdo con totales conocidos de población.
- Creación de pesos para réplicas, con pesos provenientes de encuestas que usen estratos y/o conglomerados. Soporta los métodos de jackknife para diseños estratificados y sin estratificar, bootstrap y muestreo de medias muestras balanceadas.
- Métodos de calibración *GREG*, véase *Särndal et al* [24] y *raking*. Estos métodos calibran los totales de variables en una regresión lineal. Para usar el método *GREG* en diseños multietápicos es necesario proporcionar los totales de cada variable auxiliar por *UPM*. El método *raking* postestratifica de manera iterativa para tratar de igualar las distribuciones marginales muestrales a las marginales conocidas de la población.
- Método para el reescalamiento de los pesos muestrales por no respuesta.

Para emplear el paquete es necesario tener la información con estructura de una tabla de base de datos, es decir, un arreglo rectangular en el que las columnas constituyen las variables y cada renglón es el vector de información de un elemento en muestra. También se requiere especificar el diseño muestral con la instrucción `svydesign`, en la que se indica el número de etapas de selección por medio de una ecuación, las variables de identificación de unidades primarias de muestreo, *UPM*, un identificador de estratificación en caso procedente, pesos muestrales, las fracciones de muestreo para cada etapa de selección, en caso de que se cuenta con ellas, así como el nombre del archivo de datos muestrales.

Una forma de usar datos en `R`, es con la lectura de un archivo de texto con formato `CSV`. Como ejemplo, a continuación se presenta la lectura de datos de uno de los archivos empleados en los ejercicios de simulación. Obsérvese que tiene activada la opción de títulos en el encabezado con `header=TRUE`. Esto es importante, ya que son los campos o variables que se usarán para el análisis de datos.

```
datos <- read.csv("v2bcshogar.csv",header=TRUE)
```

Una vez que se leen los datos, se crea el `data.frame` para almacenar la muestra, ya que es el formato empleado por el paquete *Survey* de `R`.

```
muesim <- as.data.frame(matrix(0,nrow=n_upm*n_usm,ncol=15))
```

Después se asignan las etiquetas para el `data.frame`. Estas etiquetas facilitan el uso de instrucciones relativas a modelos lineales o estimaciones puntuales de promedios o proporciones.

```
dimnames(muesim)[[2]] <- c("consec","ups","uss","otro",
       "peso","pesocalc","pob","sexo","joven","mediano",
       "grande","fpc","fpc1","fpc2","estrato")
```

Con la instrucción mostrada al final del presente párrafo, se declara un diseño de conglomerados bietápicos *(MAS,MAS)* cuando se cuenta con la información de las fracciones de muestreo para la primera y segunda etapas. El diseño bietápico se indica con la ecuación `id=~ups+uss`, en la que `ups` y `uss` se refieren a los identificadores en la base de datos de las unidades primarias y secundarias de selección respectivamente. Si se tuvieren las correcciones por población finita para las dos etapas de selección, se emplearía la ecuación `fpc=~fpc1+fpc2`. Los datos se encuentran almacenados en el arreglo `muesim` y en `nclus2` se almacena la información del objeto diseño. Con el objeto `nclus2` y los datos almacenados en `muesim`, pueden usarse las funciones del paquete `Survey` para estimar totales, ajustar un modelo de regresión, usar algún modelo lineal generalizado, etc.

```
nclus2 <- svydesign(id=~ups+uss, fpc=~fpc1+fpc2, data=muesim)
```

Para un diseño bietápico *(MAS,MAS)*, cuando solo se cuente con los factores de expansión, como es el caso de muchas bases de datos públicas, en lugar de especificar las correcciones por pobalción finita para las dos etapas de selección, se declaran los pesos o factores de expansión con la ecuación `weights=~peso`.

```
nclus1 <- svydesign(id=~ups+uss,weights=~peso,data=muesim)
```

Si se deseare ajustar un modelo logístico para la variable respuesta `pob` con variable explicativa `sexo` , cuando solo se cuenta con los factores de expansión, se emplea la siguiente instrucción que se almacena en `rclus1`. La instrucción `summary` devuelve el detalle de las estimaciones puntuales de cada coeficiente en la regresión, el error estándar, el valor del estadístico t y el nivel de significancia para cada uno de los coeficientes estimados.

```
rclus1 <- summary(svyglm(pob~sexo+joven+mediano+grande,
          design=nclus1, family=quasibinomial()))
```

5. Resultados

5.1. Regresión lineal.

5.1.1. Valores poblacionales. Los valores poblacionales para los coeficientes de regresión lineal se obtuvieron al ajustar un modelo de regresión a todos los datos de la población finita con las variables mencionadas en la sección 4, usando el paquete estadístico *R* 2,4,0 con la instrucción usual de modelos lineales, como se muestra a continuación.

```
lm(formula = ingreso~edad + escol, x = TRUE)

Coefficients: Estimate  Std.Error  t value  Pr(> | t |)
(Intercept)   12.77168   2.94328     4.339  1.46e-05 ***
      edad    71.38144   0.05594  1275.957  < 2e-16 ***
     escol    41.98308   0.02881  1457.242  < 2e-16 ***
--Signif.codes: 0' ***' 0.001' ** ' 0.01'*' 0.05'.' 0.1'' 1

Residual standard error: 94.34 on 5262 degrees of freedom
Multiple R-Squared: 0.9992, Adjusted R-squared: 0.9992
F-statistic: 3.246e+06 on 2 and 5262 DF, p-value: < 2.2e-16
```

En la instrucción anterior, la opción x=TRUE devuelve la matriz de diseño, $\mathbf{x}^{'}\mathbf{x}$, la cual es:

```
             (Intercept)     edad         escol
(Intercept)        5265      162748        522380
        edad     162748     8496394      18998634
       escol     522380    18998634      64896864
```

El número de condicionamiento, calculado como la raiz cuadrada del cociente del máximo eigenvalor entre el mínimo eigenvalor, es: 262.354. Esto indica la existencia de un alto grado de multicolinealidad entre las variables independientes.

5.1.2. Resumen de resultados. A continuación se muestran las estadísticas producto de las simulaciones para cada uno de los cuatro diseños muestrales: estimador puntual, sesgo observado relativo del estimador puntual, cobertura observada del estimador puntual, sesgo observado relativo para la estimación de la varianza estimada y error cuadrático medio observado. Después de los cuadros, se encuentran diagramas de caja y brazos para cada uno de los cuatro diseños correspondientes a los estimadores puntuales del ingreso, del intercepto b_0 y de la escolaridad b_1. No se incluyeron las gráficas de los estimadores puntuales del coeficiente de escolaridad b_2, ya que en general se obtuvieron estimaciones del error cuadrático medio, así como sesgos observados relativos más pequeños que los correpondientes al coeficiente de

escolaridad. En cada uno de los diagramas de caja y brazos, se incluyó una línea horizontal que representa el valor poblacional por estimar. Posteriormente, se encuentran comentarios y observaciones. Con la asignación proporcional de UPM al tamaño relativo del estrato, véase cuadro 1, el número de UPM por estrato resultó de $\{3, 14, 8\}$ y $\{4, 19, 12\}$ para los tamaños de muestra 25 y 35 UPM respectivamente que aparecen en las tablas 6 a 9.

Tabla 6
Regresión lineal: estimadores puntuales
Promedio de 1,000 simulaciones

El total de ingreso está expresado en millones de pesos, en tanto que b_0,b_1, b_2 y NC se refieren al intercepto, escolaridad, edad y número de condicionamiento del modelo de regresión lineal.

Diseño Muestral	Tamaños de muestra (UPM,USM)	Esquema A, B ó C	Ingreso 33.6155	$b0$ 12.772	$b1$ 71.381	$b2$ 41.983	NC 262.354
(MAS,MAS)	(25,5)	A	33.6702	13.182	71.389	41.978	268.116
(MAS,MAS)		C	33.7450				
(H,MAS,MAS)		A	33.6650	12.997	71.400	41.978	267.900
(H,MAS,MAS)		C	33.6373				
(ppt,MAS)		B	33.6243	13.250	71.383	41.978	267.489
(H,ppt,MAS)		B	33.5902	13.445	71.372	41.980	267.250
(MAS,MAS)	(25,10)	A	33.5591	12.759	71.373	41.988	265.110
(MAS,MAS)		C	33.5275				
(H,MAS,MAS)		A	33.5832	13.410	71.376	41.981	265.356
(H,MAS,MAS)		C	33.5435				
(ppt,MAS)		B	33.6427	13.088	71.388	41.983	264.552
(H,ppt,MAS)		B	33.5464	12.891	71.386	41.982	264.831
(MAS,MAS)	(35,5)	A	33.6126	12.651	71.365	41.988	266.445
(MAS,MAS)		C	33.6401				
(H,MAS,MAS)		A	33.6858	13.369	71.374	41.981	267.327
(H,MAS,MAS)		C	33.7186				
(ppt,MAS)		B	33.5722	13.335	71.392	41.975	265.449
(H,ppt,MAS)		B	33.6408	13.319	71.378	41.981	266.619
(MAS,MAS)	(35,10)	A	33.6114	12.665	71.384	41.986	264.511
(MAS,MAS)		C	33.6204				
(H,MAS,MAS)		A	33.6350	12.073	71.388	41.987	265.540
(H,MAS,MAS)		C	33.6520				
(ppt,MAS)		B	33.6019	13.072	71.371	41.986	264.545
(H,ppt,MAS)		B	33.6500	13.314	71.396	41.978	264.844

Tabla 7
Regresión lineal: estimadores puntuales del error cuadrático medio
Promedio de 1,000 simulaciones

Para el ingreso las cantidades están expresadas en 10^7 y $calib = 1$ se refiere a la fórmula del esquema A, pero usando una aproximación a la varianza estimada aproximada usando el calibrador g igual a uno.

	Tamaños de muestra (UPM,USM)	Esquema A, B ó C	Ingreso	Ingreso $calib = 1$	b0	b1	b2
(MAS,MAS)	(25,5)	A	387,664.27	387,536.46	494.194	0.272	0.078
(MAS,MAS)		C	558,164.84		350.512	0.138	0.033
(H,MAS,MAS)		A	381,912.78	381,785.32	497.976	0.266	0.079
(H,MAS,MAS)		C	548,488.37		352.793	0.132	0.033
(ppt,MAS)		B	126,061.27		396.244	0.161	0.041
(H,ppt,MAS)		B	123,556.83		375.669	0.138	0.034
(MAS,MAS)	(25,10)	A	386,829.45	386,779.02	156.541	0.083	0.019
(MAS,MAS)		C	444,782.56		178.240	0.071	0.017
(H,MAS,MAS)		A	379,340.01	379,289.58	156.688	0.081	0.018
(H,MAS,MAS)		C	437,289.12		178.583	0.070	0.016
(ppt,MAS)		B	125,845.07		204.336	0.083	0.020
(H,ppt,MAS)		B	121,888.70		194.782	0.073	0.018
(MAS,MAS)	(35,5)	A	266,114.49	266,022.04	513.812	0.286	0.093
(MAS,MAS)		C	381,254.12		238.703	0.094	0.023
(H,MAS,MAS)		A	265,306.61	265,214.13	509.756	0.285	0.091
(H,MAS,MAS)		C	380,503.80		239.299	0.093	0.022
(ppt,MAS)		B	86,004.75		272.009	0.113	0.027
(H,ppt,MAS)		B	89,366.69		272.649	0.100	0.024
(MAS,MAS)	(35, 10)	A	264,852.48	264,816.14	120.135	0.069	0.017
(MAS,MAS)		C	304,650.05		121.626	0.048	0.011
(H,MAS,MAS)		A	267,086.67	267,050.31	123.744	0.072	0.018
(H,MAS,MAS)		C	306,945.04		124.932	0.051	0.012
(ppt,MAS)		B	88,954.27		142.330	0.061	0.014
(H,ppt,MAS)		B	89,217.89		139.580	0.052	0.012

Tabla 8
Regresión lineal: cobertura observada al 95 % para el ingreso y los coeficientes de regresión

$calib = 1$ se refiere a la fórmula del esquema A, pero usando una aproximación al error cuadrático medio usando el calibrador g igual a uno.

	Tamaños de muestra (UPM,USM)	Esquema A, B ó C	Ingreso	Ingreso $calib = 1$	b0	b1	b2
(MAS,MAS)	(25,5)	A	93.2	93.2	96.8	97.5	99.2
(MAS,MAS)		C	93.2		91.2	89.4	91.1
(H,MAS,MAS)		A	92.3	92.3	95.8	97.8	99.0
(H,MAS,MAS)		C	93.4		91.4	88.7	90.8
(ppt,MAS)		B	99.6		93.6	90.8	92.1
(H,ppt,MAS)		B	96.3		92.7	88.3	91.7
(MAS,MAS)	(25,10)	A	94.6	94.6	89.0	92.8	91.6
(MAS,MAS)		C	93.0		91.2	90.1	89.8
(H,MAS,MAS)		A	92.7	92.7	89.3	93.2	92.4
(H,MAS,MAS)		C	93.2		91.2	89.8	90.6
(ppt,MAS)		B	99.5		94.4	89.6	93.4
(H,ppt,MAS)		B	94.5		91.8	90.5	92.3
(MAS,MAS)	(35,5)	A	92.7	92.7	98.8	99.1	100.0
(MAS,MAS)		C	93.2		91.9	88.7	93.0
(H,MAS,MAS)		A	95.0	95.0	98.4	99.2	100.0
(H,MAS,MAS)		C	95.0		92.2	88.4	91.8
(ppt,MAS)		B	100.0		92.4	91.3	93.0
(H,ppt,MAS)		B	96.7		93.5	92.9	91.5
(MAS,MAS)	(35,10)	A	93.9	93.9	91.7	95.4	96.4
(MAS,MAS)		C	94.0		92.0	89.8	92.2
(H,MAS,MAS)		A	94.4	94.4	92.8	95.5	97.5
(H,MAS,MAS)		C	94.7		93.3	91.3	92.1
(ppt,MAS)		B	100.0		91.9	93.6	92.4
(H,ppt,MAS)		B	95.7		93.0	91.2	92.0

Tabla 9
Regresión lineal: sesgo observado relativo en porcentajes para el estimador del ingreso, los coeficientes de regresión y el número de condicionamiento, NC
Promedio de 1,000 simulaciones

Diseño muestral	Tamaños de muestra (UPM,USM)	Esquema A, B ó C	Ingreso	b0	b1	b2	NC
(MAS,MAS)	(25,5)	A	0.163	3.213	0.010	-0.013	2.196
(MAS,MAS)		C	0.385				
(H,MAS,MAS)		A	0.147	1.765	0.026	-0.011	2.114
(H,MAS,MAS)		C	0.065				
(ppt,MAS)		B	0.026	3.745	0.002	-0.013	1.957
(H,ppt,MAS)		B	-0.075	5.273	-0.014	-0.007	1.866
(MAS,MAS)	(25,10)	A	-0.168	-0.100	-0.012	0.013	1.051
(MAS,MAS)		C	-0.262				
(H,MAS,MAS)		A	-0.096	4.996	-0.008	-0.005	1.144
(H,MAS,MAS)		C	-0.214				
(ppt,MAS)		B	0.081	2.479	0.009	-0.001	0.838
(H,ppt,MAS)		B	-0.206	0.938	0.007	-0.002	0.944
(MAS,MAS)	(35,5)	A	-0.009	-0.942	-0.023	0.011	1.559
(MAS,MAS)		C	0.073				
(H,MAS,MAS)		A	0.209	4.675	-0.010	-0.005	1.896
(H,MAS,MAS)		C	0.307				
(ppt,MAS)		B	-0.129	4.408	0.014	-0.018	1.180
(H,ppt,MAS)		B	0.075	4.283	-0.005	-0.004	1.626
(MAS,MAS)	(35,10)	A	-0.012	-0.836	0.004	0.006	0.822
(MAS,MAS)		C	0.015				
(H,MAS,MAS)		A	0.058	-5.469	0.009	0.009	1.214
(H,MAS,MAS)		C	0.108				
(ppt,MAS)		B	-0.041	2.353	-0.015	0.007	0.835
(H,ppt,MAS)		B	0.102	4.244	0.020	-0.012	0.949

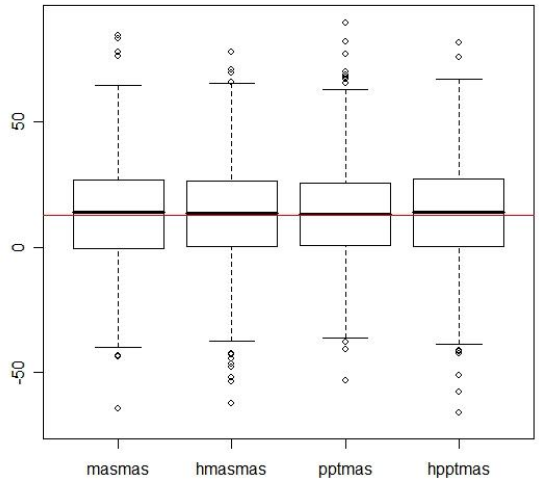

FIGURA 2. Estimadores puntuales de b0, UPM=25, USM=5

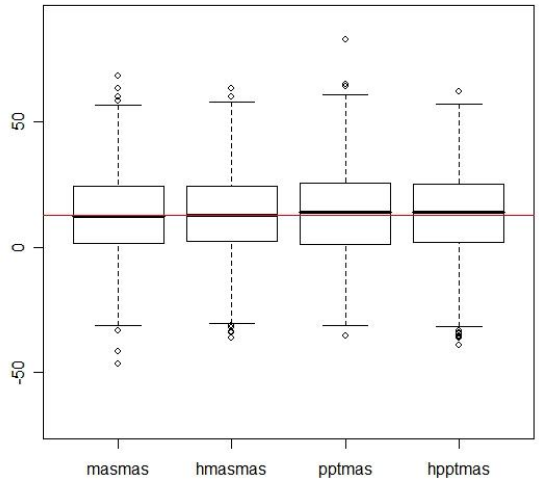

FIGURA 3. Estimadores puntuales del intercept b0, UPM=35, USM=5

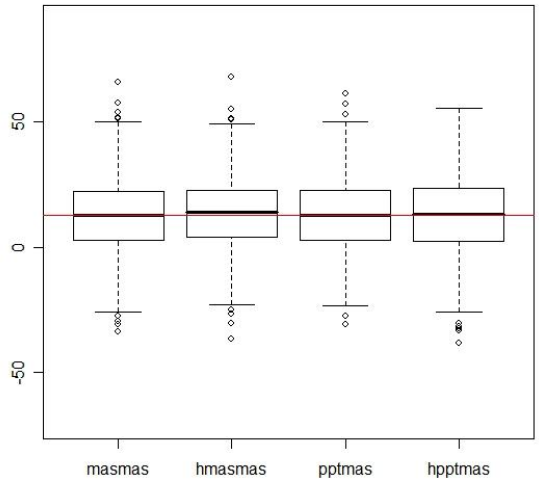

FIGURA 4. Estimadores puntuales del intercepto b0, UPM=25, USM=10

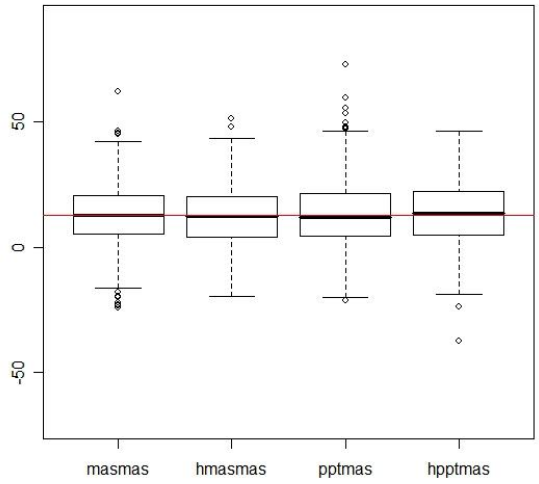

FIGURA 5. Estimadores puntuales del intercepto b0, UPM=35, USM=10

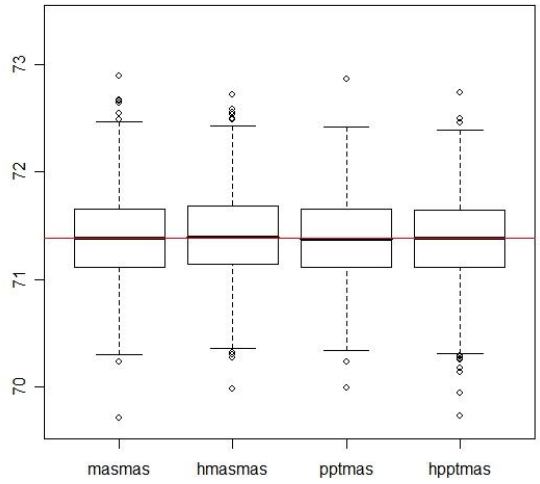

FIGURA 6. Estimadores puntuales de la variable escolaridad b1, UPM=25, USM=5

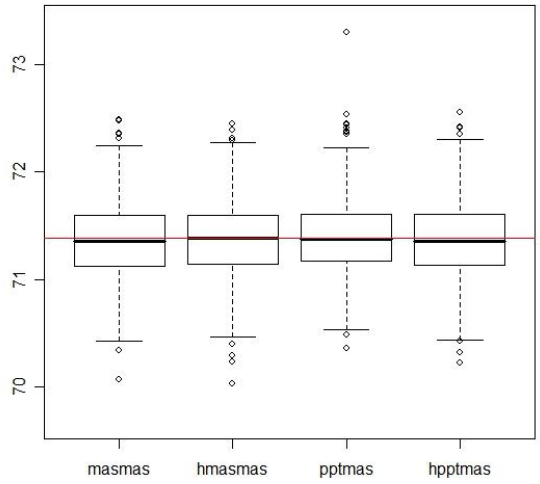

FIGURA 7. Estimadores puntuales de la variable escolaridad b1, UPM=35, USM=5

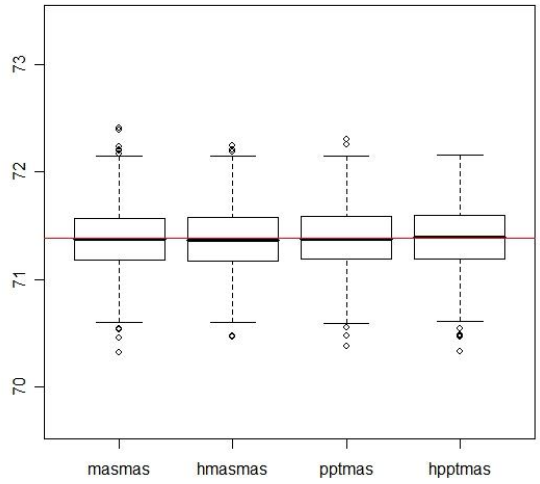

FIGURA 8. Estimadores puntuales de la variable escolaridad b1, UPM=25, USM=10

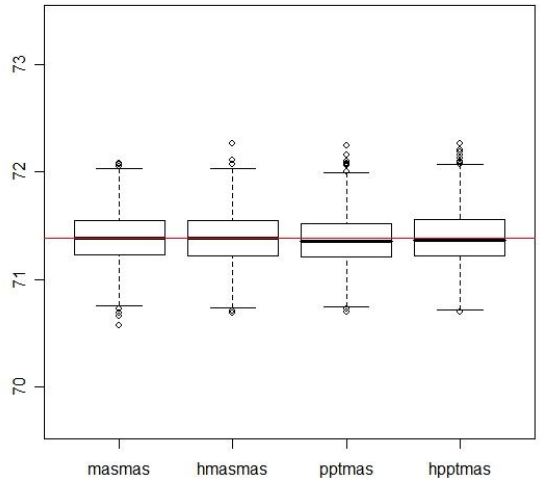

FIGURA 9. Estimadores puntuales de la variable escolaridad b1, UPM=35, USM=10

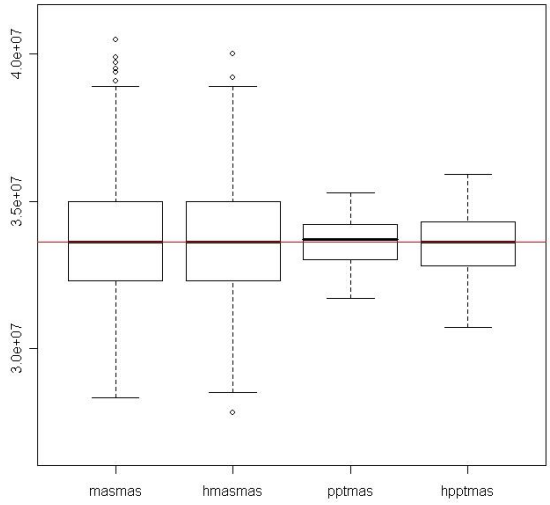

FIGURA 10. Estimadores puntuales de regresión para el ingreso, UPM=25, USM=5

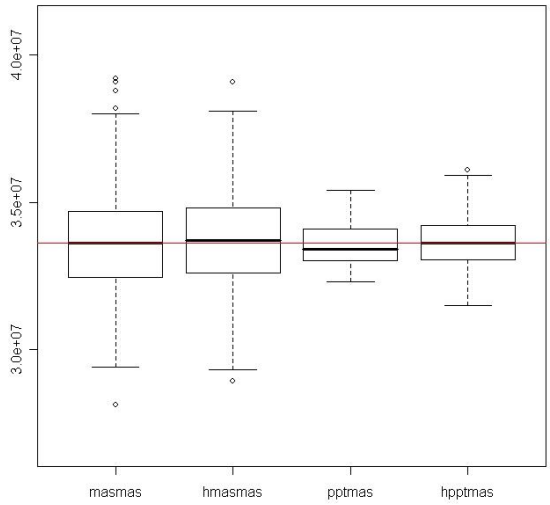

FIGURA 11. Estimadores puntuales de regresión para el ingreso, UPM=35, USM=5

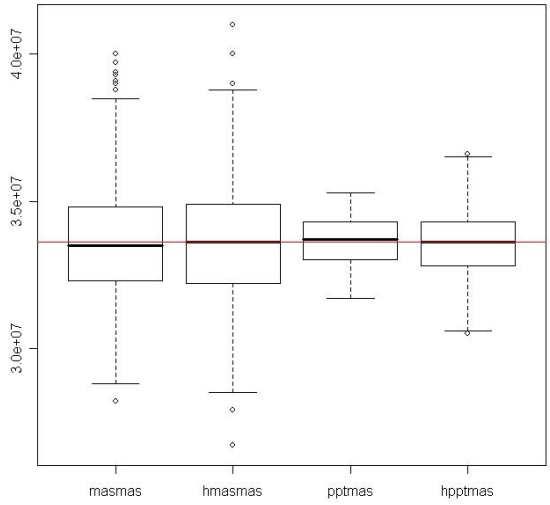

FIGURA 12. Estimadores puntuales de regresión para el ingreso, UPM=25, USM=10

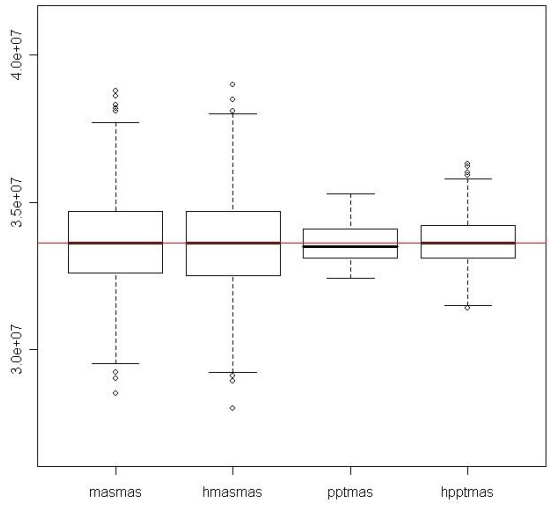

FIGURA 13. Estimadores puntuales de regresión para el ingreso, UPM=35, USM=10

Por lo que concierne a los estimadores puntuales del ingreso, en las tablas 6 y 9, así como en las figuras 10 a 13, se observa que para todos los diseños muestrales usados y tamaños de muestra, se obtuvieron estimaciones cercanas al valor poblacional. En particular, el sesgo observado relativo observado fue menor que 0.4 % en valor absoluto, como se aprecia en la tabla 9. De la tabla 7, se ve que las estimaciones puntuales del error cuadrático medio basadas en la aproximación del Esquema C, el uso de una expansión usando solo las probabilidades de inclusión de primer orden, son mayores que las del esquema A, el estimador de regresión generalizado que incorpora en la estimación la información auxiliar. Por otra parte, en esta misma tabla 7 se aprecia que con la estratificación, para cada esquema y tamaño de muestra, se obtuvo un error cuadrático medio observado menor que en el diseño bietápico (MAS, MAS). También se ve que los errores cuadráticos medios observados más pequeños se obtuvieron con los diseños *ppt*.

En cuanto a los estimadores puntuales del error cuadrático medio para la edad y escolaridad, con los diseños (H, MAS, MAS) y (MAS, MAS) se obtuvieron los valores más grandes para los cuatro diseños y tamaños de muestra, salvo en el caso del tamaño de muestra con 25 UPM y 10 hogares de submuestreo para b_1, en el que los diseños *ppt* presentaron estimaciones más grandes del error cuadrático medio. Empero, se observa que las estimaciones puntuales de b_1 y de su error cuadrático medio, fueron similares para los cuatro diseños y distintos tamaños de muestra, véase las figuras 6 al 9, así como las tablas 6 y 7. En cuanto al intercepto, en general se observó un sesgo relativo grande comparado con las estimaciones de los coeficientes de edad y escolaridad, para todos los diseños. En las figuras 2 al 5 se aprecia el rango de valores observados en las estimaciones del intercepto; en particular, el 25 % de las estimaciones puntuales fueron negativas con el tamaño de submuestreo igual a cinco. Por otra parte, en la tabla 7 se ve que los promedios de las estimaciones del error cuadrático medio del intercepto se encuentran entre 120 y 510, para todos los tamaños de submuestreo, lo cual es bastante grande comparado con las estimaciones puntuales de la tabla 6.

En relación con la cobertura de los intervalos, en la tabla 8 se ve que los diseños estratificados para el tamaño de muestra de 35 UPM casi alcanzaron o superaron la cobertura nominal, de hecho los diseños (ppt, MAS) y el (H, ppt, MAS) lograron una cobertura observada superior al 95 % para el ingreso en todos los tamaños de muestra. También se observa que el incremento en el tamaño del submuestreo, de 5 a 10 USM por conglomerado en muestra, mejoró la cobertura observada para la variable ingreso. En general, fue similar la cobertura observada para los esquemas A y C en relación con el ingreso. Por otra parte, la cobertura observada de los coeficientes de regresión en general fue superior para el esquema A comparado con los otros dos, los cuales en muchos casos adquieren valores alrededor del 90 %.

En todas las tablas se ve que la aproximación empleada al hacer el calibrador $g = 1$ en (3.34), no varía los resultados. Esto se debe a que la expansión del estimador del total por conglomerado muestreado de cada variable auxiliar es cercana al valor poblacional. Cuando se emplean calibradores, es importante verificar que éstos no sean negativos, como se mencionó en la sección 3. Para verificar esta situación, durante las simulaciones se contó el número de calibradores negativos y se obtuvieron 2 y 6 calibradores negativos para los diseños (MAS, MAS) y (H, MAS, MAS) respectivamente. Esta cantidad fué tan pequeña que no se consideró necesario emplear métodos para remediarlo[11].

Una parte importante al usar un modelo de regresión lineal se refiere a la dependencia entre las variables de la matriz de diseño. Una medida de la inestabilidad de los estimadores se tiene al calcular el número de condicionamiento del estimador de la matriz de diseño, ver *Lehtonen* [14]. El valor poblacional indica que hay un problema de multicolinealidad, ya que el número de condicionamiento de la matriz de diseño para el modelo poblacional es igual a *262.354*. En los cuadros 1 y 4 se aprecia que los diseños muestrales, salvo el (H, ppt, MAS), estimaron con poco sesgo dicho valor. Además, el grado de multicolinealidad no afectó el resultado de las estimaciones basadas en el estimador de regresión, ya que los estimadores no se vieron afectados en su capacidad predictiva. No obstante lo anterior, el modelo usado en (3.20) no sería adecuado si se utilizara en términos causales.

5.2. Regresión logística. Los valores poblacionales para los coeficientes de regresión se obtuvieron al ajustar un modelo de regresión logística a todos los datos de la población finita con las variables mencionadas en la sección 4, usando el paquete estadístico *R* 2,4,0 con la instrucción usual de modelos lineales generalizados, como se muestra a continuación.

```
m1m2 <- glm(pob~sexo+ed1+ed2+ed3,family=binomial,data=datos)
modm1m2  <-summary(m1m2)

Coefficients: Estimate Std.Error z value Pr(> | z |)
   (Intercept) 0.12510  0.04819   2.596  0.009438**
         sexo  0.41291  0.05271   7.833  4.75e-15***
          ed1  0.21655  0.09241   2.343  0.019110*
          ed2  0.16914  0.04908   3.446  0.000569***
          ed3  0.25169  0.07507   3.353  0.000801**
```

En la instrucción anterior, *pob* es la variable dependiente nivel de ingreso, *data* = *datos* se refiere al nombre en el que se encuentran las cinco series de

[11]En *Tinajero y Eslava* [29] se describe detalladamente diversos métodos para garantizar que los calibradores usados en el estimador de regresión sean positivos.

datos.

Las variables $ed1$, $ed2$ y $ed3$ son las variables recodificadas según lo indicado en la sección 4 y se refieren a los grupos de edad 2, 3 y 4. Nótese que en este caso la variable $ed0$ no aparece en la salida del programa, ya que en la recodificación para evitar la sobreparametrización tiene el valor de referencia igual a 0.

A continuación se muestran las estadísticas producto de las simulaciones para cada uno de los cuatro diseños muestrales: estimador puntual, sesgo observado relativo del estimador puntual, cobertura observada del estimador puntual, sesgo observado relativo para la estimación de la varianza estimada y error cuadrático medio observado.

Tabla 10
Regresión logística: estimadores puntuales.
Cantidades por 10^{-2}
Promedio de 500 simulaciones

Diseño muestral	Tamaños de muestra (UPM) y USM	Esquema A, B ó C	b_0 12.51	b_1 41.29	b_2 21.65	b_3 16.91	b_4 25.16
(MAS,MAS)	48 y 5	A y C	14.45	38.57	25.14	18.33	30.74
(H,MAS,MAS)	(8,4,18,14,4) y 5	A y C	11.15	44.38	38.08	15.85	26.95
(ppt,MAS)	48 y 5	B	12.59	41.61	31.63	16.86	30.73
(H,ppt,MAS)	(8,4,18,14,4) y 5	B	24.33	40.86	39.18	22.57	28.75

Tabla 11
Regresión logística: estimaciones del error cuadrático medio.
Cantidades por 10^{-2}
Promedio de 500 simulaciones

Diseño muestral	Tamaños de muestra (UPM) y USM	Esquema A, B ó C	b_0	b_1	b_2	b_3	b_4
(MAS,MAS)	48 y 5	A	4.70	4.18	20.40	3.81	14.93
		C	5.23	4.62	21.54	4.39	16.33
(H,MAS,MAS)	(8,4,18,14,4)y 5	A	8.35	5.99	61.52	5.79	16.22
		C	9.86	7.12	60.63	7.00	17.29
(ppt,MAS)	48 y 5	B	7.92	5.41	37.07	6.80	13.10
(H,ppt,MAS)	(8,4,18,14,4)y 5	B	18.13	12.17	31.40	11.86	16.68

Tabla 12
Regresión logística: cobertura al 95 % para los estimadores puntuales de los coeficientes de regresión.
Promedio de 500 simulaciones

Diseño muestral	Tamaños de muestra (UPM) y USM	Esquema A, B ó C	b_0	b_1	b_2	b_3	b_4
(MAS,MAS)	48 y 5	A	89.8	89.0	91.2	90.4	90.0
		C	90.2	89.2	91.6	91.0	90.2
(H,MAS,MAS)	(8,4,18,14,4)y 5	A	90.8	91.8	89.6	90.8	89.2
		C	91.2	92.2	89.6	91.0	89.6
(ppt,MAS)	48 y 5	B	92.4	89.8	92.0	91.2	86.2
(H,ppt,MAS)	(8,4,18,14,4)y 5	B	87.6	88.6	87.0	87.6	85.2

Tabla 13
Regresión logística: sesgo observado relativo de los estimadores puntuales de los coeficientes de regresión.
Cantidades expresadas en porcentajes
Promedio de 500 simulaciones

Diseño muestral	Tamaños de muestra (UPM) y USM	Esquema A, B ó C	b_0	b_1	b_2	b_3	b_4
(MAS,MAS)	48 y 5	A y C	15.5	-6.6	16.1	8.4	22.2
(H,MAS,MAS)	(8,4,18,14,4) y 5	A y C	-10.9	7.5	75.9	-6.3	7.1
(ppt,MAS)	48 y 5	B	0.6	0.8	46.1	-0.3	22.1
(H,ppt,MAS)	(8,4,18,14,4) y 5	B	94.5	-1.0	81.0	33.5	14.3

En cuanto a los estimadores puntuales de los coeficientes de regresión, en las tablas 10 y 13, se aprecia que no hay un diseño con los valores más pequeños del sesgo observado relativo ya que el diseño (ppt, MAS) tiene el menor sesgo para el intercepto, sexo y el tercer rango de edad; en tanto que para el rango cuarto de edad el diseño (H, MAS, MAS) tiene el menor sesgo y para el segundo rango de edad el (MAS, MAS) obtuvo el sesgo más pequeño. En este aspecto, el peor diseño fue el (H, ppt, MAS) ya que para todas las variables, excepto b_1 tuvo un alto sesgo relativo, en particular para el intercepto y el sexo.

Por lo que concierne a la cobertura, como se observa en la tabla 12, en ningún caso se alcanzó el nivel nominal del 95 %, ya que las coberturas observadas adquirieron valores entre 85 % y 92 %. Salvo el diseño (H, ppt, MAS)

que tuvo las coberturas más bajas, los 3 diseños restantes tuvieron un comportamiento similar que varió alrededor del 90 %.

En cuanto a los estimadores puntuales del error cuadrático medio, en la tabla 11 se aprecia que el diseño (MAS, MAS) para los dos esquemas, A y C, obtuvo los valores más pequeños. Se ve que la estratificación no logró disminuir el error cuadrático medio estimado ya que se obtuvieron valores más grandes que el diseño bietápico y el (ppt, MAS). En este punto se observa que el efecto de la estratificación con asignación proporcional combinado con la conglomeración no siempre logra disminuir la varianza. También se ve que el diseño (ppt, MAS) tiene cuatro de cinco estimadores del error cuadrático medio observado más pequeños que los esquemas A y C para el diseño estratificado bietápico. Así, se tiene que el esquema C tiene un error cuadrático medio observado más grande que el esquema A, con o sin estratificación.

5.2.1. Cuasiseparación de datos. Con el fin de evaluar la estabilidad de los estimadores puntuales de los coeficientes de regresión logística, se extrajeron 2,000 muestras, con el mismo tamaño de muestra, 48 *UPM* y 5 *USM* por conglomerado en muestra, para los cuatro diseños empleados con 500 simulaciones. Se calcularon los estimadores puntuales de los coeficientes de regresión, así como el sesgo observado relativo. Los resultados se muestran a continuación.

Tabla 14
Regresión logística: estimadores puntuales.
Cantidades por 10^{-2}
Promedio de 2,000 simulaciones

Diseño muestral	Tamaños de muestra (UPM) y USM	Esquema A, B ó C	b_0 12.51	b_1 41.29	b_2 21.65	b_3 16.91	b_4 25.16
(MAS,MAS)	48 y 5	A y C	12.69	42.76	32.06	16.26	26.72
(H,MAS,MAS)	(8,4,18,14,4) y 5	A y C	12.97	41.34	31.86	16.49	27.35
(ppt,MAS)	48 y 5	B	13.49	41.91	28.24	16.97	33.11
(H,ppt,MAS)	(8,4,18,14,4) y 5	B	25.01	42.17	33.84	19.04	25.61

Tabla 15
Regresión logística: sesgo observado relativo de los estimadores puntuales de los coeficientes de regresión.
Cantidades expresadas en porcentajes
Promedio de 2,000 simulaciones

Diseño muestral	Tamaños de muestra (UPM) y USM	Esquema A, B ó C	b_0	b_1	b_2	b_3	b_4
(MAS,MAS)	48 y 5	A y C	1.26	3.66	47.95	-3.30	6.13
(H,MAS,MAS)	(8,4,18,14,4) y 5	A y C	3.50	0.21	47.04	-1.90	8.63
(ppt,MAS)	48 y 5	B	7.62	1.58	30.33	0.96	31.52
(H,ppt,MAS)	(8,4,18,14,4) y 5	B	99.58	2.22	56.16	13.28	1.73

Tabla 16
Regresión logística: comparación de promedios del sesgo observado relativo (SOR) de los estimadores puntuales de los coeficientes de regresión de 2,000 y 500 simulaciones.
El valor 1 indica que el valor absoluto del SOR calculado con 2,000 simulaciones fue menor que el calculado con 500 simulaciones, el valor 0 indica el caso contrario

Diseño muestral	Tamaños de muestra (UPM) y USM	Esquema A, B ó C	b_0	b_1	b_2	b_3	b_4
(MAS,MAS)	48 y 5	A y C	1	1	0	1	1
(H,MAS,MAS)	(8,4,18,14,4) y 5	A y C	1	1	1	1	0
(ppt,MAS)	48 y 5	B	0	0	1	0	0
(H,ppt,MAS)	(8,4,18,14,4) y 5	B	0	0	1	1	1

En las tablas 14 a 16 se observa que el incremento en el número de simulaciones disminuyó en general el valor absoluto del sesgo observado relativo para los diseños *(MAS,MAS)* y *(H,MAS,MAS)*, salvó los estimadores de b_2 y b_4 para el caso *(MAS,MAS)* y *(H,MAS,MAS)* respectivamente. Para el diseño *(H,ppt,MAS)* solo los estimadores de los coeficientes de la edad lograron una disminución en valor absoluto del sesgo observado relativo, en tanto que el diseño *(ppt,MAS)* fue en el que solo se logró una disminución en el estimador de b_2.

Con el fin de identificar las posibles causas del comportamiento en el sesgo observado relativo de los estimadores puntuales de los coeficientes de regresión, durante las simulaciones se contaron las muestras que tuvieran

un coeficiente estimado de b_2 mayor que 5. El número de muestras con el estimador de b_2 mayor que 5 fue de: 8, 8, 3 y 7 para los diseños *(MAS,MAS)*, *(H,MAS,MAS)*, *(ppt,MAS)* y *(H,ppt,MAS)* respectivamente. Además, para el diseño *(MAS,MAS)*, se guardaron los resultados de las 2,000 estimaciones puntales de los coeficientes de regresión y se aisló la primera muestra con un estimador de b_2 mayor que 5. A continuación se presenta la gráfica de caja y brazos para b_2.

En la gráfica se observa que existe un conjunto de estimadores con un valor alto. Estas observaciones corresponden a las muestras con el estimador de b_2 mayor que 5. Cabe mencionar que el promedio de los estimadores de los otros coeficientes tienen como máximo los siguientes valores: 1.30, 1.91, 1.26 y 2.93.

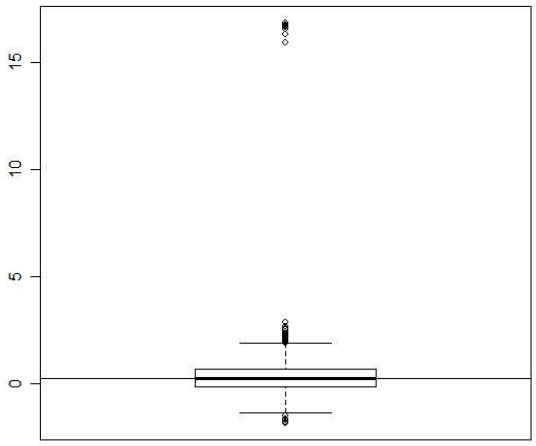

FIGURA 14. Estimadores puntuales para la variale b2, UPM=48, USM=5

Por otra parte, en la muestra que se aisló por tener un estimador de b_2 mayor que 5, los valores estimados de los coeficientes fueron:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3899     0.3466   1.125   0.2675
sexo          0.7268     0.3766   1.930   0.0609 .
joven        16.1742     0.2670  60.575   <2e-16 ***
mediano      -0.2169     0.2474  -0.877   0.3861
grande       -0.2000     0.4714  -0.424   0.6736
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Number of Fisher Scoring iterations: 16
```

En los estimadores se observa que el valor de b_2 es inusualmente alto, en tanto que los estimadores de b_3 y b_4 tuvieron un signo distinto al esperado. En opinión de *Hosmer y Lemeshow*[11], un valor inusualmente alto para estimadores de coeficientes en regresión logística, es indicativo de un problema de separación de datos. Además, el estimador de máxima verosimilitud no existe, véase *Albert y Anderson*[2] ó *Godínez y Ramírez*[7] para este problema en el caso clásico. Por otra parte, en *Albert y Anderson*[2] y *Santner y Duffy*[23] se encuentra un algoritmo para identificar la separación de datos y en *le Cessie y van Houwelingen*[13], así como en *Godínez y Ramírez*[7], se proponen estimadores que existen en presencia de separación o cuasiseparación de datos para el caso clásico. Sin embargo, con base en la literatura revisada, no se encontraron algoritmos para la identificación del problema de separación o cuasiseparación de datos, ni estimadores que existan bajo dicha condición, en muestras provenientes de diseños complejos. Por lo anterior, en la muestra aislada se contaron los elementos de la población con edad entre 12 y 24 años, la correspondiente al coeficiente b_2, que tuvieran un ingreso mensual menor o igual que 2,000 pesos. De los 240 elementos en muestra, 48 *UPM* por 5 *USM*, solamente 13 jefes de familia pertenecían a la categoría de edad entre 12 y 24 años. Los 13 tenían un ingreso mensual menor o igual que 2,000 pesos, lo cual puede deberse a que no había traslape de datos en muestra para dicha categoría o se tiene una cuasiseparación de los datos.

Es necesario mencionar que en la población, los jefes de familia pertenecientes a la categoría de edad entre 12 y 24 años con ingreso superior a 2,000 pesos, constituyen un 2.2 % de la población de jefes de hogar; en tanto que los jefes de familia en la misma categoría de edad, pero con ingreso menor o igual que 2,000 pesos, constituyen un 4.5 % de la población. Es decir, en la población no se presenta el problema para esta categoría de edad.

6. Conclusiones, recomendaciones y/o limitaciones

A pesar de que un estudio de simulación, como el efectuado en este trabajo, no es generalizable a cualquier situación, ya que se refiere a las condiciones de la población o poblaciones bajo estudio y de los diseños muestrales empleados, es posible desprender algunas conclusiones.

6.1. Regresión lineal. La estimación puntual del total de ingreso fue buena en general, ya que el sesgo observado relativo fue menor al 0.4 % para todos los diseños excepto el (H, ppt, MAS). Para los tamaños de muestra de 25 UPM el diseño (ppt, MAS) tuvo el menor sesgo relativo, en tanto que para 35 UPM el menor sesgo relativo se obtuvo con el diseño (MAS, MAS). La estimación puntual para los coeficientes de edad y escolaridad también fue buena, ya que el sesgo observado relativo para todos los diseños, fue menor al 0.09 % y no se aprecia un diseño que haya superado consistentemente a los otros con un sesgo relativo pequeño. En cuanto a la estimación del intercepto, el sesgo observado relativo fue grande, alcanzando hasta un 5 %, salvo el caso del diseño (H, ppt, MAS). No hubo un diseño que tuviese en general, un menor sesgo relativo; empero, para los tamaños de muestra $(35, 5)$ y $(35, 10)$ el diseño (MAS, MAS) presentó el menor sesgo para la mayor parte de los coeficientes. Por otra parte, a pesar de que no era el objeto de estudio, la estimación del número de condicionamiento poblacional presentó un sesgo relativo menor al 2.5 %.

Los estimadores puntuales del error cuadrático medio del ingreso presentaron grandes variaciones en magnitud, ya que los dos diseños con selección *ppt* tuvieron varianzas aproximadas estimadas un 30 % más pequeñas que los diseños bietápico y estratificado bietápico. También se verificó el resultado de que la varianza estimada del estimador de regresión para un total, en este caso el ingreso, esquema A, es menor que la expansión simple, esquema C, siempre que se tenga una correlación alta entre las variables dependientes y la independiente, y se use un tamaño de submuestreo grande. En cuanto al error cuadrático medio estimado para los coeficientes de regresión, salvo el intercepto, el esquema C presentó valores más pequeños que los esquemas A y B. Los errores cuadráticos medios de la estimación del intercepto presenta un comportamiento errático que impide una afirmación acerca del comportamiento por esquema y diseño; solo se aprecia la consistencia del estimador, ya que al aumentar el tamaño de la muestra, el error cuadrático medio se hace más pequeño. Sin embargo, para los tamaños de muestra usados, los intervalos de confianza son de tal magnitud que contienen al cero, lo cual conduciría a eliminar del modelo al intercepto.

La cobertura obsevada, no estuvo alejada del valor nominal, 95 %, para la estimación del ingreso. No obstante lo anterior, la cobertura observada para las estimaciones de los coeficientes de regresión presentó un rango más amplio de valores al obtener resultados entre 89 % y 100 %. El esquema A

presentó en general una cobertura más cercana a la nominal que el esquema C.

En conclusión, el uso del estimador bajo el esquema C presentó en general un sesgo observado relativo y una varianza estimada mayor que la del estimador de regresión del esquema A; en tanto que el sesgo observado relativo fue igual para ambos esquemas, ya que se trata del mismo estimador. Por otra parte, los estimadores del error cuadrático medio usando el esquema C tuvieron valores menores hasta en un 40 % que los estimados bajo el esquema A para los coeficientes distintos del intercepto, lo cual se vió reflejado en coberturas observadas más bajas que la nominal.

6.2. Regresión logística. Los estimadores puntuales presentaron un sesgo observado relativo grande en general, para los diseños y esquemas empleados, ya que no se obtuvo un diseño con un sesgo observado pequeño para todos los coeficientes. La cobertura observada estuvo entre un 2.5 % y casi 10 % por debajo de la nominal. Los errores cuadráticos medios observados tuvieron una amplia variación entre diseños y esquemas, siendo el asociado al intercepto el que presentó las diferencias entre estimadores, con una diferencia de casi cuatro veces entre las estimaciones del diseño (MAS, MAS) y el (H, ppt, MAS). En este caso. se cumplió con uno de los objetivos principales del trabajo, al determinar empíricamente la diferencia entre el empleo de estimadores con base en el uso del diseño complejo, esquema A, y los que se construyen usando solo los factores de expansión y suponiendo independencia, esquema C. La principal conclusión es que el uso del esquema C conduce a estimadores con un error cuadrático medio estimado más grande que al usar las estimaciones del esquema A.

6.3. Limitaciones. Una limitante en ambos casos, regresión lineal y logística, es la existencia de la multicolinealidad. En el caso de la regresión lineal, la multicolinealidad puede tratarse en el problema de la estimación usando una inversa generalizada o la inversa de Moore-Penrose, ver *Hidiroglou* [10]. Para la regresión logística, puede usarse este mismo enfoque; empero, en el caso logístico existe otro problema que se tiene al emplear variables categóricas y es el de la cuasi-separación y separación de datos, véase *Hosmer y Lemeshow* [11] y *Albert* [2]. Como se mecionó en la sección *5.2.1*, con base en lo revisado en la literatura a la fecha, no se tienen referencias de desarrollos de estimadores o estudios de simulación para casos de multicolinealidad en regresión lineal o logística con datos provenientes de encuestas complejas. En el mismo tenor, tampoco se encontraron artículos para atacar el problema de la multicolinealidad y/o cuasi-separación y separación de datos en regresión logística en encuestas complejas.

Otro punto que merece atención es el que la documentación de paquetes comerciales como *Stata* y *Sudaan* explican el método teórico de estimación, por ejemplo, para la regresión logística y mencionan que se resuelven las ecuaciones de estimación para obtener los parámetros, pero no indican el método numérico de optimización empleado, ni la forma de calcular los valores iniciales del algoritmo. Por ejemplo, en el paquete de muestreo de R 2,4,0 se conoce el método de optmización y pueden introducirse valores iniciales si así se desea.

Al efectuar la descripción de la población para la regresión lineal, se mencionó que se habían eliminado valores atípicos, grandes, del ingreso. En simulaciones iniciales se habían incluído estos valores, pero se tuvieron resultados inestables ya que algunas muestras no contenían a los valores atípicos y otras sí, resultando en sub o sobre-estimaciones grandes del valor poblacional del ingreso. Con los coeficientes de regresión también se tuvieron problemas, ya que las varianzas estimadas tenían valores inusualmente grandes. Por estos motivos, se decidió trabajar con una población menos asimétrica. Por supuesto, en la práctica estos datos pueden surgir y se han desarrollado técnicas para tratar el problema como la modificación de los valores atípicos o la modificación del peso muestral[12].

6.4. Recomendaciones. A continuación se mencionan los aspectos que requieren un análisis más detallado y que se encontraron en el desarrollo del presente trabajo:

- Análisis de las causas de sesgos relativos grandes para el caso logístico y para el intercepto en el caso de la regresión lineal.
- Explorar las causas de las coberturas observadas por debajo del valor nominal en la regresión logística.
- Revisar el estado actual del tratamiento de datos atípicos y su uso en el análisis de regresión lineal en encuestas complejas.
- Detección y estimación bajo cuasi-separación o separación de datos en regresión logística aplicadas a datos provenientes de encuestas complejas.
- Realizar simulaciones similares para otros casos de familias en modelos lineales generalizados.

[12]Una forma de estimar totales en la presencia de datos atípicos en encuestas complejas es con el uso de los métodos denominados *winsorization*.

Apéndice: ejemplo de códigos en R

```
{Regresión logística: código en R para extracción de muestra (MAS,MAS)}

  #información poblacional, lectura de datos
datos <- read.csv("v2bcshogar.csv",header=TRUE)
names(datos)
#carga el paquete de muestreo
library(survey)
#ajusta el modelo logístico a los datos poblacionales
m1m2 <- glm(pob~sexo+edad1+edad2+edad4,family=binomial,data=datos)
modm1m2  <-summary(m1m2)

#número de upm en muestra y de usm dentro de upm
tam_upm <- 907
n_upm   <- 48
tam_usm <- 10
n_usm   <- 5
n <- n_usm*n_upm
#fracciones de muestreo y peso muestral
f1 <- n_upm/tam_upm
f2 <- n_usm/tam_usm
f  <- f1*f2
wp <- 1/f
gl <-n_upm*n_usm
#extracción de upm por mas y ordenadas
mue_upm <- sort(sample(1:tam_upm,n_upm,replace=FALSE))

#se crea el data.frame para almacenar la muestra, ya que es el formato del
#paquete survey
#matriz para almacenar los datos de la muestra
muesim <- as.data.frame(matrix(0,nrow=n_upm*n_usm,ncol=15))
#etiquetas para el data.frame
dimnames(muesim)[[2]] <- c("consec","ups","uss","otro","peso","pesocalc","pob",
           "sexo","joven","mediano","grande","fpc","fpc1","fpc2","estrato")

num_sim <- 500
#matrices de almacenamiento de resultados
sal_puntual <- matrix(0,nrow=num_sim,ncol=5)
sal_varest1 <- matrix(0,nrow=num_sim,ncol=5)
sal_varest2 <- matrix(0,nrow=num_sim,ncol=5)

#variables de acumulación parcial de estimadores y coberturas
b0_mueprob <- b1_mueprob <- b2_mueprob <- b3_mueprob <- b4_mueprob <- 0
b0_mue2prob <- b1_mue2prob <- b2_mue2prob <- b3_mue2prob <- b4_mue2prob <- 0
ac_m1b0 <- ac_m1b1 <-ac_m1b2 <-ac_m1b3 <-ac_m1b4 <- 0
ac_m2b0 <- ac_m2b1 <-ac_m2b2 <-ac_m2b3 <-ac_m2b4 <- 0
for (isim in 1:num_sim)
{
  #se extraen los valores muestrales y se guardan en mueprue
  cont <- 1
  for (i in 1:n_upm)
      {
      #extrae la muestra aleatoria simple
      mue_usm <- sort(sample(1:tam_usm,n_usm,replace=FALSE))
      for (j in 1:n_usm)
          {
                muesim[cont,1] <- cont
                muesim[cont,2] <- mue_upm[i]
```

```
            muesim[cont,3] <- mue_usm[j]
            muesim[cont,4] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),1]
            muesim[cont,5] <- wp
            muesim[cont,7] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),13] #pobreza
            muesim[cont,8] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),9] #sexo
            muesim[cont,9] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),10] #edad1
            muesim[cont,10]<- datos[(mue_upm[i]*tam_usm+mue_usm[j]),11] #edad2
            muesim[cont,11]<- datos[(mue_upm[i]*tam_usm+mue_usm[j]),12] #edad4
            muesim[cont,12] <- f
            muesim[cont,13] <- f1
            muesim[cont,14] <- f2
            muesim[cont,15] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),15]
            cont <- cont+1
        }
    }
# muestra de conlomerados bietápicos (MAS,MAS) con los pesos calculados
#usando los datos de la población
nclus2 <- svydesign(id=~ups+uss, fpc=~fpc1+fpc2, data=muesim)
# muestra de conlomerados bietápicos (MAS,MAS)
# considerando que solo se tienen los factores de expansión
nclus1 <- svydesign(id=~ups+uss,weights=~peso,data=muesim)
#calcula los parámetros del modelo logístico usando diseño (MAS,MAS)
rclus2 <- summary(svyglm(pob~sexo+joven+mediano+grande, design=nclus2,
        family=quasibinomial()))
#calcula los parámetros del modelo logístico usando diseño (MAS,MAS)
#pero con base en los factores de expansión
rclus1 <- summary(svyglm(pob~sexo+joven+mediano+grande, design=nclus1,
        family=quasibinomial()))

#calcula las coberturas de los coeficientes beta para (MAS,MAS) con diseño complejo
if (modm1m2$coefficients[1] <= (rclus2$coefficients[1]+pt(0.975,gl)*2*rclus2$coefficients[6]))
  if (modm1m2$coefficients[1] >= (rclus2$coefficients[1]-pt(0.975,gl)*2*rclus2$coefficients[6]))
    ac_m1b0 <- ac_m1b0+1
if (modm1m2$coefficients[2] <= (rclus2$coefficients[2]+pt(0.975,gl)*2*rclus2$coefficients[7]))
  if (modm1m2$coefficients[2] >= (rclus2$coefficients[2]-pt(0.975,gl)*2*rclus2$coefficients[7]))
    ac_m1b1 <- ac_m1b1+1
if (modm1m2$coefficients[3] <= (rclus2$coefficients[3]+pt(0.975,gl)*2*rclus2$coefficients[8]))
  if (modm1m2$coefficients[3] >= (rclus2$coefficients[3]-pt(0.975,gl)*2*rclus2$coefficients[8]))
    ac_m1b2 <- ac_m1b2+1
if (modm1m2$coefficients[4] <= (rclus2$coefficients[4]+pt(0.975,gl)*2*rclus2$coefficients[9]))
  if (modm1m2$coefficients[4] >= (rclus2$coefficients[4]-pt(0.975,gl)*2*rclus2$coefficients[9]))
    ac_m1b3 <- ac_m1b3+1
if (modm1m2$coefficients[5] <= (rclus2$coefficients[5]+pt(0.975,gl)*2*rclus2$coefficients[10]))
  if (modm1m2$coefficients[5] >= (rclus2$coefficients[5]-pt(0.975,gl)*2*rclus2$coefficients[10]))
    ac_m1b4 <- ac_m1b4+1

#calcula las coberturas de los coeficientes beta para (MAS,MAS) con diseño complejo
if (modm1m2$coefficients[1] <= (rclus1$coefficients[1]+pt(0.975,gl)*2*rclus1$coefficients[6]))
  if (modm1m2$coefficients[1] >= (rclus1$coefficients[1]-pt(0.975,gl)*2*rclus1$coefficients[6]))
    ac_m2b0 <- ac_m2b0+1
if (modm1m2$coefficients[2] <= (rclus1$coefficients[2]+pt(0.975,gl)*2*rclus1$coefficients[7]))
  if (modm1m2$coefficients[2] >= (rclus1$coefficients[2]-pt(0.975,gl)*2*rclus1$coefficients[7]))
    ac_m2b1 <- ac_m2b1+1
if (modm1m2$coefficients[3] <= (rclus1$coefficients[3]+pt(0.975,gl)*2*rclus1$coefficients[8]))
  if (modm1m2$coefficients[3] >= (rclus1$coefficients[3]-pt(0.975,gl)*2*rclus1$coefficients[8]))
    ac_m2b2 <- ac_m2b2+1
if (modm1m2$coefficients[4] <= (rclus1$coefficients[4]+pt(0.975,gl)*2*rclus1$coefficients[9]))
  if (modm1m2$coefficients[4] >= (rclus1$coefficients[4]-pt(0.975,gl)*2*rclus1$coefficients[9]))
```

```
      ac_m2b3 <- ac_m2b3+1
  if (modm1m2$coefficients[5] <= (rclus1$coefficients[5]+pt(0.975,gl)*2*rclus1$coefficients[10]))
    if (modm1m2$coefficients[5] >= (rclus1$coefficients[5]-pt(0.975,gl)*2*rclus1$coefficients[10]))
      ac_m2b4 <- ac_m2b4+1

  #guarda los estimadores puntuales de los coeficientes de regresión
  sal_puntual[isim,1] <- rclus2$coefficients[1];sal_puntual[isim,2] <- rclus2$coefficients[2];
  sal_puntual[isim,3] <- rclus2$coefficients[3];sal_puntual[isim,4] <- rclus2$coefficients[4];
  sal_puntual[isim,5] <- rclus2$coefficients[5];

  #guarda los estimadores puntuales de las varianzas estimadas esquema A
  sal_varest1[isim,1] <- rclus2$coefficients[6]^2;sal_varest1[isim,2] <- rclus2$coefficients[7]^2;
  sal_varest1[isim,3] <- rclus2$coefficients[8]^2;sal_varest1[isim,4] <- rclus2$coefficients[9]^2;
  sal_varest1[isim,5] <- rclus2$coefficients[10]^2;

  #guarda los estimadores puntuales de las varianzas estimadas esquema C
  sal_varest2[isim,1] <- rclus1$coefficients[6]^2;sal_varest2[isim,2] <- rclus1$coefficients[7]^2;
  sal_varest2[isim,3] <- rclus1$coefficients[8]^2;sal_varest2[isim,4] <- rclus1$coefficients[9]^2;
  sal_varest2[isim,5] <- rclus1$coefficients[10]^2;

  b0_mueprob <- b0_mueprob + rclus2$coefficients[1]
  b1_mueprob <- b1_mueprob + rclus2$coefficients[2]
  b2_mueprob <- b2_mueprob + rclus2$coefficients[3]
  b3_mueprob <- b3_mueprob + rclus2$coefficients[4]
  b4_mueprob <- b4_mueprob + rclus2$coefficients[5]

  b0_mue2prob <- b0_mue2prob + rclus2$coefficients[1]
  b1_mue2prob <- b1_mue2prob + rclus2$coefficients[2]
  b2_mue2prob <- b2_mue2prob + rclus2$coefficients[3]
  b3_mue2prob <- b3_mue2prob + rclus2$coefficients[4]
  b4_mue2prob <- b4_mue2prob + rclus2$coefficients[5]

  muesim[,] <-0
  #extracción de upm por mas y ordenadas
  mue_upm <- sort(sample(1:tam_upm,n_upm,replace=FALSE))

}

b_mueprob <- c(b0_mueprob,b1_mueprob,b2_mueprob,b3_mueprob,b4_mueprob)/num_sim
print(b_mueprob)
b_mue2prob <- c(b0_mue2prob,b1_mue2prob,b2_mue2prob,b3_mue2prob,b4_mue2prob)/num_sim
print(b_mue2prob)

comp <- c(b_mueprob[1]/modm1m2$coefficients[1],b_mueprob[2]/modm1m2$coefficients[2],
          b_mueprob[3]/modm1m2$coefficients[3],b_mueprob[4]/modm1m2$coefficients[4],
          b_mueprob[5]/modm1m2$coefficients[5])
print(comp)
comp2 <- c(b_mue2prob[1]/modm1m2$coefficients[1],b_mue2prob[2]/modm1m2$coefficients[2],
          b_mue2prob[3]/modm1m2$coefficients[3],b_mue2prob[4]/modm1m2$coefficients[4],
          b_mue2prob[5]/modm1m2$coefficients[5])
print(comp2)

#coberturas observadas de los esquemas A y C
cober_m1 <- c(ac_m1b0,ac_m1b1,ac_m1b2,ac_m1b3,ac_m1b4)/num_sim
print(cober_m1)
cober_m2 <- c(ac_m2b0,ac_m2b1,ac_m2b2,ac_m2b3,ac_m2b4)/num_sim
print(cober_m2)
```

```
est_beta <- apply(sal_puntual,2,mean)
ventrebetas <- apply(sal_puntual,2,var)  #varianza entre estimadores puntuales

#sesgo observado relativo para la varianza estimada
sor_v1 <- c(mean((sal_varest1[,1]-ventrebetas[1])/ventrebetas[1]),
            mean((sal_varest1[,2]-ventrebetas[2])/ventrebetas[2]),
            mean((sal_varest1[,3]-ventrebetas[3])/ventrebetas[3]),
            mean((sal_varest1[,4]-ventrebetas[4])/ventrebetas[4]),
            mean((sal_varest1[,5]-ventrebetas[5])/ventrebetas[5]))

sor_v2 <- c(mean((sal_varest2[,1]-ventrebetas[1])/ventrebetas[1]),
            mean((sal_varest2[,2]-ventrebetas[2])/ventrebetas[2]),
            mean((sal_varest2[,3]-ventrebetas[3])/ventrebetas[3]),
            mean((sal_varest2[,4]-ventrebetas[4])/ventrebetas[4]),
            mean((sal_varest2[,5]-ventrebetas[5])/ventrebetas[5]))

#error cuadrático medio relativo observado para la varianza estimada
mseobs1 <- c(mean((sal_varest1[,1]-ventrebetas[1])^2/ventrebetas[1]^2),
            mean((sal_varest1[,2]-ventrebetas[2])^2/ventrebetas[2]^2),
            mean((sal_varest1[,3]-ventrebetas[3])^2/ventrebetas[3]^2),
            mean((sal_varest1[,4]-ventrebetas[4])^2/ventrebetas[4]^2),
            mean((sal_varest1[,5]-ventrebetas[5])^2/ventrebetas[5]^2))

mseobs2 <- c(mean((sal_varest2[,1]-ventrebetas[1])^2/ventrebetas[1]^2),
            mean((sal_varest2[,2]-ventrebetas[2])^2/ventrebetas[2]^2),
            mean((sal_varest2[,3]-ventrebetas[3])^2/ventrebetas[3]^2),
            mean((sal_varest2[,4]-ventrebetas[4])^2/ventrebetas[4]^2),
            mean((sal_varest2[,5]-ventrebetas[5])^2/ventrebetas[5]^2))

#escribe los resultados en archivo
write.csv(sal_puntual,"mmb48-500sim.csv")
write.csv(sal_varest1,"mmv148-500sim.csv")
write.csv(sal_varest2,"mmv248-500sim.csv")
```

{Regresión logística: código en R para extracción de muestra (H,ppt,MAS)}

```
#carga el paquete de muestreo
library(survey)
#carga el paquete que permite extraer muestras ppt
library(pps)
#información poblacional
datos <- read.csv("v3bcshogar.csv",header=TRUE)
names(datos)

s_phog <- cumsum(datos[,8])
#pi_phog <- datos[,8]/max(s_phog)

#calcula las betas poblacionales para el modelo
m1m2 <- glm(pob~sexo+edad1+edad2+edad4,family=binomial,data=datos)
modm1m2  <-summary(m1m2)

#número de elementos por estrato y tamaño de estratos
N1 <- sum(datos[,15]==1);N2 <- sum(datos[,15]==2);N3 <- sum(datos[,15]==3);
N4 <- sum(datos[,15]==4);N5 <- sum(datos[,15]==5); NT <- N1+N2+N3+N4+N5
W1 <- N1/NT;W2 <- N2/NT;W3 <- N3/NT;W4 <- N4/NT;W5 <- N5/NT;

#número de upm en población y de usm dentro de upm
NH1 <- datos[N1,2];NH2 <- datos[N2,2];NH3 <- datos[N3,2];
NH4 <- datos[N4,2];NH5 <- datos[N5,2];NHT <- NH1+NH2+NH3+NH4+NH5
n_upm   <- c(4,2,9,7,2)*2 #conglomerados en muestra por estrato
tam_usm <- 10  #elementos por conglomerado
n_usm   <- 5   #elementos en muestra por upm en muestra
n <- n_usm*n_upm

#calcula una medida de tamaño para las upm para extraer con pips
mos_upm <- c();cont <- 1;cont_ext <- 1
for (i in 1:907)
    {
    temp <- 0
    for (j in 1:10)
        {
           temp <- temp+datos[cont,8]
           cont <- cont+1
        }
    mos_upm[cont_ext] <- temp
    cont_ext <- cont_ext+1
    }

totpers <- sum(mos_upm)
pi_phog <- datos[,8]/totpers


#número de upm en muestra y de usm dentro de upm
tam_upm <- 907
#fracciones de muestreo y peso muestral
f2 <- n_usm/tam_usm
gl <-sum(n_upm*n_usm)

#extracción de upm por mas y ordenadas
mc1<-ppss(mos_upm[1:NH1],n_upm[1])
mc2<-ppss(mos_upm[(NH1+1):(NH1+NH2)],n_upm[2])+(NH1)
mc3<-ppss(mos_upm[(NH1+NH2+1):(NH1+NH2+NH3)],n_upm[3])+(NH1+NH2)
```

```
mc4<-ppss(mos_upm[(NH1+NH2+NH3+1):(NH1+NH2+NH3+NH4)],n_upm[4])+(NH1+NH2+NH3)
mc5<-ppss(mos_upm[(NH1+NH2+NH3+NH4+1):(NHT-1)],n_upm[5])+(NH1+NH2+NH3+NH4)
mue_upm <- sort(c(mc1,mc2,mc3,mc4,mc5))

#matriz para almacenar los datos de la muestra
muesim <- as.data.frame(matrix(0,nrow=sum(n_upm)*n_usm,ncol=15))
dimnames(muesim)[[2]] <- c("consec","ups","uss","otro","peso","pesocalc","pob","sexo",
                    "joven","mediano","grande","fpc","fpc1","fpc2","estrato")

num_sim <- 500
#matrices de almacenamiento de resultados
sal_puntual <- matrix(0,nrow=num_sim,ncol=5)
sal_varest1 <- matrix(0,nrow=num_sim,ncol=5)

#variables de acumulación parcial de estimadores y coberturas
b0_mueprob <- b1_mueprob <- b2_mueprob <- b3_mueprob <- b4_mueprob <- 0
b0_mue2prob <- b1_mue2prob <- b2_mue2prob <- b3_mue2prob <- b4_mue2prob <- 0
ac_m1b0 <- ac_m1b1 <-ac_m1b2 <-ac_m1b3 <-ac_m1b4 <- 0
ac_m2b0 <- ac_m2b1 <-ac_m2b2 <-ac_m2b3 <-ac_m2b4 <- 0

for (isim in 1:num_sim)
{
  cont <- 1
  for (i in 1:sum(n_upm))
      {
      #extrae la muestra dentro de UPM
      mue_usm <- sort(sample(1:tam_usm,n_usm,replace=FALSE))
      for (j in 1:n_usm)
          {
                muesim[cont,1] <- cont
                muesim[cont,2] <- mue_upm[i]
                muesim[cont,3] <- mue_usm[j]
                muesim[cont,4] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),1]
                muesim[cont,15] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),15]
                muesim[cont,5] <- 1/(n_upm[muesim[cont,15]]*mos_upm[mue_upm[i]]/totpers)*(1/f2)
                muesim[cont,7] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),13] #pobreza
                muesim[cont,8] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),9] #sexo
                muesim[cont,9] <- datos[(mue_upm[i]*tam_usm+mue_usm[j]),10] #edad1
                muesim[cont,10]<- datos[(mue_upm[i]*tam_usm+mue_usm[j]),11] #edad2
                muesim[cont,11]<- datos[(mue_upm[i]*tam_usm+mue_usm[j]),12] #edad4
                #muesim[cont,12] <- f
                #muesim[cont,13] <- f1
                muesim[cont,14] <- f2
                cont <- cont+1
          }
      }
  # muestra (H,ppt,MAS) con los pesos calculados
  nstclus2 <- svydesign(id=~ups+uss, weights=~peso, strata=~estrato, data=muesim)
  # calcula los parámetros del modelo logístico usando diseño (H,ppt,MAS)
  rstclus2 <- summary(svyglm(pob~sexo+joven+mediano+grande, design=nstclus2,
          family=quasibinomial()))

  #calcula las coberturas de los coeficientes beta para (H,ppt,MAS) con diseño complejo
  if (modm1m2$coefficients[1] <= (rstclus2$coefficients[1]+pt(0.975,gl)*2*rstclus2$coefficients[6]))
    if (modm1m2$coefficients[1] >= (rstclus2$coefficients[1]-pt(0.975,gl)*2*rstclus2$coefficients[6]))
      ac_m1b0 <- ac_m1b0+1
  if (modm1m2$coefficients[2] <= (rstclus2$coefficients[2]+pt(0.975,gl)*2*rstclus2$coefficients[7]))
    if (modm1m2$coefficients[2] >= (rstclus2$coefficients[2]-pt(0.975,gl)*2*rstclus2$coefficients[7]))
```

```
      ac_m1b1 <- ac_m1b1+1
  if (modm1m2$coefficients[3] <= (rstclus2$coefficients[3]+pt(0.975,gl)*2*rstclus2$coefficients[8]))
    if (modm1m2$coefficients[3] >= (rstclus2$coefficients[3]-pt(0.975,gl)*2*rstclus2$coefficients[8]))
      ac_m1b2 <- ac_m1b2+1
  if (modm1m2$coefficients[4] <= (rstclus2$coefficients[4]+pt(0.975,gl)*2*rstclus2$coefficients[9]))
    if (modm1m2$coefficients[4] >= (rstclus2$coefficients[4]-pt(0.975,gl)*2*rstclus2$coefficients[9]))
      ac_m1b3 <- ac_m1b3+1
  if (modm1m2$coefficients[5] <= (rstclus2$coefficients[5]+pt(0.975,gl)*2*rstclus2$coefficients[10]))
    if (modm1m2$coefficients[5] >= (rstclus2$coefficients[5]-pt(0.975,gl)*2*rstclus2$coefficients[10]))
      ac_m1b4 <- ac_m1b4+1

  sal_puntual[isim,1] <- rstclus2$coefficients[1];sal_puntual[isim,2] <- rstclus2$coefficients[2];
  sal_puntual[isim,3] <- rstclus2$coefficients[3];sal_puntual[isim,4] <- rstclus2$coefficients[4];
  sal_puntual[isim,5] <- rstclus2$coefficients[5];

  sal_varest1[isim,1] <- rstclus2$coefficients[6]^2;sal_varest1[isim,2] <- rstclus2$coefficients[7]^2;
  sal_varest1[isim,3] <- rstclus2$coefficients[8]^2;sal_varest1[isim,4] <- rstclus2$coefficients[9]^2;
  sal_varest1[isim,5] <- rstclus2$coefficients[10]^2;

  b0_mueprob <- b0_mueprob + rstclus2$coefficients[1]
  b1_mueprob <- b1_mueprob + rstclus2$coefficients[2]
  b2_mueprob <- b2_mueprob + rstclus2$coefficients[3]
  b3_mueprob <- b3_mueprob + rstclus2$coefficients[4]
  b4_mueprob <- b4_mueprob + rstclus2$coefficients[5]

  muesim[,] <-0
  #extracción de upm por mas y ordenadas
  mc1<-ppss(mos_upm[1:NH1],n_upm[1])
  mc2<-ppss(mos_upm[(NH1+1):(NH1+NH2)],n_upm[2])+(NH1)
  mc3<-ppss(mos_upm[(NH1+NH2+1):(NH1+NH2+NH3)],n_upm[3])+(NH1+NH2)
  mc4<-ppss(mos_upm[(NH1+NH2+NH3+1):(NH1+NH2+NH3+NH4)],n_upm[4])+(NH1+NH2+NH3)
  mc5<-ppss(mos_upm[(NH1+NH2+NH3+NH4+1):(NHT-1)],n_upm[5])+(NH1+NH2+NH3+NH4)
  mue_upm <- sort(c(mc1,mc2,mc3,mc4,mc5))
}
#promedio de estimadores puntuales
b_mueprob <- c(b0_mueprob,b1_mueprob,b2_mueprob,b3_mueprob,b4_mueprob)/num_sim
print(b_mueprob)

comp <- c(b_mueprob[1]/modm1m2$coefficients[1],b_mueprob[2]/modm1m2$coefficients[2],
          b_mueprob[3]/modm1m2$coefficients[3],b_mueprob[4]/modm1m2$coefficients[4],
          b_mueprob[5]/modm1m2$coefficients[5])
print(comp)

#cobertura observada
cober_m1 <- c(ac_m1b0,ac_m1b1,ac_m1b2,ac_m1b3,ac_m1b4)/num_sim
print(cober_m1)

est_beta <- apply(sal_puntual,2,mean)
ventrebetas <- apply(sal_puntual,2,var)  #varianza entre estimadores puntuales

#sesgo observado relativo para V-est
sor_v1 <- c(mean((sal_varest1[,1]-ventrebetas[1])/ventrebetas[1]),
            mean((sal_varest1[,2]-ventrebetas[2])/ventrebetas[2]),
            mean((sal_varest1[,3]-ventrebetas[3])/ventrebetas[3]),
            mean((sal_varest1[,4]-ventrebetas[4])/ventrebetas[4]),
            mean((sal_varest1[,5]-ventrebetas[5])/ventrebetas[5]))

#escribe los resultados en archivo
```

```
write.csv(sal_puntual,"hppsmb48-500sim.csv")
write.csv(sal_varest1,"hppsmv148-500sim.csv")
```

Referencias

[1] Agresti, A., *Categorical Data Analysis*, John Wiley and Sons, New York, 1990.

[2] Albert, A. and Anderson, J.A., *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika **2**71, (1984), 1-10.

[3] Bebbington, A. and Smith, T.M.F, *The Effect of Survey Design on Multivariate Analysis* in The Analysis of Survey Data (Vol. 2), eds. O'Muircheartaigh, C. and Payne, C., John Wiley and Sons, New York, 1978.

[4] Binder, D.A., *On the variances of asymptotically normal estimators from complex surveys*, International Statistical Review 51, (1983), 279-292.

[5] Cochran, W. *Técnicas de Muestreo*, Ed. CECSA, México, 1986.

[6] Des Raj, *Teoría del Muestreo*, Fondo de Cultura Económica, México, 1992.

[7] Godínez, F. y Ramírez, G., *Intervalos de confianza en el modelo de regresión logística, en presencia de separación de los datos y colinealidad entre las variables explicatorias. XX Foro Nacional de estadística*, Universidad Autónoma de Guerrero, Acapulco, 2006.

[8] *Handbook of Statistics 9: Computational Statistics*, ed. by C.R. Rao, Amsterdam: North-Holland, 1993.

[9] Heinze, G. and Schemper, M., *A solution to the problem of separation in logistic regression*, Statistics in Medicine 21, (1984), 2409-2419.

[10] Hidiroglou, M.A., *Estimation of regression parameters for finite populations: A Monte-Carlo Study*, Journal of Official Statistics, Vol. II, (1986b), 3-11.

[11] Hosmer, D.W. and Lemeshow, S., *Applied Logistic Regression*, Second Edition, John Wiley and Sons, New York, 2000.

[12] Horvitz, D.G. and Thompson, D.J., *A generalization of sampling without replacement from a finite universe*, Journal of the American Statistical Association, Vol. 47, No. 260, (Dec. 1952), 663-685.

[13] le Cessie, S. and van Houwelingen, J.C., *Ridge estimators in logistic regression*, Applied Statistics 41(1), (1992), 191-201.

[14] Lehtonen, R. and Pahkinen, E.J., *Practical Methods for Design and Analysis of Complex Surveys*, John Wiley and Sons, New York, 1995.

[15] Liang, K.Y. and Zeger, S.L., *Longitudinal data analysis using generalized linear models*, Biometrika **2**73, (1986), 13-22.

[16] Liu, K., *Using Liu-type estimator to combat collinearity*, Communications in Statistics, Theory and Methods, Vol. 32, No. 5, (2003), 1009-1020.

[17] Lohr, S., *Muestreo: Diseño y Análisis*, International Thomson Editores, México, 2000.

[18] Lumley, T., *Survey Package*, R-Statistical Sofrware.

[19] McCulloch, C. E. and Searle, S.R., *Generalized, Linear, and Mixed Models*, John Wiley and Sons, Wiley Series in Probability and Statistics, New York, 2001.

[20] Méndez, I. y Romero, P., *Comparación de estimadores de varianza para diseños de muestra bietápicos. XVII Foro Nacional de estadística*, Universidad de las Américas, Puebla, 2002.

[21] Mood, A.M., Graybill, F.A. and Boes, D.C., *Introduction to the Theory of Statistics*, Third Edition, McGraw-Hill, 1985.

[22] Nordberg, L., *Generalized linear modeling of sample survey data*, Journal of Official Statistics, Vol. 5, No. 3, (1989), 223-239.

[23] Santner, T.J. and Duffy, D.E., *A note on A. Albert and J.A. Anderson´s conditions for the existence of maximum likelihood estimates in logistic regression models*, Biometrika 73, (1986), 755-758.

[24] Särndal, C.E., B. Swensson and J.H. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, New York, 1992.

[25] *Stata Survey Data Reference Manual*, Release 8. Stata Corp., 2003, Stata Statistical Software: Release 8.0.

[26] Research Triangle Institute (2001)., *SUDAAN User´s Manual*, Release 8.0 Research Triangle Park, NC: Research Triangle Institute.

[27] Thompson, M.E., *Theory of Sample Surveys*, Chapman and Hall, London, 1997.

[28] Tillè, Yves, *Sampling Algorithms*, Springer-Verlag, New York, 2006.

[29] Tinajero, B.M. y Eslava, G.G., *Calibración en muestreo: una aplicación a la encuesta nacional de ingresos y gastos en los hogares 1992 y 1996*, Monografías, IIMAS, UNAM, Vol. 9, No. 21, (2000).

[30] Valliant, R., Dorfman, A. and Royall, R. *Finite Population Sampling and inference: a prediction approach*, John Wiley ans Sons, New York, 2000.

[31] Wolter, K.M., *Introduction to Variance Estimation*, Springer-Verlag, New York, 1985.