

*Estimación de un total para un muestreo de tres
ocasiones*

LAURA BUSTAMANTE ATEHORTÚA
ESTADÍSTICA



UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA
BOGOTÁ, D.C.
JUNIO DE 2018

*Estimación de un total para un muestreo de tres
ocasiones*

LAURA BUSTAMANTE ATEHORTÚA
ESTADÍSTICA

DISERTACIÓN PRESENTADA PARA OPTAR AL TÍTULO DE
MAGISTER EN CIENCIAS - ESTADÍSTICA

DIRECTOR
LUZ MERY GONZÁLEZ GARCÍA, PH.D.
DOCTORA EN ESTADÍSTICA



UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA
BOGOTÁ, D.C.
JUNIO DE 2018

Título en español

Estimación de un total para un muestreo de tres ocasiones.

Title in English

Total estimation for sampling on three occasions.

Resumen: En este trabajo se presenta una adecuación de la metodología expuesta para el estimador del total en un muestreo de dos ocasiones a un diseño muestral de tres ocasiones. Con una variación en la aproximación realizada de las variables, utilizando ecuaciones de estimación generalizadas. Además, se presenta un ejemplo práctico aplicando la metodología propuesta al conjunto de datos disponibles de la Encuesta Longitudinal Colombiana de la Universidad de Los Andes, adecuándolo a la naturaleza de los datos. Asimismo, mediante un proceso de simulación se evaluó el estimador propuesto bajo un muestreo aleatorio simple, considerando diferentes escenarios y comparándolo con el estimador de Horvitz-Thompson.

Abstract: In this work, we present an adaptation of the total estimator methodology from a sampling on two occasions to a sampling on three occasions. Also, a variation in the approximation to the study variable was made using generalized estimating equations. In addition, a practical example is presented applying the proposed methodology to the set of data available from the Colombian Longitudinal Survey of Andes University, adapting it to the nature of the data. The simulation process was evaluated using a simple random sampling, under different scenarios and comparing it with the Horvitz-Thompson estimator.

Palabras clave: Muestreo en dos ocasiones, Ecuaciones de estimación generalizadas, Muestreo aleatorio simple, Estimador de Horvitz-Thompson.

Keywords: Sampling on two occasions, Generalized estimating equations, Simple random sampling, Horvitz-Thompson estimator.

Índice general

Índice general	II
Índice de tablas	IV
Índice de figuras	V
1. Introducción	1
2. Marco Conceptual	4
2.1. Diseño Muestral	4
2.1.1. Diseño muestral en etapas	5
2.1.2. Diseño muestral en fases	6
2.1.3. Muestreo en dos ocasiones	7
2.2. No Respuesta	9
2.2.1. Submuestreo de no respondientes	9
2.2.2. Imputación múltiple	10
2.3. Ecuaciones de Estimación Generalizadas	11
3. Muestreo en tres ocasiones	14
3.1. Caso particular: Muestreo Aleatorio Simple	19
4. Encuesta Longitudinal Colombiana de la Universidad de los Andes	21
4.1. Antecedentes	21
4.2. Análisis descriptivo: Población ocupada	25
4.3. Análisis del salario 2010, 2013 y 2016	27
4.4. Estimación del salario total para el año 2016	30
4.5. Conclusiones	32

5. Simulación	36
5.1. Conclusiones	38
6. Conclusiones y trabajo futuro	42
A. Demostraciones	44
A.1. Esperanza	45
A.2. Varianza	46
A.3. Covarianza	50
B. Tamaños de muestra para la tercera ocasión	53
C. Código desarrollado	55
D. Bibliografía	58

Índice de tablas

2.1. Técnicas de imputación implementadas en el algoritmo MICE, de acuerdo al tipo de variable.	12
3.1. Tamaños de muestra para s_{m_2} , s_{m_3} y s_n , obtenidos en la tercera ocasión. Ver anexo B.	19
4.1. Total de encuestas esperadas y efectivas por región dentro de la zona urbana para el año 2010.	22
4.2. Total de encuestas esperadas y efectivas por subregión dentro de la zona rural para el año 2010.	23
4.3. Cobertura para el año 2013 de acuerdo a la zona. (1): Total de hogares de seguimiento de la muestra 2010. (2): Total de hogares de (1) encuestados. (3): Número de encuestas en 2013. (4): Total de encuestas en 2013 teniendo en cuenta la migración urbano/rural y rural/urbano.	24
4.4. Cobertura para el año 2016 de acuerdo a la zona. (1): Total de hogares encuestados en 2010 con interés de seguimiento para el 2016. (2): Total de hogares objeto de seguimiento en 2013 y encuestados. (3): Total de hogares de (1) encuestados. (4): Total de encuestas completas realizadas en 2016 teniendo en cuenta la migración urbano/rural y rural/urbano.	24
4.5. Distribución de la población ocupada por estado civil. Años: 2010, 2013 y 2016. ELCA.	25
4.6. Tipo de ocupación en la población ocupada. Años: 2010, 2013 y 2016. ELCA.	26
4.7. Estimaciones de los parámetros para el modelo cuasi gamma aplicado a los individuos de la ELCA con observaciones en los años 2010 y 2016 y/o 2013.	31
4.8. Estimaciones de los parámetros para el modelo cuasi gamma aplicado a los individuos de la ELCA con observaciones en los años 2013 y 2016.	31
4.9. Valor de los $\hat{w}'s$ para cada imputación, con los datos de la ELCA.	32
A.1. Probabilidades de inclusión de primer orden, segundo orden y Δ_{kl} para cada muestra.	44

Índice de figuras

1.1. Comportamiento de las muestras en la primera y segunda ocasión.	2
1.2. Comportamiento de las muestras en la tercera ocasión. Muestra s_{m_2} seleccionada de s_a , muestra s_{m_3} seleccionada de s_u y muestra s_n seleccionada de $s^c = (s_a \cup s_u)^c$	3
3.1. Muestra s_a obtenida en la primera ocasión.	14
3.2. Muestras de la segunda ocasión: s_{m_1} seleccionada de s_a y muestra s_u seleccionada de $s_a^c = U - s_a$	15
3.3. Comportamiento de las muestras en la tercera ocasión. Muestra s_{m_2} seleccionada de s_a , muestra s_{m_3} seleccionada de s_u y muestra s_n seleccionada de $s^c = (s_a \cup s_u)^c$	16
4.1. Distribución de la edad en la población ocupada. Años: 2010, 2013 y 2016. ELCA.	26
4.2. Distribución del salario recibido en la población ocupada. Años: 2010, 2013 y 2016. ELCA.	27
4.3. Distribución del salario recibido en la población ocupada, una vez indexados los salarios de los años 2013 y 2016 con base al IPC. Años: 2010, 2013 y 2016. ELCA.	28
4.4. Salario en los años 2010, 2013 y 2016. ELCA.	28
4.5. Cociente de los salarios de los años 2013 y 2016 con respecto al año 2010. ELCA.	29
4.6. Convergencia de las cadenas de Markov para la ELCA 2010, 2013 y 2016.	29
4.7. Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2010 y 2016 y/o 2013.	34
4.8. Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2013 y 2016.	35
5.1. Sesgo relativo de \widehat{t}_Y variando error muestral y proporción de muestra emparejada (μ) para cada uno de los universos.	39

5.2. Contribución relativa del sesgo al error cuadrático medio de \widehat{t}_Y variando error muestral y proporción de muestra emparejada (μ) para cada uno de los universos.	40
5.3. Coeficiente de variación para \widehat{t}_π , \widehat{t}_Y y \widehat{t}_Y variando error muestral y proporción de muestra emparejada (μ) para cada uno de los universos.	41
B.1. Muestras de la tercera ocasión relacionadas con las ocasiones anteriores.	53

CAPÍTULO 1

Introducción

Considerando que los estudios longitudinales dan la posibilidad de observar el cambio o evolución en las dinámicas y decisiones de los individuos de interés, el Centro de Estudios Sobre Desarrollo Económico (CEDE) y la Facultad de Economía de la Universidad de los Andes crearon la Encuesta Longitudinal Colombiana de la Universidad de los Andes (ELCA) (CEDE & Facultad Economía U.Andes, 2011), buscando realizar un seguimiento durante 12 años a los hogares e individuos colombianos que fueron seleccionados de acuerdo al diseño muestral descrito en CEDE & Facultad Economía U. Andes (2010).

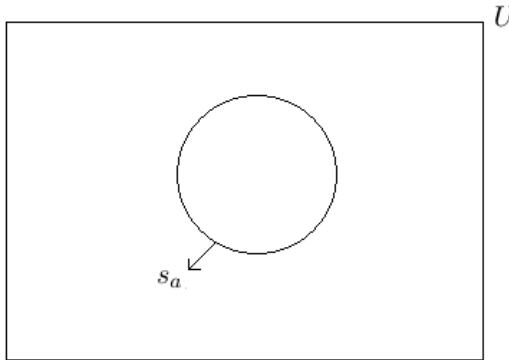
Con tres levantamientos de información hasta el momento, la ELCA ha abordado líneas de investigación en los ámbitos social, demográfico y económico en diferentes grupos poblacionales. Sin embargo, el interés del presente trabajo estuvo enfocado en las preguntas que caracterizaron la fuerza laboral de hogares colombianos, pertenecientes a los estratos uno al cuatro, de la zona urbana en las regiones Atlántica, Central, Oriental, Pacífica y Bogotá.

En estudios de seguimiento como la ELCA, es frecuente que los sujetos seleccionados no se logren observar en todas las ocasiones, causando que los investigadores remuevan individuos por falta de continuidad de los mismos o realicen estimaciones enfocadas en la información de un solo levantamiento, perdiendo de vista el total de información que ha sido recolectada para la variable de estudio. Ésto se puede ver en publicaciones como CEDE & Facultad Economía U. Andes (2011),(2014) y Fuertes et al. (2016), donde los análisis fueron hechos de manera transversal y no incluyeron, en algunos casos, el diseño muestral bajo el cual se basó la recolección de información.

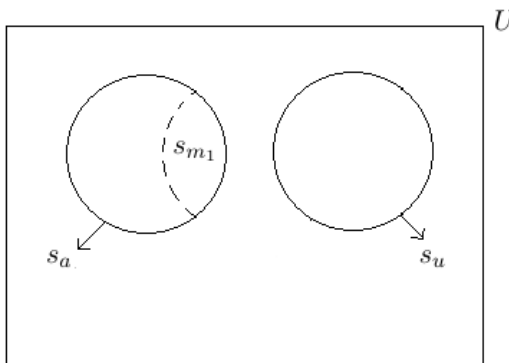
Por tal motivo, buscando no eliminar individuos o datos ya recolectados, y considerando que en la ELCA hubo una proporción de individuos con información de años anteriores y otros no, para el trabajo se utilizó la metodología desarrollada para un muestreo de dos ocasiones de Särndal (1992)[p.368] extendiéndola a tres ocasiones, la cual tiene en cuenta las proporciones de individuos que coinciden y los pesos muestrales para el cálculo de diferentes estimadores, en este caso el estimador del salario total.

Así, sea U una población finita de donde se obtienen las muestras. Para la primera ocasión se selecciona una muestra s_a , de donde se observa la variable de interés, Figura 1.1(a). Para la segunda ocasión se seleccionan dos muestras independientes, la primera s_{m_1}

considerada un subconjunto de la muestra s_a , y s_u una muestra de $s_a^c = U - s_a$, Figura 1.1(b).



(a) Muestra s_a obtenida en la primera ocasión.



(b) Muestras de la segunda ocasión: s_{m_1} seleccionada de s_a y muestra s_u seleccionada de $s_a^c = U - s_a$.

FIGURA 1.1. Comportamiento de las muestras en la primera y segunda ocasión.

De aquí se construyen dos estimadores insesgados del total para la segunda ocasión, el primero \hat{t}_1 que utiliza la información de s_a (primera ocasión) y s_{m_1} (primera y segunda ocasión), y \hat{t}_2 que se forma a partir de s_u (segunda ocasión). Tanto \hat{t}_1 como \hat{t}_2 son utilizados para construir el estimador insesgado del total quedando, $\hat{t}_y = w_1\hat{t}_1 + w_2\hat{t}_2$ donde w_1 y w_2 son pesos constantes no negativos, tal que $w_1 + w_2 = 1$.

Ahora, como los datos utilizados de la ELCA fueron recogidos en *tres ocasiones*, la última ocasión se incluyó a la metodología ya descrita, mediante el enfoque de un muestreo en fases. Luego, el estimador del total para la tercera ocasión queda *actualizado* con la información de años anteriores.

Tres muestras independientes son obtenidas en la tercera ocasión. La primera de ellas denotada por s_{m_2} , es considerada un subconjunto de s_a y por consiguiente sus individuos cuentan con información de la primera y tercera ocasión. Algunos de ellos, incluso tendrán información de la segunda ocasión debido a que cayeron anteriormente en la muestra

s_{m_1} . La segunda muestra es un subconjunto de s_u y se denota como s_{m_3} , sus individuos cuentan con información de la segunda y tercera ocasión. Finalmente, la muestra s_n es un subconjunto de $s^c = (s_a \cup s_u)^c$ y por lo tanto sus individuos solo tienen información de la tercera ocasión. Ver Figura 1.2.

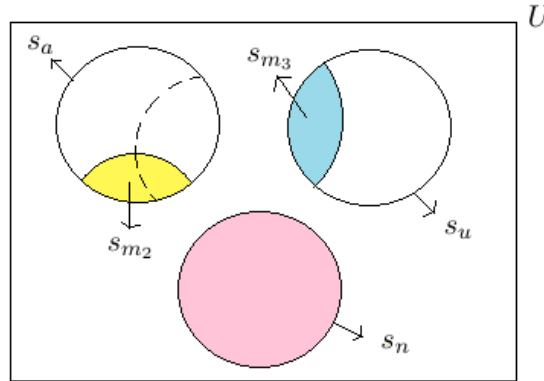


FIGURA 1.2. Comportamiento de las muestras en la tercera ocasión. Muestra s_{m_2} seleccionada de s_a , muestra s_{m_3} seleccionada de s_u y muestra s_n seleccionada de $s^c = (s_a \cup s_u)^c$.

Algunos estudios como la Encuesta de Población Actual que realiza la Oficina del Censo de Estados Unidos (Cheng & et. al, 2017) o la Encuesta de Fuerza Laboral llevada a cabo en Australia (Bell, 2001), incorporan para el cálculo de estimaciones la información de las muestras que coinciden en diferentes años, y a su vez la información que brinda aquella muestra que no coincide con anteriores, mediante una combinación de las mismas.

Por último, dado que los datos usados eran incompletos debido a la no respuesta en la variable de mayor interés (salario), se utilizó la técnica de imputación múltiple descrita por Rubin (1987, 1977, 1978) para tener un conjunto de datos completos. Ésta permite introducir el conocimiento del investigador para la conformación del modelo a usar en la imputación de los datos, incrementa la eficiencia de las estimaciones y a su vez, posibilita análisis de sensibilidad de las inferencias de los modelos.

En el capítulo 2 se presenta una breve revisión de conceptos y metodologías necesarias para el desarrollo del estimador del total en un diseño muestral de tres ocasiones. Posteriormente, en el capítulo 3 se tiene el desarrollo del estimador para un total en un muestreo de tres ocasiones. En el capítulo 4, se muestra la información relacionada con el conjunto de datos de la ELCA, antecedentes de la encuesta, análisis descriptivos, implementación de la técnica de imputación múltiple y la aplicación del estimador propuesto en el capítulo 3. En el capítulo 5, se presenta un trabajo de simulación que evidencia la calidad del estimador propuesto y a su vez, una comparación con el estimador clásico de Horvitz-Thompson. Finalmente, en el capítulo 6 se muestran conclusiones y recomendaciones para trabajo futuro.

Marco Conceptual

2.1. Diseño Muestral

Sea $U = \{1, \dots, k, \dots, N\}$ una población finita de tamaño N y Y una variable de estudio, donde Y_k denota el valor de Y para el k -ésimo elemento de la población. Con interés en estimar el total poblacional de Y ,

$$t = \sum_U Y_k,$$

se selecciona un subconjunto de la población conocido como muestra y notado por s , donde el valor de Y_k es observado para cada elemento k seleccionado.

Aunque la muestra puede ser cualquier subconjunto de la población U , se considera que ésta es una realización de un esquema de selección probabilístico que tiene asociada una función $p(\cdot)$, la cual asigna una probabilidad de selección bajo el esquema utilizado. Esta función se conoce como diseño muestral.

Para un diseño muestral dado $p(\cdot)$, sea s una realización de la variable aleatoria S con función de probabilidad definida por $p(\cdot)$ y \mathcal{S} el conjunto de todas las muestras. Se tiene que

$$P(S = s) = p(s),$$

para cualquier $s \in \mathcal{S}$. Además,

$$p(s) \geq 0, \quad \text{para } s \in \mathcal{S}$$

y

$$\sum_{s \in \mathcal{S}} p(s) = 1.$$

El número de elementos en s se conoce como tamaño muestral y se denota por n_s .

Sea $p(s)$ la probabilidad de seleccionar s bajo un diseño muestral dado. La inclusión del elemento k en la muestra es un evento aleatorio definido por la variable aleatoria I_k , como

$$I_k = \begin{cases} 1 & \text{si } k \in S; \\ 0 & \text{en otro caso.} \end{cases}$$

La probabilidad que el elemento k sea incluido en la muestra se denota por π_k ,

$$\pi_k = P(k \in S) = P(I_k = 1) = \sum_{s \ni k} p(s).$$

La probabilidad que dos elementos k y l se incluyan en la muestra se denota como π_{kl} y está dada por

$$\pi_{kl} = P(k \& l \in S) = P(I_k I_l = 1) = \sum_{s \ni k \& l} p(s).$$

Note que $\pi_{kl} = \pi_{lk}$ y $\pi_{kk} = \pi_k$.

Como los estimadores de interés pueden ser expresados en función de I_k , sus propiedades son importantes y pueden ser consultadas en Särndal (1992, p.36).

Para la estimación, se utiliza el π -estimador o estimador de Horvitz-Thompson que es un estimador insesgado para el total, y es dado por

$$\hat{t}_\pi = \sum_s \frac{Y_k}{\pi_k}.$$

Sus propiedades se pueden ver en Särndal (1992, p.43).

2.1.1.1. Diseño muestral en etapas

Se caracteriza por la selección de una muestra probabilística de conglomerados. No obstante, para disminuir costos, en lugar de encuestar todos los elementos dentro de los conglomerados seleccionados, se realiza una muestra de elementos dentro de éstos.

Se particiona la población finita U en N_I subpoblaciones disyuntas, conocidas como Unidades Primarias de Muestreo (UPM's), denotadas por $U_1, \dots, U_i, \dots, U_{N_I}$ y $U_I = \{1, \dots, i, \dots, N_I\}$ el conjunto de todas ellas. N_i el tamaño de U_i y por tanto, $N = \sum_{U_I} N_i$.

El procedimiento en el caso más sencillo, diseño muestral en dos etapas, se define como (Särndal, 1992, p.126):

- **Primera etapa.** Una muestra s_I de UPM's es seleccionada de U_I , de acuerdo a un diseño $p_I(\cdot)$.
- **Segunda etapa.** Para cada $i \in s_I$, una muestra de elementos s_i se selecciona de U_i , de acuerdo a un diseño $p_i(\cdot | s_I)$. Así, la muestra de elementos se compone por $s = \bigcup_{i \in s_I} s_i$.

Bajo el diseño muestral en etapas, dos propiedades son requeridas:

- **Invarianza** de la segunda etapa. El diseño muestral a realizar en la i -ésima UPM es independiente de las muestras seleccionadas de UPM's.

$$p_i(\cdot | s_I) = p_I(\cdot)$$

- **Independencia.** El diseño muestral aplicado en la i -ésima UPM es independiente del diseño muestral aplicado en la j -ésima UPM ($i \neq j$).

$$P\left(\bigcup_{i \in s_I} s_i \mid s_I\right) = \prod_{i \in s_I} P(s_i \mid s_I)$$

En caso de no cumplirse alguna de las dos propiedades anteriores, se tendría un diseño muestral en fases.

Para un diseño muestral de r etapas ($r \geq 2$), un estimador insesgado para el total poblacional t está dado por (Särndal, 1992)[p.144]

$$\hat{t} = \sum_{s_I} \frac{\hat{t}_i}{\pi_{Ii}}$$

donde $E(\hat{t}_i \mid s_I) = t_i$. Denotando $\check{t}_i = t_i/\pi_{Ii}$, la varianza se puede escribir como

$$V_{rst}(\hat{t}) = \sum \sum_{U_I} \Delta_{Iij} \check{t}_i \check{t}_j + \sum_{U_I} \frac{V_i}{\pi_{Ii}}$$

donde el primer término de la derecha representa la varianza debido a la primera etapa y el segundo término combina la varianza contribuida por el resto de etapas. Finalmente, un estimador insesgado para la varianza está dado por

$$\hat{V}_{rst}(\hat{t}) = \sum \sum_{s_I} \check{\Delta}_{Iij} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}},$$

donde $\check{\Delta}_{Iij} = \Delta_{Iij}/\pi_{Iij}$.

2.1.2. Diseño muestral en fases

La técnica de muestreo en dos fases se convierte en una solución cuando surgen problemas como tener un marco muestral con poca información de la población o hay presencia de no respuesta. Su procedimiento se define como (Särndal, 1992, p.344):

- **Fase I.** Se selecciona una muestra s_a de U de acuerdo a un diseño muestral $p_a(\cdot)$, de tamaño n_{s_a} . Para todo $k \in s_a$ se recoge la información auxiliar requerida.

Las correspondientes probabilidades de inclusión están dadas por:

$$\pi_{ak} = P(k \in s_a), \quad \pi_{akl} = P(k, l \in s_a), \quad k \neq l.$$

- **Fase II.** A partir de la información auxiliar recolectada en la fase I, se selecciona una muestra s de s_a de acuerdo a un diseño muestral $p(\cdot \mid s_a)$. La variable de interés Y_k se observa para $k \in s$.

Las probabilidades de inclusión bajo este diseño se denotan como:

$$\pi_{k|s_a} = P(k \in s \mid k \in s_a), \quad \pi_{kl|s_a} = P(k, l \in s \mid k, l \in s_a), \quad k \neq l.$$

Una vez definidas las propiedades del diseño, se plantea como candidato natural para el estimador del total t , el π -estimador,

$$\hat{t}_\pi = \sum_s \frac{Y_k}{\pi_k},$$

donde

$$\pi_k = P(k \in s) = E(I_k) = \sum_{s_a \ni k} p_a(s_a) \pi_{k|s_a}.$$

No obstante, se presentan dos situaciones:

- **Situación 1.** $\pi_{k|s_a}$ no depende de s_a , teniendo así un diseño en dos etapas.

$$\pi_k = \pi_{k|s_a} \sum_{s_a \ni k} p_a(s_a) = \pi_{k|s_a} \pi_{ak}.$$

- **Situación 2.** $\pi_{k|s_a}$ depende de s_a , luego

$$\pi_k = \sum_{s_a \ni k} p_a(s_a) \pi_{k|s_a} \neq \pi_{k|s_a} \pi_{ak}.$$

Luego, se propone un nuevo estimador conocido como π^* -estimador (Särndal, 1992, p.347), definido como

$$\hat{t}_{\pi^*} = \sum_{k \in s} \frac{Y_k}{\pi_{ak} \pi_{k|s_a}}$$

con las siguientes propiedades:

- $E(\hat{t}_{\pi^*}) = t$
- $Var(\hat{t}_{\pi^*}) = \sum \sum_U \Delta_{akl} \check{Y}_{ak} \check{Y}_{al} + E_{s_a} \left(\sum \sum_{s_a} \Delta_{kl|s_a} \frac{Y_k}{\pi_k^*} \frac{Y_l}{\pi_l^*} \right)$
- $\widehat{Var}(\hat{t}_{\pi^*}) = \sum \sum_s \frac{\Delta_{akl}}{\pi_{kl}^*} \check{Y}_{ak} \check{Y}_{al} + \sum_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \frac{Y_k}{\pi_k^*} \frac{Y_l}{\pi_l^*}$

donde $\check{Y}_k = \frac{Y_k}{\pi_{ak}}$, $\pi_k^* = \pi_{ak} \pi_{k|s_a}$, $\Delta_{akl} = \pi_{akl} - \pi_{ak} \pi_{al}$ y $\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a} \pi_{l|s_a}$.

2.1.3. Muestreo en dos ocasiones

Dependiendo de los objetivos del estudio, usualmente la misma población es encuestada en diferentes momentos, recogiendo en todos éstos información de alguna variable de interés. Para la primera ocasión, la variable de estudio se denota como Y_1 y en la segunda ocasión como Y_2 (Särndal, 1992, p.369).

- **Primera ocasión.** Una muestra s_a es seleccionada mediante el diseño $p_a(\cdot)$ y en todos los elementos se observa Y_1 , ver Figura 1.1(a). Las probabilidades de inclusión se notan π_{ak} y π_{akl} , luego $\Delta_{akl} = \pi_{akl} - \pi_{ak} \pi_{al}$.

- **Segunda ocasión.** Dos muestras independientes son seleccionadas. Una *muestra emparejada*, denotada por s_{m_1} , seleccionada de s_a de acuerdo a un diseño $p_{m_1}(\cdot | s_a)$ y una *muestra no emparejada* s_u , seleccionada de $s_a^c = U - s_a$ mediante un diseño $p_u(\cdot | s_a^c)$ e independiente de s_{m_1} . Ver Figura 1.1(b).

Las cantidades

$$\pi_{k|s_a}, \pi_{kl|s_a}, \Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a},$$

son asociadas a $p_{m_1}(\cdot | s_a)$ y

$$\pi_{k|s_a^c}, \pi_{kl|s_a^c}, \Delta_{kl|s_a^c} = \pi_{kl|s_a^c} - \pi_{k|s_a^c}\pi_{l|s_a^c},$$

son las cantidades asociadas a $p_u(\cdot | s_a^c)$.

Además se tiene que

$$\pi_{ak}^c = 1 - \pi_{ak},$$

$$\pi_{akl}^c = 1 - \pi_{ak} - \pi_{al} + \pi_{akl}$$

y

$$\Delta_{akl}^c = \Delta_{akl}.$$

La variable Y_2 es observada en s_{m_1} y s_u . La muestra total de la segunda ocasión estaría dada por $s = s_{m_1} \cup s_u$.

Cuando la misma variable es observada en diferentes tiempos, estimaciones acerca del total presente, total previo, cambio absoluto o suma de totales pueden llevarse a cabo. Para el presente trabajo el interés está en la estimación del total presente y funciones de éste como cambio absoluto entre tiempos. El desarrollo presentado a continuación será para el estimador del total.

Se asume que Y_{k2} se puede aproximar mediante $Y_{k2}^0 = KY_{k1}$, donde K es una constante conocida proveniente de algún estudio previo o del conocimiento de un experto. Usando s_a, s_{m_1} y la diferencia $D_k = Y_{k2} - Y_{k2}^0$, se construye un estimador de diferencia insesgado para el total presente,

$$\hat{t}_1 = \hat{t}_{Y_2^0 s_a} + \hat{t}_{D s_{m_1}},$$

donde,

$$\hat{t}_{Y_2^0 s_a} = \sum_{s_a} \frac{Y_{k2}^0}{\pi_{ak}}, \quad \hat{t}_{D s_{m_1}} = \sum_{s_{m_1}} \frac{D_k}{\pi_{ak}\pi_{k|s_a}}.$$

Un segundo estimador insesgado para el total presente se define como,

$$\hat{t}_2 = \hat{t}_{Y_2 s_u} = \sum_{s_u} \frac{Y_{k2}}{\pi_{ak}^c \pi_{k|s_a^c}}.$$

Así, el estimador insesgado para el total, conocido como *estimador compuesto* y que combina tanto la muestra emparejada como la no emparejada, se obtiene a partir de una combinación lineal entre \hat{t}_1 y \hat{t}_2 ,

$$\hat{t}_Y = w_1 \hat{t}_1 + w_2 \hat{t}_2,$$

donde w_1 y w_2 son pesos constantes no negativos tal que $w_1 + w_2 = 1$.

Sea $V_1 = V(\hat{t}_1)$, $V_2 = V(\hat{t}_2)$ y $C = C(\hat{t}_1, \hat{t}_2)$. La varianza del estimador está dada por, ver (Särndal, 1992)[p.371],

$$\hat{V}(\hat{t}_Y) = w_1^2 V_1 + w_2^2 V_2 + 2w_1 w_2 C$$

donde

$$V_1 = \sum \sum_U \Delta_{akl} \frac{Y_{k2} Y_{l2}}{\pi_{ak} \pi_{al}} + E \left(\sum \sum_{s_a} \Delta_{kl|s_a} \frac{D_k}{\pi_{ak} \pi_{k|s_a}} \frac{D_l}{\pi_{al} \pi_{l|s_a}} \right),$$

$$V_2 = \sum \sum_U \Delta_{akl}^c \frac{Y_{k2} Y_{l2}}{\pi_{ak}^c \pi_{al}^c} + E \left(\sum \sum_{s_a^c} \Delta_{kl|s_a^c} \frac{Y_{k2}}{\pi_{ak}^c \pi_{k|s_a^c}} \frac{Y_{l2}}{\pi_{al}^c \pi_{l|s_a^c}} \right)$$

y

$$C = - \sum \sum_U \Delta_{akl} \frac{Y_{k2} Y_{l2}}{\pi_{ak} \pi_{al}^c}.$$

Buscando minimizar la varianza del estimador y con la restricción de $w_1 + w_2 = 1$, se tiene que

$$w_1 = \frac{V_2 - C}{V_1 + V_2 - 2C}.$$

2.2. No Respuesta

Cuando se tienen problemas de no respuesta, entendiendo la no respuesta como “la información deseada que no se obtiene del conjunto de elementos s designado para observar” (Särndal, 1992, p.556), surgen problemas como estimaciones menos eficientes por el tamaño de muestra reducido, restricción en el uso de técnicas estándar y un posible sesgo debido a que el comportamiento de los respondientes puede ser sistemáticamente diferente al de los no respondientes (Rubin, 1987, p.1).

Por ello, entre las técnicas planteadas para abordar este problema y que se presentan a continuación están: el submuestreo de los no respondientes, la cual es basada en la teoría presentada en la sección 2.1.2 y es una de las técnicas más usadas. Además, imputación de los datos que en este trabajo se enfocará a la imputación múltiple.

2.2.1. Submuestreo de no respondientes

La idea general de esta técnica se basa en tomar una submuestra de los no respondientes y realizar todo el esfuerzo posible para obtener la información requerida para todos los sujetos de esta submuestra. El procedimiento es similar al desarrollado en muestreo en dos fases y se explica a continuación:

- **Primera fase.** Una muestra s_a de tamaño n_a se selecciona de acuerdo a un diseño $p_a(\cdot)$, con probabilidades de inclusión π_{ak} y π_{akl} . No respuesta para la variable de interés Y se evidencia en s_a , creándose un grupo de respondientes s_{a1} de tamaño n_{a1} y un segundo grupo de no respondientes s_{a2} de tamaño n_{a2} .
- **Segunda fase.** Una muestra s_2 se selecciona de s_{a2} de acuerdo a un diseño $p(\cdot | s_{a2})$, con probabilidades de inclusión $\pi_{k|s_{a2}}$ y $\pi_{kl|s_{a2}}$. Para s_2 se debe realizar el mayor esfuerzo para obtener la información de interés.

Sea $s = s_{a1} \cup s_{a2}$ el conjunto para el cual se observó Y . El estimador del total poblacional se define como

$$\hat{t} = \sum_s \frac{Y_k}{\pi_k^*},$$

donde

$$\pi_k^* = \begin{cases} \pi_{ak} & \text{si } k \in s_{a1}; \\ \pi_{ak}\pi_{k|s_{a2}} & \text{si } k \in s_{a2}. \end{cases}$$

Propiedades del estimador pueden ser consultados en Särndal (1992, p.566).

2.2.2. Imputación múltiple

Técnica caracterizada por imputar $m > 1$ veces los valores perdidos dentro de las variables de interés, creando un total de m bases completas con las cuales se pueden trabajar las técnicas estándar. Su proceso empieza con la imputación de los valores, seguido de la estimación de los parámetros de interés en cada una de las bases generadas, finalizando con la agrupación de las m estimaciones en una sola y estimando su varianza (van Buuren, 2012)[p.16].

Entre los problemas presentes a la hora de imputar datos multivariados están (van Buuren & Groothuis-Oudshoorn, 2011):

- Los predictores utilizados en el proceso de imputación están incompletos.
- Posible correlación entre las variables.
- Tamaño de muestra pequeño y gran número de variables, llevando a problemas de colinealidad.
- Relación no lineal entre la variable a imputar y los predictores.

Debido a la complejidad presente en los datos reales y los diferentes tipos de variables a imputar, lo ideal es especificar el modelo a utilizar para la imputación de cada variable por separado. Aunque hay diversas formas de implementar la imputación condicionado a la especificación de modelos, el de interés para el trabajo se conoce como algoritmo MICE (Multivariate Imputation by Chained Equations).

Siguiendo la notación presentada por van Buuren & Groothuis-Oudshoorn (2011), sea Y_j con $j = 1, \dots, p$ una de las variables incompletas, donde $Y = (Y_1, \dots, Y_p)^T$. El conjunto de variables con datos faltantes se denota por $Y^{mis} = (Y_1^{mis}, \dots, Y_p^{mis})^T$ y aquellas con todas las observaciones, $Y^{obs} = (Y_1^{obs}, \dots, Y_p^{obs})^T$. El número total de imputaciones es $m > 1$. El h -ésimo conjunto de datos imputados se denota $Y^{(h)}$ con $h = 1, \dots, m$ y la colección de las $p - 1$ variables excepto Y_j está dado por $Y_{-j} = (Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_p)^T$. A la cantidad de interés se le denota por Q .

Asumiendo que la distribución multivariada para Y está completamente definida por el vector de parámetros desconocido θ , el problema se traslada a obtener la distribución multivariada de θ . Para ello, el algoritmo MICE obtiene la distribución posteriori de θ muestreando de forma iterativa de las distribuciones condicionales

$$P(Y_1 | Y_{-1}, \theta_1)$$

$$\begin{aligned} & \vdots \\ & P(Y_p | Y_{-p}, \theta_p) \end{aligned}$$

Los parámetros $\theta_1, \dots, \theta_p$ son específicos a cada distribución condicional.

Para la t -ésima iteración de las ecuaciones encadenadas (concatenación de procedimientos univariados), un muestreador de Gibbs se utiliza a partir de las distribuciones marginales observadas

$$\begin{aligned} \theta_1^{*(t)} &\sim P(\theta_1 | Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}) \\ Y_1^{*(t)} &\sim P(Y_1 | Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}, \theta_1^{*t}) \\ & \vdots \\ \theta_p^{*(t)} &\sim P(\theta_p | Y_p^{obs}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}) \\ Y_p^{*(t)} &\sim P(Y_p | Y_p^{obs}, Y_1^{(t)}, \dots, Y_p^{(t)}, \theta_p^{*(t)}) \end{aligned}$$

donde $Y_j^{(t)} = (Y_j^{obs}, Y_j^{*(t)})$ es la j -ésima variable imputada en la iteración t .

La convergencia del método puede darse relativamente rápido, con alrededor de 10 a 20 imputaciones es suficiente.

Dado que la escogencia del modelo de imputación es el paso más importante dentro de la técnica, a continuación se citan algunos pasos a tener en cuenta:

- Identificar el comportamiento de los datos faltantes. Pérdida al azar (MAR, por sus siglas en inglés) cuando la probabilidad de ser un dato perdido es la misma dentro de los grupos definidos por los datos observados, o pérdida no al azar (MNAR, por sus siglas en inglés) cuando la probabilidad de ser un dato faltante varía por razones desconocidas.
- Seleccionar el modelo de imputación para cada variable a imputar. Las opciones disponibles dentro del algoritmo se encuentran en la Tabla 2.1.
- Seleccionar el conjunto de variables predictoras incluidas en el modelo de imputación.
- Decidir si es necesaria alguna transformación de las variables.
- Definir el número de iteraciones.
- Escoger m , el número de imputaciones a realizar.

2.3. Ecuaciones de Estimación Generalizadas

De acuerdo a Paula (2013, p.374) cuando los datos están correlacionados y la distribución marginal de los datos no es normal, existen dos formas de proseguir. La primera de ellas, ignorando la estructura de correlación para luego aplicar un modelo lineal generalizado y obtener, de esta manera, estimaciones consistentes y asintóticamente normales, perdiendo muchas veces eficiencia.

TABLA 2.1. Técnicas de imputación implementadas en el algoritmo MICE, de acuerdo al tipo de variable.

Método	Descripción	Tipo de variable
pmm	Predictive mean matching	Numérica
norm	Regresión lineal bayesiana	Numérica
norm.nob	Regresión lineal no bayesiana	Numérica
mean	Imputación media incondicional	Numérica
2L.norm	Modelo lineal de dos niveles	Numérica
logreg	Regresión logística	Factor, 2 niveles
polyreg	Modelo logit multinomial	Factor, > 2 niveles
polr	Modelo logit ordenado	Factor ordenado, > 2 niveles
lda	Análisis discriminante lineal	Factor
sample	Muestra aleatoria de los datos observados	Cualquiera

Una segunda manera es incluyendo la estructura de correlación a la función score. Así, siguiendo la notación de Paula (2013, p.373), sea $Y_k = (Y_{k1}, \dots, Y_{kr_k})^T$ un vector de respuesta multivariada para el k -ésimo individuo, con $k = 1, \dots, n$ y $t = 1, \dots, r_k$ donde r_k es el tiempo máximo de observación del k -ésimo individuo. Asumiendo que sólo se conoce la distribución marginal de Y_{kt} , dada por

$$f(y; \theta_{kt}, \phi) = \exp[\phi\{y\theta_{kt} - b(\theta_{kt})\} + c(y, \phi)],$$

donde $b(\cdot)$ y $c(\cdot)$ son funciones reales, se plantea el siguiente modelo:

$$\begin{cases} E(Y_k) = \mu_k, \\ Var(Y_k) = \phi^{-1}V_k^{1/2}R_k(\rho)V_k^{1/2} = \Omega_k, \\ g(\mu_k) = X_k^T\beta, \end{cases}$$

con $V_k = \text{diag}\{V_{k1}, \dots, V_{kr_k}\}$, $V_{kt} = d\mu_{kt}/d\theta_{kt} = db'(\theta_{kt})/d\theta_{kt}$ ¹ la función de varianza, $\phi^{-1} > 0$ el parámetro de dispersión, β el vector de parámetros, X_k matriz de variables explicativas observadas para el k -ésimo individuo, $g(\cdot)$ la función de enlace y $R_k(\cdot)$ la matriz con la estructura de correlación existente entre las mediciones.

Para estimar β se resuelve el sistema de ecuaciones

$$S(\beta) = 0,$$

conocido como ecuaciones de estimación generalizadas (EEG), donde

$$S(\beta) = \sum_{k=1}^n D_k^T \Omega_k^{-1} (Y_k - \mu_k) = \sum_{k=1}^n D_k^T \left[\phi^{-1} V_k^{1/2} R_k(\rho) V_k^{1/2} \right]^{-1} (Y_k - \mu_k),$$

con $D_k = W_k^{1/2} V_k^{1/2} X_k$, $W_k = \text{diag}\{w_{k1}, \dots, w_{kr_k}\}$ una matriz de pesos con $w_{kt} = (d\mu_{kt}/d\eta_{kt})^2 / V_{kt}$ y $\eta_{kt} = X_{kt}^T \beta$.

Al no ser $S(\beta)$ un sistema lineal, no tiene solución cerrada. Por tanto la estimación se realiza por métodos iterativos como Newton Raphson o Fisher Scoring (Paula, 2013,

¹Por condiciones de regularidad de la familia exponencial lineal, se tiene que $\mu_{kt} = b'(\theta_{kt})$

p.376). Asumiendo que $\hat{\phi}$ y $\hat{\rho}$ son estimadores consistentes para ϕ y ρ , se tiene que

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N_p(0, \Sigma),$$

$$\text{con } \Sigma = \lim_{n \rightarrow \infty} \left[n \left(\sum_{k=1}^n D_k^T \Omega_k^{-1} D_k \right)^{-1} \left(\sum_{k=1}^n D_k^T \Omega_k^{-1} \text{Var}(Y_k) \Omega_k^{-1} D_k \right) \left(\sum_{k=1}^n D_k^T \Omega_k^{-1} D_k \right)^{-1} \right].$$

Un estimador robusto para $\text{Var}(\hat{\beta})$ está dado por (Paula, 2013, p.376)

$$\hat{V}_\beta = H_1^{-1}(\hat{\beta}) H_2(\hat{\beta}) H_1^{-1}(\hat{\beta}),$$

$$\text{con } H_1(\hat{\beta}) = \sum_{k=1}^n \hat{D}_k^T \hat{\Omega}_k^{-1} \hat{D}_k \text{ y } H_2(\hat{\beta}) = \sum_{k=1}^n \hat{D}_k^T \hat{\Omega}_k^{-1} (Y_k - \hat{\mu}_k) (Y_k - \hat{\mu}_k)^T \hat{\Omega}_k^{-1} \hat{D}_k.$$

Aunque diferentes estructuras de correlación pueden utilizarse, la no estructurada genera interés debido a que son $r_k(r_k - 1)/2$ parámetros a estimar, luego

$$\hat{R}_{jj'} = \frac{1}{n} \sum_{k=1}^n \frac{(Y_{kj} - \hat{\mu}_{kj})}{\sqrt{\hat{V}_{kj}}} \frac{(Y_{kj'} - \hat{\mu}_{kj'})}{\sqrt{\hat{V}_{kj'}}}.$$

Entre las técnicas diagnóstico para las EEG se encuentran los residuos de Pearson y una versión de la distancia de Cook para evaluar datos influyentes o medidas de apalancamiento. Mayor información ver Paula (2013, p.378).

Muestreo en tres ocasiones

Siguiendo la idea presentada en la sección 2.1.3, se considera un muestreo de tres ocasiones para una población finita $U = \{1, \dots, k, \dots, N\}$, compuesta de los mismos elementos en los tres momentos. La variable de estudio que se mide en los elementos seleccionados se define como, Y_1 para la primera ocasión, Y_2 para la segunda ocasión y Y_3 para la tercera ocasión.

- **Primera ocasión.** Se selecciona una muestra s_a mediante un diseño $p_a(\cdot)$ y en los elementos se observa Y_1 . Las probabilidades de inclusión de primer y segundo orden se notan como

$$\pi_{ak}$$

y

$$\pi_{akl},$$

respectivamente. Ver Figura 3.1.

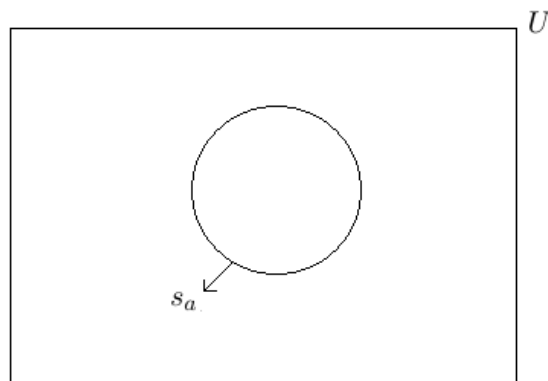


FIGURA 3.1. Muestra s_a obtenida en la primera ocasión.

- **Segunda ocasión.** Se seleccionan dos muestras de manera independiente, ver Figura 3.2. La primera, una *muestra emparejada*, denotada por s_{m_1} y seleccionada de s_a

mediante $p_{m_1}(\cdot | s_a)$. Sus probabilidades de inclusión de primer y segundo orden son

$$\pi_{k|s_a}$$

y

$$\pi_{kl|s_a}.$$

La segunda, una *muestra no emparejada*, denotada como s_u y seleccionada de $s_a^c = U - s_a$ a partir de un diseño muestral $p_u(\cdot | s_a^c)$. Las probabilidades de inclusión de primer orden se notan por

$$\pi_{k|s_a^c}$$

y de segundo orden

$$\pi_{kl|s_a^c}.$$

En ambas muestras se observa Y_2 , pero solo los elementos que están en s_{m_1} tienen también información de Y_1 .

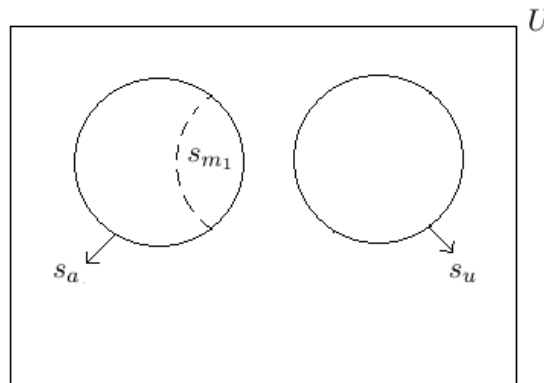


FIGURA 3.2. Muestras de la segunda ocasión: s_{m_1} seleccionada de s_a y muestra s_u seleccionada de $s_a^c = U - s_a$.

- **Tercera ocasión.**¹ Tres muestras independientes son seleccionadas de U , dos de ellas *emparejadas* con muestras anteriores y una *no emparejada*, ver Figura 3.3.

Sea s_{m_2} una muestra de s_a , seleccionada a partir de $p_{m_2}(\cdot | s_a)$. Sus probabilidades de inclusión de primer y segundo orden son respectivamente,

$$\pi_{k|s_{m_2}}$$

y

$$\pi_{kl|s_{m_2}}.$$

Luego,

$$\Delta_{kl|s_{m_2}} = \pi_{kl|s_{m_2}} - \pi_{k|s_{m_2}}\pi_{l|s_{m_2}}.$$

¹**Nota:** Inicialmente se consideraron cuatro muestras seleccionadas de manera independiente y como estimador del total una combinación lineal de cuatro estimadores. No obstante, al generar mayor sesgo y varianza que el estimador propuesto, éste se descartó.

Los elementos de esta muestra cuentan con información de Y_1 y Y_3 , aunque algunos por haber sido seleccionados en s_{m_1} también tienen información de Y_2 .

Una segunda muestra es s_{m_3} , seleccionada de s_u mediante un diseño muestral $p_{m_3}(\cdot | s_u)$, y con probabilidades de inclusión de primer y segundo orden dadas respectivamente por

$$\pi_{k|s_u}$$

y

$$\pi_{kl|s_u}.$$

Así,

$$\Delta_{kl|s_u} = \pi_{kl|s_u} - \pi_{k|s_u}\pi_{l|s_u}.$$

Los elementos de esta muestra además de observarles Y_3 , tienen información de Y_2 .

Siendo $s = s_a \cup s_u$, con s_a disyunto de s_u , la muestra no emparejada se denota por s_n y es obtenida de $s^c = (s_a \cup s_u)^c$ mediante un diseño muestral $p_n(\cdot | s^c)$. Sus probabilidades de inclusión de primer orden son

$$\pi_{k|s^c}$$

y de segundo orden son

$$\pi_{kl|s^c}.$$

Por tanto,

$$\Delta_{kl|s^c} = \pi_{kl|s^c} - \pi_{k|s^c}\pi_{l|s^c}.$$

Cabe resaltar que el conjunto s tiene asociado unas probabilidades de inclusión denotadas por π_{sk} y π_{skl} . Para los elementos de s_n solo se tiene información de Y_3 .

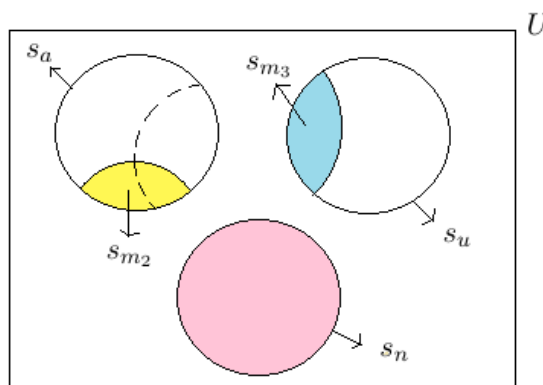


FIGURA 3.3. Comportamiento de las muestras en la tercera ocasión. Muestra s_{m_2} seleccionada de s_a , muestra s_{m_3} seleccionada de s_u y muestra s_n seleccionada de $s^c = (s_a \cup s_u)^c$.

Definido lo anterior y con interés de actualizar el estimador del total con información recolectada en tiempos anteriores, se definen tres estimadores insesgados para el total a partir de las tres muestras obtenidas en la tercera ocasión y la combinación lineal de éstos permitirá la construcción del estimador total.

A diferencia de lo propuesto en la sección 2.1.3, donde se aproxima Y_{k2} a través de KY_{k1} , con K en algunos casos el coeficiente de correlación entre variables, se propone ajustar y utilizar ecuaciones de estimación generalizadas para cada muestra obtenida. A la aproximación hecha para Y_{k3} se le denota Y_{k3}^0 y permite la construcción de $D_k = Y_{k3} - Y_{k3}^0$, teniendo en cuenta la información previa de cada individuo y así, la muestra a la que pertenece.

Haciendo uso de la muestra s_a , la muestra de la tercera ocasión s_{m_2} y las diferencias D_k , se forma un estimador insesgado del total quedando como,

$$\hat{t}_1 = \sum_{s_a} \frac{Y_{k3}^0}{\pi_{ak}} + \sum_{s_{m_2}} \frac{D_k}{\pi_{ak}\pi_{k|s_{m_2}}}. \quad (3.1)$$

Un segundo estimador insesgado del total se forma con la muestra s_u , la muestra de la tercera ocasión s_{m_3} y las diferencias D_k obtenidas para este caso. Queda expresado como,

$$\hat{t}_2 = \sum_{s_u} \frac{Y_{k3}^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_{m_3}} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}}, \quad (3.2)$$

donde $(\pi_{ak})^c = 1 - \pi_{ak}$.

El último estimador insesgado del total, se construye con la muestra s_n y se escribe como,

$$\hat{t}_3 = \sum_{s_n} \frac{Y_{k3}}{(\pi_{sk})^c \pi_{k|s^c}}, \quad (3.3)$$

donde $(\pi_{sk})^c = 1 - \pi_{sk}$.

Como se muestra en la sección A.1 del apéndice A, tanto \hat{t}_1 , \hat{t}_2 y \hat{t}_3 son estimadores insesgados para el total. Luego, mediante una combinación lineal de los mismos se tiene un nuevo estimador insesgado del total

$$\hat{t}_Y = w_1 \hat{t}_1 + w_2 \hat{t}_2 + w_3 \hat{t}_3,$$

donde w_1 , w_2 y w_3 son constantes no negativas conocidas, tal que $w_1 + w_2 + w_3 = 1$.

Sean $V_1 = V(\hat{t}_1)$, $V_2 = V(\hat{t}_2)$, $V_3 = V(\hat{t}_3)$, $C_{12} = C(\hat{t}_1, \hat{t}_2)$, $C_{13} = C(\hat{t}_1, \hat{t}_3)$ y $C_{23} = C(\hat{t}_2, \hat{t}_3)$. La varianza del estimador \hat{t}_Y está dada por

$$V(\hat{t}_Y) = w_1^2 V_1 + w_2^2 V_2 + w_3^2 V_3 + 2w_1 w_2 C_{12} + 2w_1 w_3 C_{13} + 2w_2 w_3 C_{23},$$

donde

$$V_1 = \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{\pi_{al}} + E_{s_a} \left(\sum \sum_{s_a} \Delta_{kl|s_{m_2}} \frac{D_k}{\pi_{ak}\pi_{k|s_{m_2}}} \frac{D_l}{\pi_{al}\pi_{l|s_{m_2}}} \right), \quad (3.4)$$

$$V_2 = \sum \sum_U \Delta_{kl|s_a^c} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l}{(\pi_{al})^c \pi_{l|s_a^c}} + E_{s_u} \left(\sum \sum_{s_u} \Delta_{kl|s_u} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \frac{D_l}{(\pi_{al})^c \pi_{l|s_a^c} \pi_{l|s_u}} \right), \quad (3.5)$$

$$V_3 = \sum \sum_U \Delta_{skl}^c \frac{Y_k}{(\pi_{sk})^c} \frac{Y_l}{(\pi_{sl})^c} + E_{s^c} \left(\sum \sum_{s^c} \Delta_{kl|s^c} \frac{Y_k}{(\pi_{sk})^c \pi_{k|s^c}} \frac{Y_l}{(\pi_{sl})^c \pi_{l|s^c}} \right), \quad (3.6)$$

$$C_{12} = - \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c}, \quad (3.7)$$

$$C_{13} = - \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c}, \quad (3.8)$$

$$C_{23} = - \sum \sum_U \Delta_{kl|s_a^c} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l}{(\pi_{al})^c (\pi_{l|s_a^c})^c}. \quad (3.9)$$

La demostración se presenta en las secciones A.2 y A.3 del apéndice A.

La óptima escogencia de w_1 , w_2 y w_3 se hace a partir de la minimización de la varianza. Haciendo uso de la metodología de multiplicadores de Lagrange (Mardsen & Tromba, 1991, p. 265), el sistema de ecuaciones queda expresado como,

$$\begin{aligned} 2w_1V_1 + 2w_2C_{12} + 2w_3C_{13} + \lambda &= 0 \\ 2w_1C_{12} + 2w_2V_2 + 2w_3C_{23} + \lambda &= 0 \\ 2w_1C_{13} + 2w_2C_{23} + 2w_3V_3 + \lambda &= 0 \\ w_1 + w_2 + w_3 &= 1. \end{aligned}$$

Luego,

$$\lambda = \frac{h_1}{h_1h_4 - h_2h_3}, \quad (3.10)$$

$$w_3 = \frac{-\lambda h_2}{2h_1}, \quad (3.11)$$

$$w_2 = \frac{-2w_3 \left(\frac{V_1V_3 - C_{12}C_{13}}{V_1} \right) - \lambda \left(\frac{V_1 - C_{12}}{V_1} \right)}{\frac{2(V_1V_2 - C_{12}^2)}{V_1}}, \quad (3.12)$$

$$w_1 = \frac{-\lambda - 2w_2C_{12} - 2w_3C_{13}}{2V_1}, \quad (3.13)$$

donde,

$$\begin{aligned} h_1 &= \frac{V_1V_3 - C_{13}^2}{V_1} - \frac{(V_1V_3 - C_{12}C_{13})(V_1C_{23} - C_{12}C_{13})}{V_1(V_1V_2 - C_{12}^2)}, \\ h_2 &= \frac{V_1 - C_{13}}{V_1} - \frac{(V_1 - C_{12})(V_1C_{23} - C_{12}C_{13})}{V_1(V_1V_2 - C_{12}^2)}, \\ h_3 &= \frac{2V_1 - C_{13}}{2V_1} - \frac{2(V_1V_3 - C_{12}C_{13})(2V_1 - C_{12})}{V_1(2V_1V_2 - C_{12}^2)}, \\ h_4 &= \frac{-1}{2V_1} - \frac{(V_1 - C_{12})(2V_1 - C_{12})}{2V_1V_2 - C_{12}^2}. \end{aligned}$$

Debido a que las expresiones (3.10), (3.11), (3.12) y (3.13) dependen de las varianzas y covarianzas de \hat{t}_1 , \hat{t}_2 y \hat{t}_3 , la construcción de $\hat{w}'s$ es posible obteniendo las expresiones para \hat{V}_1 , \hat{V}_2 , \hat{V}_3 , \hat{C}_{12} , \hat{C}_{13} y \hat{C}_{23} , lo que lleva a tener un estimador del total con estimaciones de

w_1 , w_2 y w_3 , expresado como

$$\widehat{t}_Y = \widehat{w}_1 \widehat{t}_1 + \widehat{w}_2 \widehat{t}_2 + \widehat{w}_3 \widehat{t}_3. \quad (3.14)$$

3.1. Caso particular: Muestreo Aleatorio Simple

Asumiendo para cada ocasión que el diseño muestral empleado fue un muestreo aleatorio simple (MAS), las expresiones anteriores se simplifican de la siguiente manera.

- **Primera ocasión.** Sea s_a una muestra de tamaño n_1 seleccionada de U , su fracción muestral se denota por

$$f_1 = \frac{n_1}{N}.$$

De lo anterior, la muestra s_a^c queda de tamaño $N - n_1$ de U .

- **Segunda ocasión.** Siendo μ la proporción de muestra emparejada, la muestra s_{m_1} seleccionada de s_a es de tamaño μn_2 y su fracción muestral es

$$f_{12} = \frac{\mu n_2}{n_1}.$$

La muestra s_u seleccionada de s_a^c tiene un tamaño igual a $(1 - \mu)n_2$ y la fracción muestral asociada a esta muestra es

$$f_2 = \frac{(1 - \mu)n_2}{N - n_1}.$$

- **Tercera ocasión.** Asumiendo que la proporción de muestra emparejada es igual a μ y buscando garantizar un tamaño para las muestras s_{m_2} , s_{m_3} y s_n , manteniendo unas proporciones establecidas en el anexo B, en la Tabla 3.1 se definen los tamaños de muestra de interés.

TABLA 3.1. Tamaños de muestra para s_{m_2} , s_{m_3} y s_n , obtenidos en la tercera ocasión. Ver anexo B.

Muestra	Tamaño de muestra
s_n	$(1 - \mu)n_3$
s_{m_3}	$a = \frac{\mu n_3 (1 - \mu) n_2}{n_1 + (1 - \mu) n_2}$
s_{m_2}	$b = \frac{\mu a}{1 - \mu} \left(\frac{n_1 - \mu n_2}{\mu n_2} + 1 \right)$

Las fracciones muestrales asociadas a s_{m_2} , s_{m_3} y s_n son respectivamente,

$$f_{13} = \frac{b}{n_1},$$

$$f_{23} = \frac{a}{(1 - \mu)n_2}$$

y

$$f_3 = \frac{(1 - \mu)n_3}{N - n_1 - (1 - \mu)n_2}.$$

Establecido lo anterior, las expresiones (3.1), (3.2) y (3.3) quedan definidas como,

$$\hat{t}_1 = N(\bar{Y}_{s_a}^0 + \bar{D}_{s_{m_2}}), \quad (3.15)$$

$$\hat{t}_2 = N(\bar{Y}_{s_u}^0 + \bar{D}_{s_{m_3}}), \quad (3.16)$$

$$\hat{t}_3 = N\bar{Y}_{s_n}. \quad (3.17)$$

Asimismo, las varianzas y covarianzas de las fórmulas (3.4) a (3.9) quedan expresadas como,

$$V_1 = N^2 \left(\frac{1 - f_1}{n_1} S_{YU}^2 + \frac{1 - f_{13}}{b} S_{DU}^2 \right), \quad (3.18)$$

$$V_2 = N^2 \left(\frac{1 - f_2}{(1 - \mu)n_2} S_{YU}^2 + \frac{1 - f_{23}}{a} S_{DU}^2 \right), \quad (3.19)$$

$$V_3 = N^2 \left(\frac{f_1 + f_2}{(N - n_1)^2 + Nn_2(1 - \mu)} + \frac{1 - f_3}{(1 - \mu)n_3} \right) S_{YU}^2, \quad (3.20)$$

$$C_{12} = -NS_{YU}^2, \quad (3.21)$$

$$C_{13} = -NS_{YU}^2, \quad (3.22)$$

$$C_{23} = -\frac{N^2}{N - n_1} S_{YU}^2. \quad (3.23)$$

Con las expresiones obtenidas en (3.17) a (3.22), obtener los w 's descritos en (3.11), (3.12) y (3.13) es posible, y con ello, el estimador del total para un muestreo de tres ocasiones usando MAS. Sin embargo, para situaciones donde la única información disponible es la obtenida a partir de la muestra seleccionada, es de interés trabajar con las estimaciones de las fórmulas anteriores. Así,

$$\hat{V}_1 = N^2 \left(\frac{1 - f_1}{n_1} S_{Y_{s_a}}^2 + \frac{1 - f_{13}}{b} S_{D_{s_{m_2}}}^2 \right), \quad (3.24)$$

$$\hat{V}_2 = N^2 \left(\frac{1 - f_2}{(1 - \mu)n_2} S_{Y_{s_u}}^2 + \frac{1 - f_{23}}{a} S_{D_{s_{m_3}}}^2 \right), \quad (3.25)$$

$$\hat{V}_3 = N^2 \left(\frac{f_1 + f_2}{(N - n_1)^2 + Nn_2(1 - \mu)} + \frac{1 - f_3}{(1 - \mu)n_3} \right) S_{Y_{s_n}}^2, \quad (3.26)$$

$$\hat{C}_{12} = -NS_{Y_{s_a}}^2, \quad (3.27)$$

$$\hat{C}_{13} = -NS_{Y_{s_a}}^2, \quad (3.28)$$

$$\hat{C}_{23} = -\frac{N^2}{N - n_1} S_{Y_{s_u}}^2. \quad (3.29)$$

Encuesta Longitudinal Colombiana de la Universidad de los Andes

4.1. Antecedentes

Con el fin de conocer las dinámicas y cambios existentes dentro de la población colombiana en temas de interés, el Centro de Estudios Sobre Desarrollo Económico (CEDE) y la Facultad de Economía de la Universidad de los Andes crearon la ELCA, con la iniciativa de realizar un seguimiento durante 12 años a los hogares e individuos pertenecientes al panel de ésta (CEDE & Facultad Economía U.Andes, 2011). Así, al presente año cuentan con tres levantamientos de información, llevados a cabo en gran parte de territorio colombiano y que han permitido caracterizar en diferentes aspectos a la población colombiana (CEDE & Facultad Economía U. Andes, 2011, 2014 y 2017).

Siendo la selección de los sujetos de seguimiento, fundamental para la generación de resultados y conclusiones válidas a lo largo del estudio, se definen los siguientes criterios para la selección de muestra de hogares, consultar CEDE & Facultad Economía U. Andes (2010),

- **Universo de estudio:** En la zona urbana son los hogares pertenecientes a los estratos uno al cuatro, agrupados en regiones y con diferentes niveles de urbanización. Para la zona rural son los hogares de pequeños productores (principalmente estrato uno) pertenecientes a subregiones homogéneas internamente a nivel agropecuario.
- **Marco muestral:** Del Censo General de Población y Vivienda de 2005, para la zona urbana se obtiene un inventario cartográfico, archivos de vivienda, hogares y personas agregados por manzanas y centros poblados. Para la zona rural, se utiliza el inventario cartográfico de los Planes de Ordenamiento Territorial o los archivos desarrollados en cada municipio por Secretaria de Planeación.
- **Diseño muestral:** Muestra probabilística, estratificada, multietápica y de conglomerados.
 - De conglomerados: Para el universo urbano están constituidos por cabeceras municipales (Unidades Primarias de Muestreo UPM), manzanas (Unidades Secundarias de Muestreo USM) y segmentos de viviendas contiguas (Unidades

Terciarias de Muestreo UTM). Cada segmento de vivienda contiguo seleccionado se observó completamente.

Para el universo rural se definen los municipios (UPM), veredas (USM) y segmentos (UTM).

- Estratificada: Se presentó una estratificación de acuerdo a variables altamente correlacionadas con el estudio como región, departamento, nivel de urbanización (tamaño de la población) e Índice de Necesidades Básicas Insatisfechas (NBI). Dentro de los estratos definidos se hizo la selección de UPM's. Detalle de los estratos se puede ver en CEDE & Facultad Economía U. Andes (2010).
- Multietápica: Para la zona urbana dentro de cada estrato definido en la primera etapa se hizo selección de municipios, en la segunda etapa se seleccionaron manzanas y en la tercera etapa, segmentos dentro de cada manzana.

Para la zona rural, la primera etapa hace selección de las zonas rurales de cada municipio, en la segunda etapa se seleccionaron veredas, y en la última etapa se seleccionaron segmentos.

- **Tamaño de muestra:** Debido a costos operativos y limitaciones financieras la región Andén Pacífico fue excluida del levantamiento de información, dejando así una muestra para la zona urbana de 5000 hogares igualmente distribuidos entre Bogotá y las regiones: Atlántica, Central, Oriental y Pacífica. Además, una sobremuestra del 20 % debido a posibles rechazos, quedando un total de 6000 encuestas distribuidas de a 1200 en cada región.

Para la zona rural se definió una muestra de 4000 encuestas repartidas en las subregiones de Centro-Oriente, Atlántica Media, Eje Cafetero y Cundiboyacense. Al igual que en la zona urbana se consideró una sobremuestra del 20 % dejando un total de 4800 hogares igualmente distribuidos en las cuatro subregiones.

De acuerdo a las características planteadas anteriormente acerca de la selección de la muestra, para el primer año de recolección de información (2010) la cobertura final obtenida para la zona urbana se encuentra resumida en la Tabla 4.1 y para la zona rural en la Tabla 4.2 (CEDE & Facultad Economía U. Andes, 2011).

TABLA 4.1. Total de encuestas esperadas y efectivas por región dentro de la zona urbana para el año 2010.

Región	Muestra	Encuestas Completas
Atlántica	1200	1126
Oriental	1200	1081
Central	1200	1164
Pacífica	1200	1101
Bogotá	1200	976
Total	6000	5448

Teniendo que la ELCA fue construida para responder a objetivos como: identificar el comportamiento de la pobreza, establecer el impacto de choques como violencia en los hogares, describir el mercado laboral y establecer la inversión en primera infancia a través de varios años, el cuestionario de hogares para el 2010 constaba de 272 preguntas en la zona urbana y 364 en la zona rural. Esto permitió recoger información acerca de la “estructura del hogar y características demográficas de todos sus miembros; información detallada del

TABLA 4.2. Total de encuestas esperadas y efectivas por subregión dentro de la zona rural para el año 2010.

Región	Muestra	Encuestas Completas
Atlántica Media	1200	1180
Cundiboyacense	1200	1203
Eje Cafetero	1200	1209
Centro-Oriente	1200	1128
Total	4800	4720

jefe de hogar, cónyuge y menores de diez años sobre educación, salud, empleo, ingresos, participación social y comunitaria, e información del hogar sobre choques, activos, ahorros, deudas, transferencias, condiciones de la vivienda y del hogar” y para la zona rural, se obtuvo además información sobre “tenencia y uso de tierras, producción agropecuaria y no agropecuaria y uso del tiempo” (CEDE & Facultad Economía U.Andes, 2011).

Para el año 2013, en que se llevó a cabo el segundo levantamiento de información, varios temas fueron evaluados antes de empezar el operativo, entre ellos:

- Se definió como persona de seguimiento al jefe de hogar y cónyuge no mayor de 65 años, y niños menores de diez años en el 2010 identificados como hijos, hijastros, nietos o bisnietos del jefe de hogar o cónyuge. En caso de no encontrarse la persona en el mismo hogar del año 2010, se indagó por su ubicación y se buscó con el fin de encuestar a esa persona y los integrantes del nuevo hogar (CEDE & Facultad Economía U.Andes, 2014).
- Dentro de cada nuevo hogar encuestado por haber recibido o haber sido conformado por una persona de seguimiento, se consideran para el seguimiento a futuro de la ELCA aquellas personas reconocidas como jefe de hogar, cónyuge, hijos, nietos o bisnietos (menores a trece años para el año 2013) de alguno de ellos (CEDE & Facultad Economía U.Andes, 2014).
- Se definió un criterio de búsqueda de acuerdo al tiempo de desplazamiento de los encuestadores y el número de hogares en zonas cercanas (CEDE & Facultad Economía U.Andes, 2014).

Así, según los criterios anteriores la cobertura para el año 2013 en la zona urbana y rural se evidencia en la Tabla 4.3 (CEDE & Facultad Economía U.Andes, 2014).

Adicional a los datos de cobertura, cabe resaltar que para el 2013 los cuestionarios fueron modificados antes de empezar el operativo debido a resultados observados del 2010, como la falta de claridad de algunas preguntas, la necesidad de cambiar los periodos de referencia, la molestia de responder algunos temas por parte de los encuestados, la edad de los niños de seguimiento y eventos como la ola invernal (2010-2011), la implementación de la estrategia “De cero a siempre” (2011), la expedición de la ley de víctimas y restitución de tierras y el inicio de negociaciones entre el Gobierno Nacional y las FARC-EP (Fuertes et al., 2016).

Por ejemplo, se cambió el periodo de referencia a los últimos tres años para indagar acerca del desempleo o la no cobertura de seguridad social, se eliminó la sección de percepción de salud y se incluyó una nueva que diera respuesta a enfermedades crónicas y hábitos

TABLA 4.3. Cobertura para el año 2013 de acuerdo a la zona. (1): Total de hogares de seguimiento de la muestra 2010. (2): Total de hogares de (1) encuestados. (3): Número de encuestas en 2013. (4): Total de encuestas en 2013 teniendo en cuenta la migración urbano/rural y rural/urbano.

Zona	Hogares objeto de seguimiento (1)	Hogares de seguimiento encuestados en 2013 (2)	Encuestas efectivas ELCA 2013 (3)	Total hogares ELCA 2013 (4)
Urbana	5275	4430	4681	4911
Rural	4555	4418	4581	4351
Total	9830	8848	9262	9262

de vida, se ajustaron las preguntas del módulo de gastos del hogar y valor de activos debido a la alta no respuesta, se agregó un nuevo módulo para los niños de seguimiento, se adicionaron preguntas en el módulo de choques para identificar los hogares afectados por la ola invernal y se creó un módulo de información básica, de empleo y educación para las personas del hogar que no iban a ser consideradas de seguimiento (Bernal et al., 2014).

Para el año 2016 en que se llevó a cabo el tercer levantamiento de información, los resultados de cobertura se presentan en la Tabla 4.4 (Fuertes et al., 2017).

TABLA 4.4. Cobertura para el año 2016 de acuerdo a la zona. (1): Total de hogares encuestados en 2010 con interés de seguimiento para el 2016. (2): Total de hogares objeto de seguimiento en 2013 y encuestados. (3): Total de hogares de (1) encuestados. (4): Total de encuestas completas realizadas en 2016 teniendo en cuenta la migración urbano/rural y rural/urbano.

Zona	Hogares objeto de seguimiento (1)	Total hogares ELCA 2013 (2)	Hogares de seguimiento encuestados en 2016 (3)	Encuestas completas ELCA 2016 (4)
Urbano	5275	4681	4394	4359
Rural	4578	4581	4424	4397
Total	9853	9262	8818	8756

Durante el operativo del año 2016, los cuestionarios sufrieron cambios en algunos módulos pero principalmente para los enfocados en los niños de seguimiento, que a la fecha sería una cohorte de niños entre 6 y 16 años, un ajuste a las pruebas socioemocionales acorde a la edad de los niños y la inclusión de temas de postconflicto en dimensiones como el bienestar del hogar, capital social y percepción sobre el proceso y reconciliación (Fuertes et al., 2017).

Finalmente, en búsqueda de mantener representatividad y un bajo porcentaje de no respuesta, entre las acciones implementadas, estuvo el no incluir a hogares y personas de seguimiento que se encontraran a más de 90 minutos de desplazamiento del municipio registrado en la ronda anterior, visitando así 235 municipios, incluso cuando la muestra de línea de base era de 80 municipios, obteniendo una cobertura del 89,5% de hogares (CEDE & Facultad Economía U.Andes, 2017).

Para el presente trabajo el interés está centrado exclusivamente en el módulo concierne a la fuerza laboral. Este módulo identifica a los encuestados que se les hará segui-

miento como ocupados, desocupados o inactivos, y dependiendo al grupo perteneciente varían las preguntas realizadas. Aquellos que al momento de la encuesta, en la semana pasada habían trabajado de forma remunerada por lo menos una hora, como ayudante familiar sin remuneración por lo menos una hora, no habían trabajado pero tenían un empleo de por lo menos una hora o habían trabajado por lo menos una hora y buscaron un empleo se clasificaron como *ocupados*. Los incapacitados de forma permanente se identificaron inmediatamente en el grupo de *inactivos* y finalmente, los que no cumplían con ninguna de estas características pasaron a ser los *desocupados*.

4.2. Análisis descriptivo: Población ocupada

Siendo la población de interés para el trabajo desarrollado, los encuestados identificados como ocupados, en primer lugar se hizo un análisis descriptivo de la información obtenida a partir de los datos recolectados para estos individuos ¹, y con el fin de ver la representatividad de la muestra, se realizó una comparación de la información obtenida en el año 2010 con lo reportado por la Gran Encuesta Integrada de Hogares (GEIH) realizada por el Departamento Administrativo Nacional de Estadística (DANE) en el mismo año. Ésto se realizó solo para el año 2010 dado que, para los siguientes levantamientos, lo que se hizo fue un seguimiento a ciertos hogares del año base.

Para el año 2010, mientras la GEIH reportó que el 39.2 % de la población era ocupada y de estratos uno al cuatro, la ELCA identificó de las 22179 personas encuestadas un 27.5 % como ocupados (6107), tres años después el porcentaje de ocupados aumentó a 43.9 % (9042 personas) y en el año 2016, se mantuvo el porcentaje de personas ocupadas en 44.9 % (8673 personas).

En este grupo la proporción de hombres y mujeres se mantuvo relativamente constante durante los tres años, teniendo 45.2 % de mujeres en el 2010, 45.5 % y 46.1 % en los años 2013 y 2016, respectivamente. Esta proporción se mantuvo en la GEIH para el año 2010, donde reportó un total de 46.5 % de mujeres.

En cuanto a la edad de los ocupados, tanto en el año 2010 como en el 2013 y 2016 se presentó un comportamiento similar, donde la mayoría de sujetos tenían edades entre los 25 y 50 años (ver Figura 4.1). La GEIH reportó para el año 2010 que la mayoría de sujetos tenían edades entre 28 y 48 años.

La Tabla 4.5 muestra el estado civil de los ocupados para los tres años de levantamiento de información. Donde se evidencia, que la mayoría de ellos se encuentran en unión libre a la hora de ser encuestados y la menor proporción se concentra en viudos. No obstante, al comparar la información del año 2010 con lo reportado por la GEIH, se evidencia una gran diferencia en las proporciones, teniendo las siguientes cifras por cada estado civil: casado 27.5 %, separado 14.5 %, viudo 2.5 %, soltero 27.9 % y en unión libre 27.6 %.

TABLA 4.5. Distribución de la población ocupada por estado civil. Años: 2010, 2013 y 2016. ELCA.

Año	Casado	Separado	Viudo	Soltero	Unión libre	Total base
2010	37 %	11 %	2 %	7 %	44 %	6107
2013	25 %	13 %	2 %	28 %	32 %	9042
2016	25 %	13 %	2 %	27 %	32 %	8673

¹Datos disponibles en <https://encuestalongitudinal.uniandes.edu.co/es/>

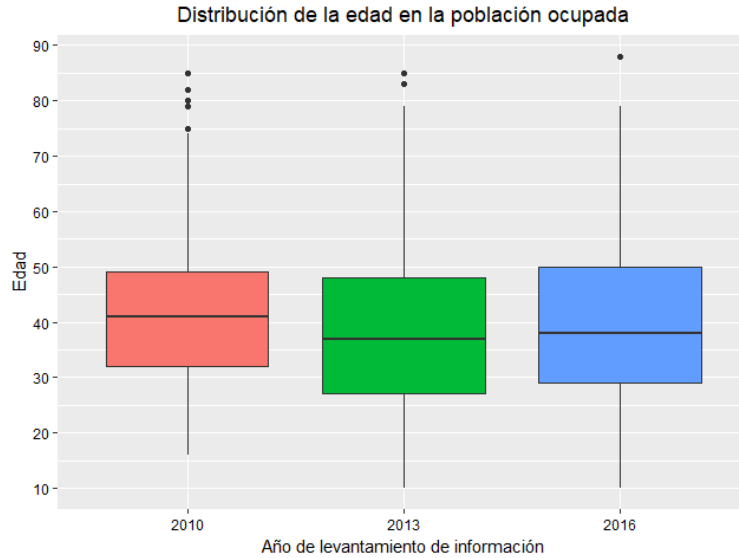


FIGURA 4.1. Distribución de la edad en la población ocupada. Años: 2010, 2013 y 2016. ELCA.

La actividad principal registrada para los ocupados fue “trabajó en forma remunerada durante una hora” y su ocupación, tal como lo muestra la Tabla 4.6, fue asalariado de empresa particular seguida de trabajador por cuenta propia. Cabe resaltar, que tan solo un 1 % de los empleados domésticos eran hombres. Las cifras del DANE para el año 2010 son similares a las obtenidas por la ELCA, teniendo un 41.1 % de asalariados de empresas particulares, 4.6 % asalariados del gobierno, 0.2 % de jornaleros o peones, 3.8 % empleados domésticos, 42.1 % trabajadores por cuenta propia, 4.5 % patrones o empleadores, 3.4 % trabajadores familiares sin remuneración económica y 0.3 % otro.

TABLA 4.6. Tipo de ocupación en la población ocupada. Años: 2010, 2013 y 2016. ELCA.

Ocupación	2010	2013	2016
Asalariado de empresa particular	40 %	42 %	42 %
Asalariado del gobierno	5 %	6 %	5 %
Jornalero o peón	3 %	4 %	3 %
Empleado doméstico	3 %	4 %	4 %
Trabajador por cuenta propia	41 %	39 %	41 %
Patrón o empleador	3 %	1 %	1 %
Trabajador de su propia finca independientemente de la forma de tenencia	0 %	1 %	1 %
Trabajador familiar sin remuneración	2 %	3 %	2 %
Otro	3 %	1 %	0 %

Entre las preguntas realizadas al grupo de ocupados, de las más importantes fue el valor del salario recibido en el empleo principal, cuyo comportamiento se puede ver en la Figura 4.2 y tal como lo evidencia, la distribución tiende a ser asimétrica con la mayoría de los sujetos concentrados en valores muy bajos durante los tres años y con valores desde 2'000.000 pesos ya considerados como datos atípicos. Esto mismo se evidencia en las cifras

de la GEIH con un valor promedio de 657.910 pesos, el 25 % de los individuos con salarios de hasta 246.000 y el 75 % con salarios de máximo 800.000 pesos.

Cabe resaltar que de la información obtenida de la ELCA para el año 2010, solo un 5.4 % de los ocupados pertenecen a estrato socioeconómico cuatro y en las cifras del DANE de los ocupados para el año 2010 pertenecientes a estratos uno al cuatro, solo 8.4 % es de estrato socioeconómico cuatro.

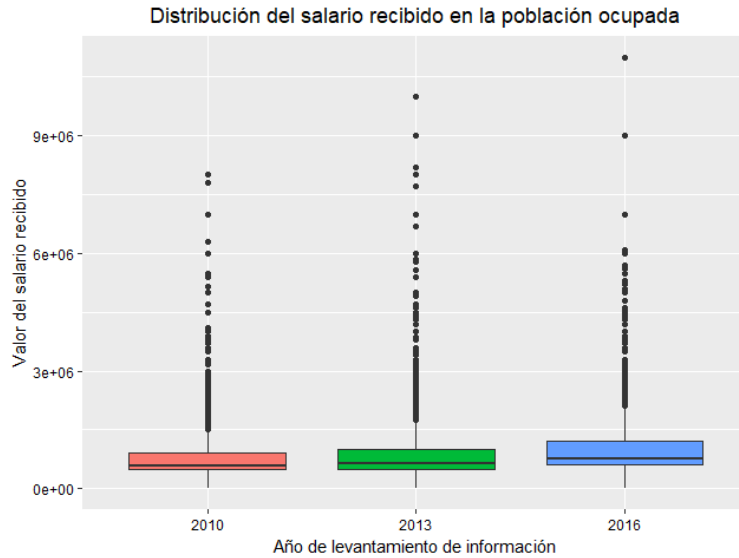


FIGURA 4.2. Distribución del salario recibido en la población ocupada. Años: 2010, 2013 y 2016. ELCA.

4.3. Análisis del salario 2010, 2013 y 2016

Dada la motivación de aplicar la metodología desarrollada en el capítulo 3 a la información brindada por la ELCA y estimar el salario total del año 2016, con el uso de información de años anteriores, en la población identificada como ocupada. Primero se indexaron los salarios del año 2013 y 2016 con base al Índice de Precios al Consumidor (IPC) para llevarlos a una misma unidad de valor, en este caso pesos del años 2010, y ser comparables. Lo anterior, mediante las siguientes fórmulas:

$$SalarioIndexado_{2013} = \frac{IPC_{2010}}{IPC_{2013}} Salario_{2013} = \frac{102.001}{111.815} Salario_{2013}$$

$$SalarioIndexado_{2016} = \frac{IPC_{2010}}{IPC_{2016}} Salario_{2016} = \frac{102.001}{126.149} Salario_{2016}$$

Una vez indexados los salarios y viendo la Figura 4.3 se puede notar como la cantidad de datos atípicos sigue siendo alta, pero a diferencia de la Figura 4.2 los límites son diferentes y los valores máximos entre tiempos no difieren tanto.

Una vez esto, se observó el comportamiento del salario a lo largo de los años en todos los individuos, identificando en algunos de ellos saltos bruscos entre años (Ver Figura 4.4). Por tal motivo y buscando una aproximación del cambio salarial, se realizó el cociente de los salarios 2013 y 2016 con respecto al salario del año 2010, permitiendo establecer

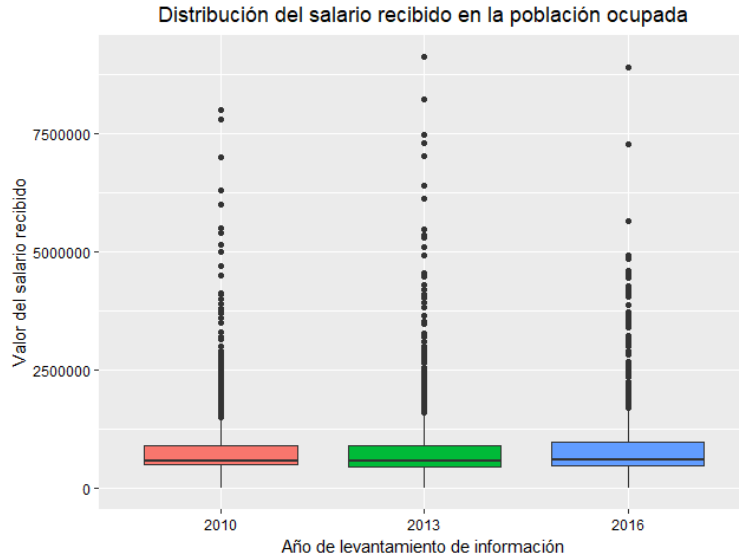


FIGURA 4.3. Distribución del salario recibido en la población ocupada, una vez indexados los salarios de los años 2013 y 2016 con base al IPC. Años: 2010, 2013 y 2016. ELCA.

como individuos atípicos aquellos que recibieron más de 9.5 veces el salario recibido en el año 2010 o los que recibieron 9% o menos del salario de 2010. La Figura 4.5 muestra los 31 individuos con este comportamiento que fueron retirados para los análisis posteriores, debido a la posible mala digitación de su información o por ser respuestas “sospechosas”.

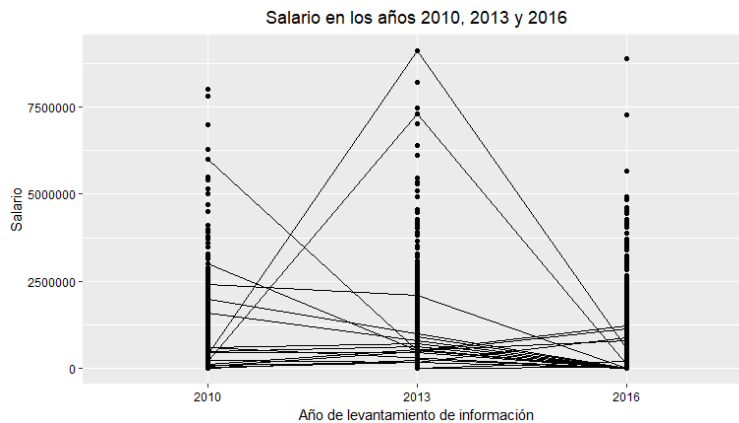


FIGURA 4.4. Salario en los años 2010, 2013 y 2016. ELCA.

Imputación de los salarios faltantes

Considerando la pregunta “El mes pasado, cuánto ganó ... en su empleo principal (incluyendo propinas, comisiones y bonificaciones; excluyendo viáticos, horas extras y pagos)” como una *pregunta sensible*, la presencia de no respuesta es natural. Para cada uno de los años de recolección de información de la ELCA los porcentajes de no respuesta registrados fueron altos, teniendo para el año 2010 una no respuesta del 49.8%, en el 2013 del 67.2% y en el 2016 del 69.2%.

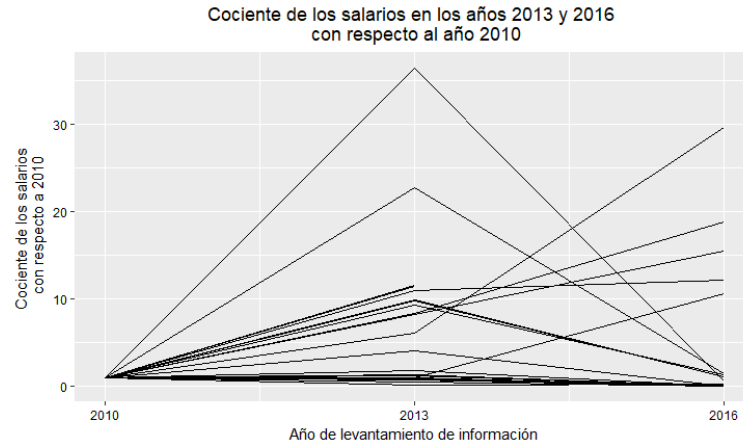
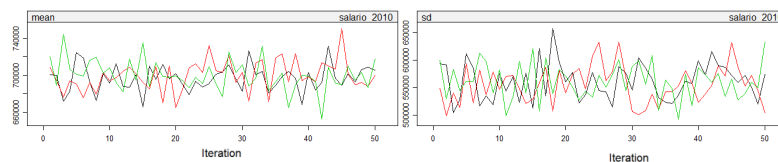


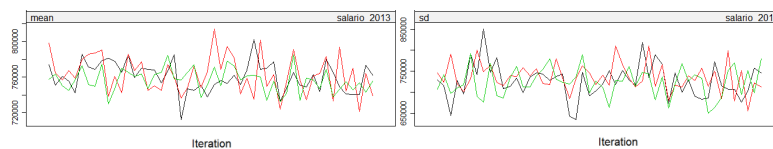
FIGURA 4.5. Cociente de los salarios de los años 2013 y 2016 con respecto al año 2010. ELCA.

Por ello, aunque la imputación múltiple es utilizada para contrarrestar este tipo de problemas, en este caso hacer uso de la metodología mencionada, es generar más del 50% de información de la variable de interés. *Así, el trabajo desarrollado con esta información se presenta como un ejemplo ilustrativo de la teoría presentada en el capítulo 3, mas no como una aplicación de la cual sea válido sacar conclusiones del comportamiento salarial en la población colombiana.*

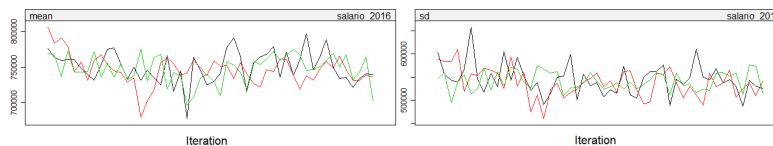
Aclarado lo anterior, se hizo uso de la teoría presentada en la sección 2.2.2 y del paquete `mice` de R (van Buuren, 2017) para generar tres imputaciones por cada valor faltante y a su vez, tres bases completas para trabajar. La Figura 4.6 muestra que a pesar de la variabilidad presente, hay un buen ajuste de acuerdo a la convergencia de las cadenas.



(a) Convergencia de cadenas de Markov ELCA año 2010.



(b) Convergencia de cadenas de Markov ELCA año 2013.



(c) Convergencia de cadenas de Markov ELCA año 2016.

FIGURA 4.6. Convergencia de las cadenas de Markov para la ELCA 2010, 2013 y 2016.

El método seleccionado para imputar el salario fue *predictive mean matching*. Caracterizado porque a cada dato faltante, el método le forma un conjunto de candidatas a partir de los datos observados que sí tienen valor de la variable a imputar (van Buuren, 2012) y por tanto se considera un ejemplo de *hot-deck*.

Para la selección de variables predictoras dentro del modelo, se utilizó la función `quickpred()` del paquete `mice` de R (van Buuren, 2017). En el caso práctico, las variables predictoras utilizadas fueron: edad, género, estrato socioeconómico, nivel educativo y tipo de ocupación.

4.4. Estimación del salario total para el año 2016

Ajuste usando Ecuaciones de Estimación Generalizadas

Tal como se mencionó en el capítulo 3, Y_{k3}^0 es una aproximación de Y_{k3} obtenida a partir del ajuste de una EEG, permitiendo la construcción de $D_k = Y_{k3} - Y_{k3}^0$.

Denotando Y_{kj} el salario del k -ésimo individuo en la j -ésima ocasión, $i = 1, \dots, 15144$ y $j = 1, 2, 3$. Se asume $Y_{kj} \sim G(\mu_{kj}, \phi)$ con la parte sistemática del modelo dada por

$$\mu_{kj} = \beta_0 + \beta_1 \text{Tiempo}_k + \beta_2 \text{Edad}_k + \beta_3 \text{Genero}_k + \beta_4 \text{Estrato}_k^{(2)} + \beta_5 \text{Estrato}_k^{(3)}, \quad (4.1)$$

en que Tiempo_k denota el momento en que se levantó la información, Edad_k es la edad en años, Genero_k ($= 0$ hombre, $= 1$ mujer), $\text{Estrato}_k^{(2)}$ ($= 1$ bajo, $= 0$ no bajo), $\text{Estrato}_k^{(3)}$ ($= 1$ medio, $= 0$ no medio). Asumiendo una estructura de correlación no estructurada para las respuestas de cada sujeto, se tiene que $\text{Corr}(Y_{k1}, Y_{k2}) = \rho_{12}$, $\text{Corr}(Y_{k1}, Y_{k3}) = \rho_{13}$, $\text{Corr}(Y_{k2}, Y_{k3}) = \rho_{23}$ y $\text{Corr}(Y_{kj}, Y_{kj'}) = 1$ para $j = j'$.

Para el conjunto de individuos con observaciones en los años 2010 y 2016 y/o 2013, es decir, los pertenecientes a la muestra s_{m_2} , las estimaciones de los parámetros del modelo con estructura de correlación no estructurada se presentan en la Tabla 4.7. Observando la correlación estimada entre tiempos se tiene para los tres conjuntos de imputación que la mayor correlación se da entre tiempos 1 y 2, y la menor entre tiempos 1 y 3. Sin embargo, no llega a ser mayor de 0.4, lo cual no es muy alto.

Identificando en la Tabla 4.7 los parámetros significativos a un nivel de confianza aproximadamente del 95% con *, se tiene que los parámetros significativos para los tres conjuntos de imputación son el intercepto, el **Tiempo** y el **Genero**. Luego, usando la estimación de β_3 de la imputación 1 se tiene que, el salario promedio de una mujer es 170575.117 pesos menor que el de un hombre.

Para el conjunto de individuos con observaciones en los años 2013 y 2016, es decir, los pertenecientes a la muestra s_{m_3} , las estimaciones de los parámetros del modelo con estructura de correlación no estructurada se presentan en la Tabla 4.8. Nótese que la estimación de correlación no es alta para ninguna de las imputaciones realizadas.

Al igual que lo realizado en la Tabla 4.7, en la Tabla 4.8 se identificaron los parámetros significativos a un nivel de confianza aproximadamente del 95% con *. Por tanto, los parámetros significativos para los tres conjuntos de imputación son el intercepto, el **Tiempo**, el **Genero**, la **Edad** y el **Estrato**. Luego, usando la estimación de β_4 de la imputación 1 se

TABLA 4.7. Estimaciones de los parámetros para el modelo cuasi gamma aplicado a los individuos de la ELCA con observaciones en los años 2010 y 2016 y/o 2013.

Parámetro	Imputación 1		Imputación 2		Imputación 3	
	Estimación	z robusto	Estimación	z robusto	Estimación	z robusto
β_0	942050.395	21.626*	715506.403	9.464*	554516.663	25.279*
β_1	38629.654	10.443*	39356.778	10.825*	35001.607	9.831*
β_2	-1006.403	-2.573*	-352.804	-0.906	-817.886	-2.058*
β_3	-170575.117	-19.517*	-158606.704	-18.308*	-182599.577	-20.580*
β_4	-285687.210	-7.115*	-95598.921	-1.293	97402.861	11.911*
β_5	-111202.116	-2.722*	104067.713	1.402	289388.181	31.246*
ρ_{12}	0.385		0.386		0.402	
ρ_{13}	0.258		0.260		0.268	
ρ_{23}	0.301		0.294		0.308	
ϕ^{-1}	0.252		0.245		0.261	

tiene que el salario promedio de una persona perteneciente al estrato bajo es 689327.374 pesos menor que el de un sujeto perteneciente a estrato alto.

TABLA 4.8. Estimaciones de los parámetros para el modelo cuasi gamma aplicado a los individuos de la ELCA con observaciones en los años 2013 y 2016.

Parámetro	Imputación 1		Imputación 2		Imputación 3	
	Estimación	z robusto	Estimación	z robusto	Estimación	z robusto
β_0	1302221.509	42.583*	1021288.514	10.001*	1176253.416	20.669*
β_1	46496.708	5.311*	61950.785	7.225*	38476.297	4.257*
β_2	-1708.943	-4.034*	-1052.776	-2.635*	-991.827	-2.306*
β_3	-162839.152	-15.743*	-145012.669	-14.745*	-180737.801	-17.812*
β_4	-689327.374	-57.240*	-487256.443	-4.870*	-577066.714	-11.502*
β_5	-509907.692	-41.430*	-282693.187	-2.803*	-363057.751	-7.223*
ρ_{12}	0.173		0.136		0.101	
ϕ^{-1}	0.256		0.239		0.272	

Finalmente, tanto las Figuras 4.7 y 4.8 muestran el diagnóstico realizado a los modelos ajustados, graficando la distancia de Cook y los residuales de Pearson para cada caso de imputación. Aunque es posible ver en todos los casos sujetos con distancias de Cook altas, posiblemente relacionados con los residuales de Pearson que se encuentran por encima de dos, la mayoría de los residuales se concentran entre $[-1, 2]$ sugiriendo un ajuste adecuado del modelo.

Estimador del total

Debido a que la ELCA no cumplía con los supuestos mínimos para la aplicación de la metodología propuesta en el capítulo 3 y por las limitaciones presentadas en la base de datos, para el presente trabajo en el ejemplo ilustrativo se asumió un diseño muestral aleatorio simple sin reemplazamiento y así utilizar las fórmulas (3.14)-(3.16) y (3.23)-(3.28) de la sección 3.1.

Reemplazando las fórmulas (3.23)-(3.28) en las expresiones (3.10)-(3.13), la construcción de $\widehat{w}'s$ es posible. Para el caso particular, son presentados en la Tabla 4.9.

TABLA 4.9. Valor de los $\widehat{w}'s$ para cada imputación, con los datos de la ELCA.

	Imputación 1	Imputación 2	Imputación 3
\widehat{w}_1	0.568	0.586	0.579
\widehat{w}_2	0.305	0.309	0.280
\widehat{w}_3	0.127	0.105	0.141

Calculando el estimador definido en (3.14) para cada conjunto de imputaciones, se tiene que

$$\begin{aligned}\widehat{t}_{Y1} &= 1.013 \times 10^{13} \\ \widehat{t}_{Y2} &= 1.027 \times 10^{13} \\ \widehat{t}_{Y3} &= 9.947 \times 10^{12}\end{aligned}$$

Por tanto, según lo descrito en la sección 2.2.2, el estimador del total del salario para el año 2016 en la población identificada como ocupada es igual a

$$\widehat{t}_Y = \frac{1.013 \times 10^{13} + 1.027 \times 10^{13} + 9.947 \times 10^{12}}{3} = 1.012 \times 10^{13},$$

con una varianza estimada igual a

$$\widehat{V}(\widehat{t}_Y) = 1.348 \times 10^{13}.$$

El estimador del salario promedio para el año 2016 en los ocupados es igual a

$$\frac{\widehat{t}_Y}{N} = \frac{1.012 \times 10^{13}}{14722000} = 687173.5,$$

y su varianza estimada es

$$\widehat{V}\left(\frac{\widehat{t}_Y}{N}\right) = 918494.8.$$

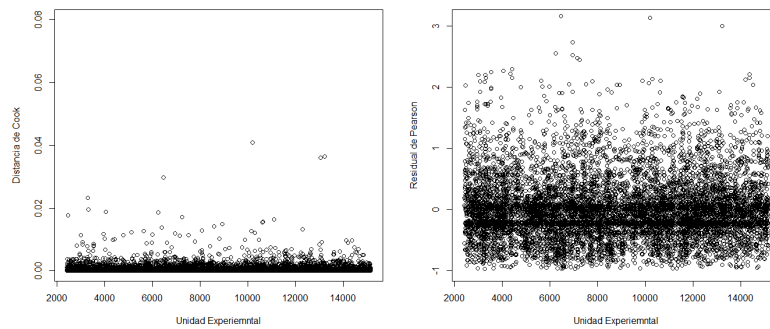
4.5. Conclusiones

Lo primero que cabe notar del ejercicio realizado con la ELCA es la calidad de los datos mismos, puesto que para la población identificada como ocupada en la variable más importante relacionada con el salario, la no respuesta fue de más del 50% y aún así se presentó en la base sin ningún tratamiento. Por lo cual, se decidió adoptar la metodología de imputación múltiple para obtener al menos un conjunto de datos completos.

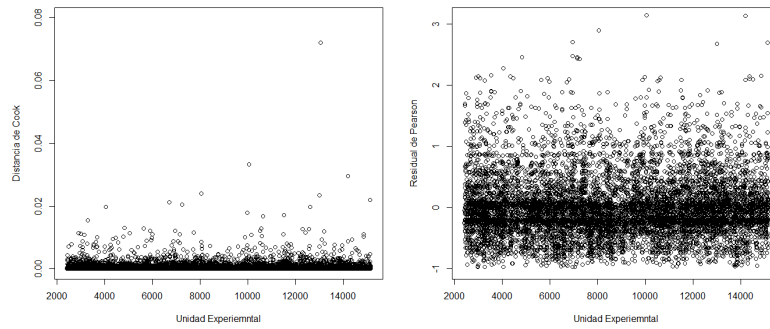
Además de esto, al comparar las cifras obtenidas de la ELCA con lo reportado por la GEIH para el año 2010, año en que se realizó un diseño muestral para el levantamiento de información buscando obtener una muestra representativa de la población colombiana de estratos uno al cuatro, las cifras difieren tanto en la proporción reportada de ocupados,

como en las características sociodemográficas de esta población, por ejemplo el estado civil.

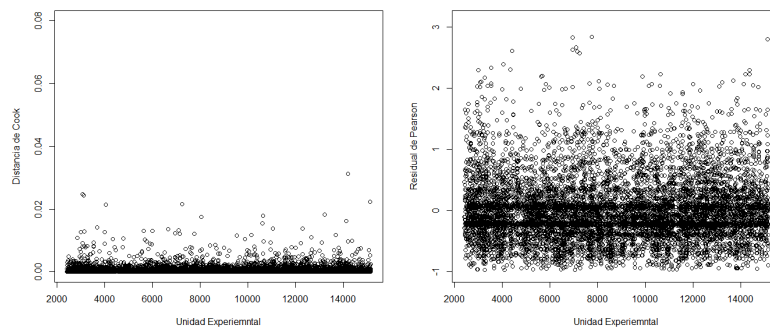
Finalmente, al ser la ELCA una encuesta que no cumple con los requisitos mínimos para la aplicación de la metodología y con lo expuesto anteriormente se llevó a desarrollar más que una aplicación, un *ejemplo práctico e ilustrativo* de la metodología propuesta en el capítulo 3, para lograr visualizar el resultado del estimador del total para un muestreo de tres ocasiones en datos reales.



(a) Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2010 y 2016 y/o 2013. Individuos eliminados: 7726, 12545, 14828, 14185. Imputación 1.

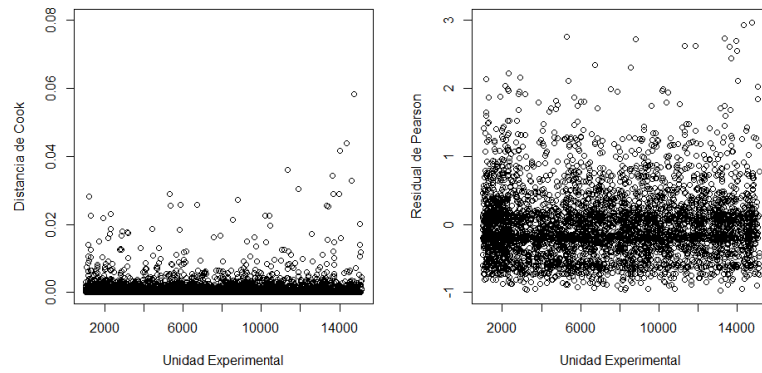


(b) Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2010 y 2016 y/o 2013. Individuos eliminados: 3998, 12545, 14828. Imputación 2.

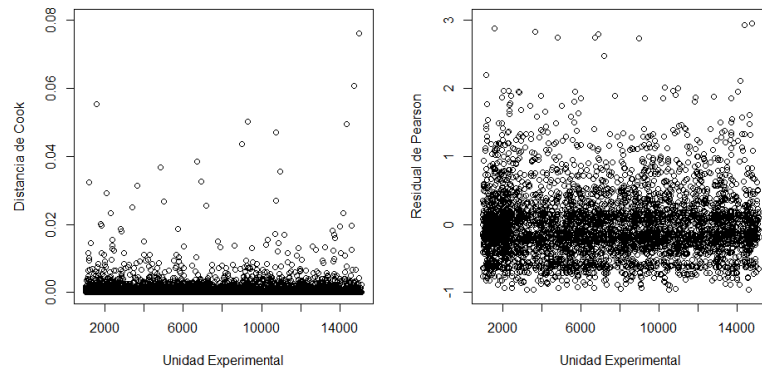


(c) Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2010 y 2016 y/o 2013. Individuos eliminados: 3996, 12545, 13027, 14828, 12549. Imputación 3.

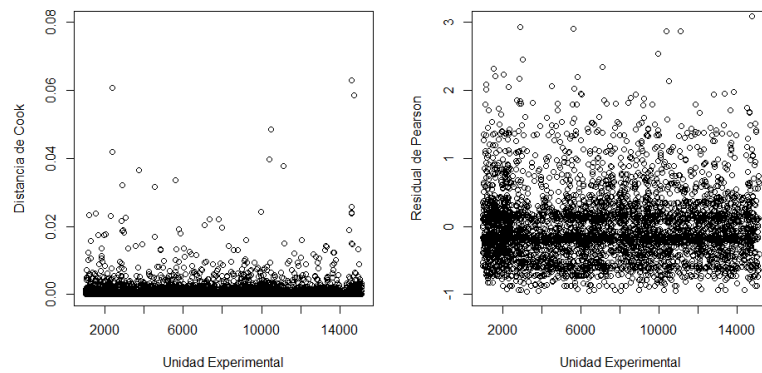
FIGURA 4.7. Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2010 y 2016 y/o 2013.



(a) Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2013 y 2016. Individuos eliminados: 1810, 3742. Imputación 1.



(b) Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2013 y 2016. Individuos eliminados: 3742, 14608, 14610. Imputación 2.



(c) Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2013 y 2016. Imputación 3.

FIGURA 4.8. Distancia de Cook y residuales de Pearson para el modelo aplicado a los individuos de la ELCA con observaciones en los años 2013 y 2016.

Simulación

Con el objetivo de conocer el comportamiento del estimador propuesto en el capítulo 3, se crearon tres universos con las variables: edad, género, tipo de ocupación y salario para tres años, basado en la información inicial obtenida de la ELCA, y variando entre cada universo la correlación presente entre tiempos. Así, el primer universo tenía una correlación alta ($\rho_{12} = 0.74, \rho_{13} = 0.61, \rho_{23} = 0.75$), el segundo una correlación baja ($\rho_{12} = 0.08, \rho_{13} = 0.11, \rho_{23} = 0.11$) y el tercero una correlación alta entre dos tiempos y baja entre los otros dos ($\rho_{12} = 0.74, \rho_{13} = 0.32, \rho_{23} = 0.13$).

Además de lo anterior, se crearon diferentes escenarios modificando el error muestral y la proporción de muestra emparejada (μ). Luego, para cada universo se trabajó con errores muestrales de 0.03, 0.04, 0.05 y valores de μ de 0.5, 0.6 y 0.7, generando un total de nueve escenarios por cada universo.

Aunque inicialmente se consideró trabajar con el total de la población de ocupados en Colombia, debido a limitaciones computacionales se trabajó con el total de personas ocupadas en la ciudad de Bogotá para el año 2017, obtenido de la Gran Encuesta Integrada de Hogares llevada a cabo por el Departamento Administrativo Nacional de Estadística en el año 2017.

Haciendo uso de la función `ss4m()` del paquete `samplesize4surveys` de R, se sacaron los tamaños de muestra de la primera, segunda y tercera ocasión (n_1, n_2 y n_3 respectivamente), con base en el error muestral definido y la desviación del salario en los tiempos uno, dos y tres de la información de la ELCA. Asimismo, para garantizar la existencia de s_{m_2}, s_{m_3} y s_n dentro de la simulación, se utilizaron los tamaños de muestra definidos en la Tabla 3.1.

Aplicando un diseño muestral aleatorio simple sin reemplazo para la selección de muestras en cada ocasión, usando las fórmulas presentadas en la sección 3.1 y obteniendo Y_{k3}^0 mediante una ecuación de estimación generalizada de respuesta cuasi gamma, matriz de correlación no estructurada y covariables: tiempo, edad, género y tipo de ocupación, a continuación se presentan los resultados obtenidos al comparar los siguientes estimadores,

- El estimador del total de Horvitz-Thompson para la tercera ocasión: \hat{t}_π
- El estimador del total para un muestreo de tres ocasiones: \hat{t}_Y

- El estimador del total para un muestreo de tres ocasiones con $\hat{w}'s: \hat{t}_Y$

Desde la teoría, tanto \hat{t}_π como \hat{t}_Y son insesgados. Sin embargo, debido a que \hat{t}_Y se construye usando $\hat{w}'s$ (ver (3.14)), éste presenta un sesgo. Luego, el interés es ver para \hat{t}_Y el comportamiento del sesgo relativo

$$Sesgo_{relativo} = \frac{B(\hat{\theta})}{\theta} = \frac{E(\hat{\theta}) - \theta}{\theta}$$

y la contribución relativa del sesgo al error cuadrático medio

$$\frac{B(\hat{\theta})^2}{ECM(\hat{\theta})} = \frac{B(\hat{\theta})^2}{V(\hat{\theta}) + B(\hat{\theta})^2}.$$

Mediante un proceso de simulación y con un total de 10000 réplicas, las fórmulas anteriores quedan expresadas para el proceso como,

$$Sesgo_{relativo} = \frac{\hat{B}(\hat{\theta})}{\theta} = \frac{\hat{E}(\hat{\theta}) - \theta}{\theta}$$

y

$$\frac{\hat{B}(\hat{\theta})^2}{\hat{V}(\hat{\theta}) + \hat{B}(\hat{\theta})^2},$$

en que $\hat{E}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_r$ y $\hat{V}(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^R [\hat{\theta}_r - \hat{E}(\hat{\theta})]^2$, donde $\hat{\theta}_r$ es la estimación de θ obtenida a partir de la r -ésima réplica de la simulación.

La Figura 5.1 evidencia la existencia de un sesgo en el estimador \hat{t}_Y . Para los tres universos, con un error muestral de 0.03 este sesgo llega a ser a lo más del orden de -0.00058, para un error muestral de 0.04 el sesgo relativo es máximo de -0.00041 mientras que para un error muestral de 0.05 alcanza a ser del orden de -0.0025.

La Figura 5.2 ayuda a contrastar lo visto anteriormente, mostrando que la contribución relativa del sesgo al error cuadrático medio es a lo más de -0.0035. Por lo tanto, es posible considerar el sesgo despreciable para el estimador \hat{t}_Y .

Bajo la consideración del sesgo despreciable para el estimador \hat{t}_Y , se realizó una comparación del coeficiente de variación

$$CV = \frac{\sqrt{ECM(\hat{\theta})}}{\hat{\theta}} \times 100$$

para los tres estimadores: \hat{t}_π , \hat{t}_Y y \hat{t}_Y . Para la simulación queda expresado como,

$$CV = \frac{\sqrt{\hat{V}(\hat{\theta})}}{\hat{E}(\hat{\theta})} \times 100,$$

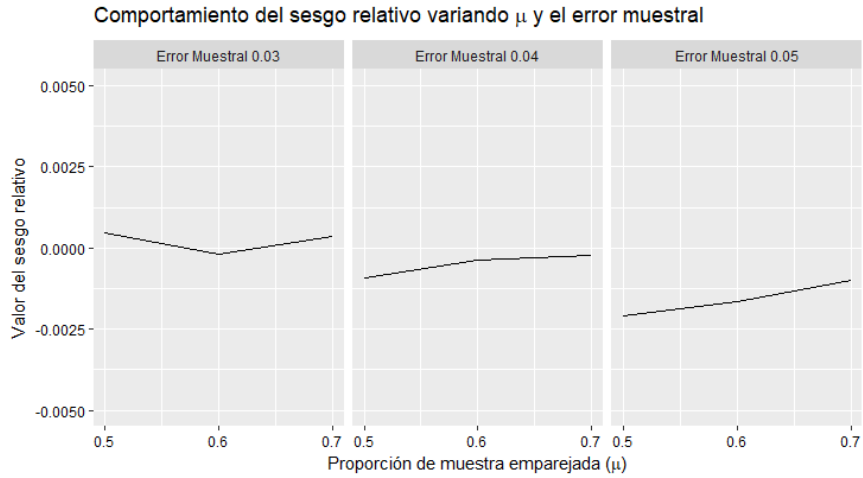
donde $\hat{E}(\hat{\theta})$ y $\hat{V}(\hat{\theta})$ están definidos anteriormente.

La Figura 5.3 muestra que sin importar el escenario el estimador de Horvitz-Thompson es el que mejor propiedades presenta. No obstante, a medida que aumenta la proporción de muestra emparejada, el estimador \hat{t}_Y se acerca bastante al estimador \hat{t}_π y más aún para el universo donde se consideran correlaciones entre tiempos altas y errores muestrales más bajos.

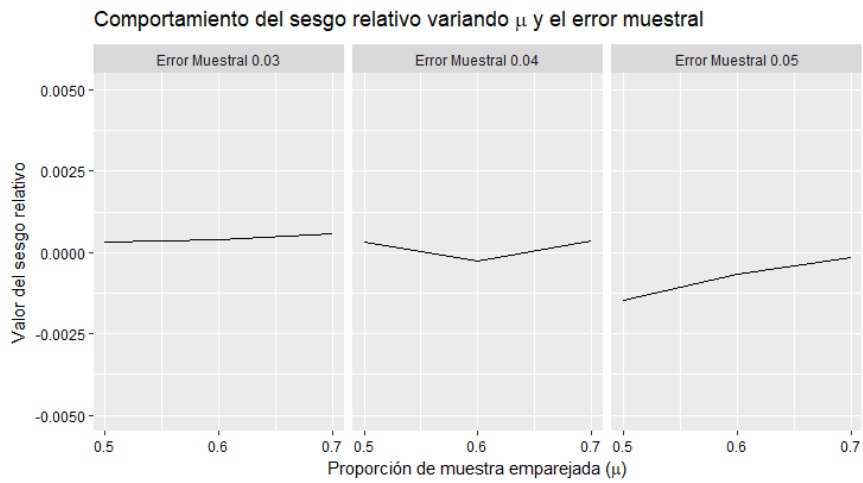
5.1. Conclusiones

Teniendo en cuenta que dentro del ejercicio de simulación se trabajó con un diseño muestral aleatorio simple sin reemplazo, una ecuación de estimación generalizada específica, y limitados valores de errores muestrales y proporciones de muestra emparejada, los resultados son:

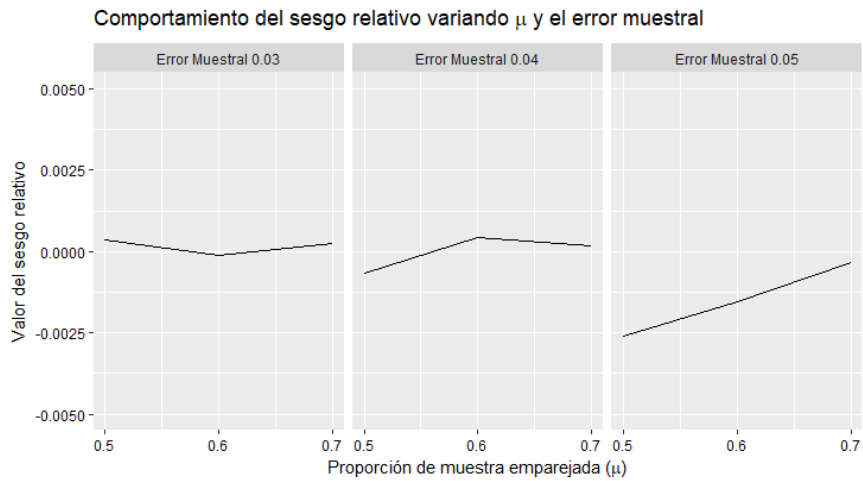
- Al estudiar la presencia de sesgo en el estimador para el total en un muestreo de tres ocasiones con $\hat{w}'s$, \hat{t}_Y , se mostró que hay presencia de sesgo aunque casi despreciable. Lo anterior se pudo ver a través del sesgo relativo y la contribución relativa del sesgo al error cuadrático medio.
- Bajo los escenarios simulados, al realizar una comparación del coeficiente de variación entre \hat{t}_π , \hat{t}_Y y $\hat{\hat{t}}_Y$, se observó que para la estimación del total en la tercera ocasión, el mejor estimador es el de Horvitz-Thompson.



(a) Sesgo relativo para el universo con correlaciones altas entre tiempos.

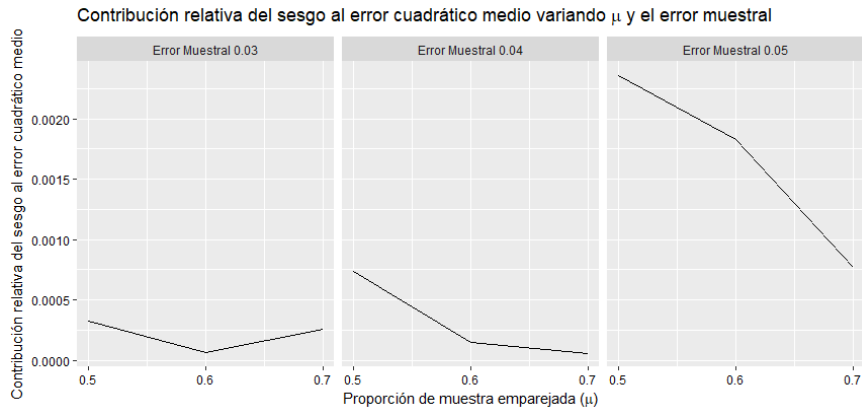


(b) Sesgo relativo para el universo con correlaciones bajas entre tiempos.

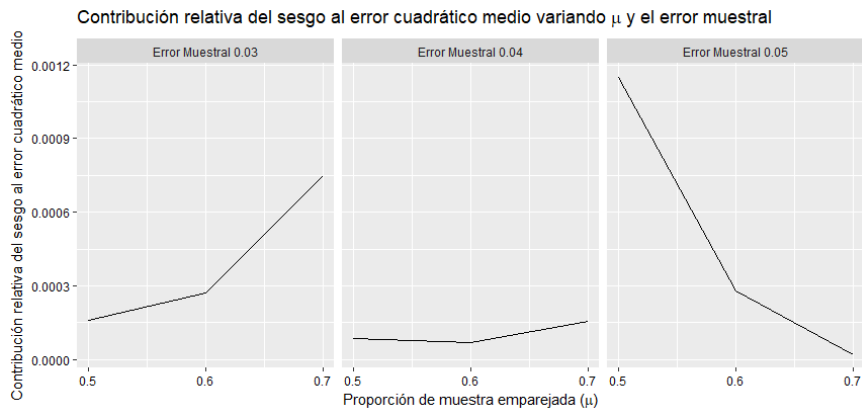


(c) Sesgo relativo para el universo con correlaciones bajas y altas entre tiempos.

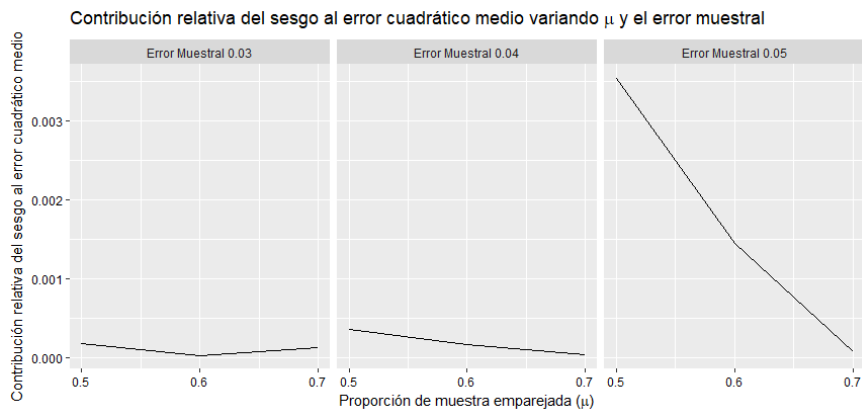
FIGURA 5.1. Sesgo relativo de \widehat{t}_Y variando error muestral y proporción de muestra emparejada (μ) para cada uno de los universos.



(a) Contribución relativa del sesgo al error cuadrático medio para el universo con correlaciones altas entre tiempos.

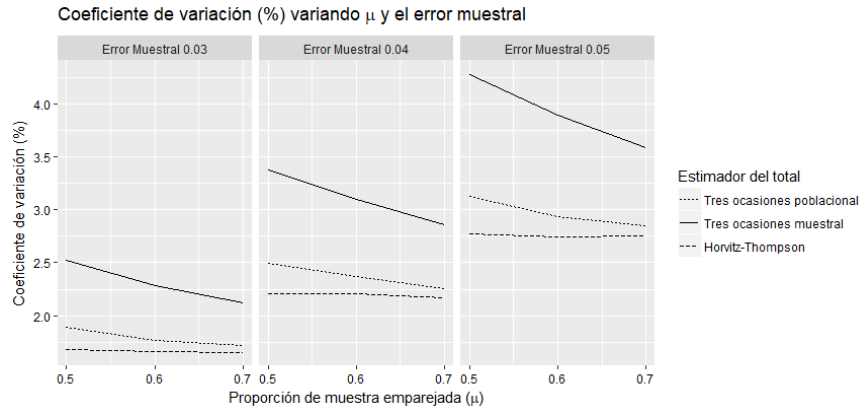


(b) Contribución relativa del sesgo al error cuadrático medio para el universo con correlaciones bajas entre tiempos.

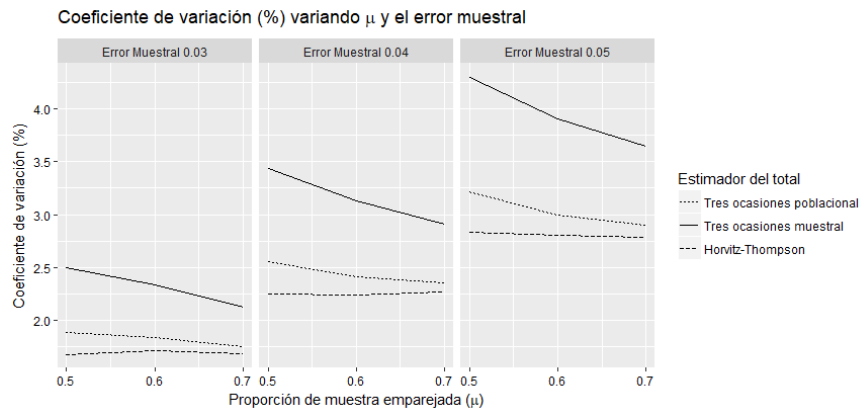


(c) Contribución relativa del sesgo al error cuadrático medio para el universo con correlaciones bajas y altas entre tiempos.

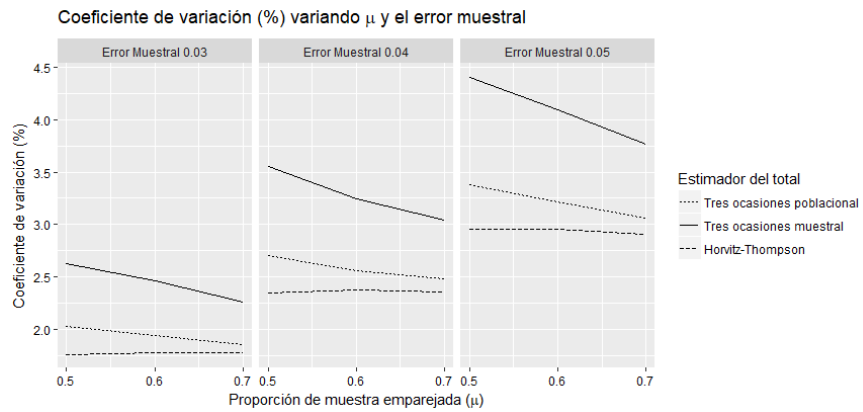
FIGURA 5.2. Contribución relativa del sesgo al error cuadrático medio de \hat{t}_Y variando error muestral y proporción de muestra emparejada (μ) para cada uno de los universos.



(a) Coeficiente de variación para el universo con correlaciones altas entre tiempos.



(b) Coeficiente de variación para el universo con correlaciones bajas entre tiempos.



(c) Coeficiente de variación para el universo con correlaciones bajas y altas entre tiempos.

FIGURA 5.3. Coeficiente de variación para \hat{t}_π , \hat{t}_Y y $\hat{\hat{t}}_Y$ variando error muestral y proporción de muestra emparejada (μ) para cada uno de los universos.

Conclusiones y trabajo futuro

En este trabajo presentamos una adaptación del estimador del total para un muestreo de dos ocasiones propuesto por Särndal (1992)[p. 369] a un diseño muestral de tres ocasiones, motivados por la información disponible de la Encuesta Longitudinal Colombiana de la Universidad de los Andes (ELCA). Dentro de las modificaciones hechas, estuvo el ajuste de Ecuaciones de Estimación Generalizadas (EEG) propuestas por Liang & Zeger (1986) para predecir la variable de interés, en vez de utilizar una aproximación lineal como la descrita por Särndal (1992)[p. 370].

El estimador propuesto fue aplicado al conjunto de datos disponibles, definiendo como variable de interés el salario de la población identificada como ocupada. No obstante, al ser el salario una pregunta sensible, la no respuesta fue evidente en el manejo de la información y como solución para contrarrestar este problema, se usó la metodología de imputación múltiple desarrollada por Rubin (1977).

Además, por la naturaleza de los datos y el proceso de recolección de los mismos, la aplicación se enfocó más hacia un ejemplo práctico y no un trabajo del cual fuera posible realizar inferencias poblacionales en donde el diseño muestral utilizado fuese un muestreo aleatorio simple sin reemplazamiento (MAS).

Otro punto importante a resaltar es la simulación realizada en el capítulo 5 para ver las propiedades del estimador propuesto y a su vez compararlo con el estimador de Horvitz-Thompson. En ella, diferentes universos fueron simulados variando la correlación entre tiempos y generando diversos escenarios al cambiar el error muestral y la proporción de muestra emparejada. Entre lo observado está que el sesgo del estimador \hat{t}_Y es tan pequeño que puede considerarse despreciable.

Al realizar la comparación entre estimadores para ver cuál tenía menor coeficiente de variación, en todos los escenarios el estimador de Horvitz-Thompson tuvo mejor comportamiento. Sin embargo, cabe aclarar que esta comparación no es adecuada debido a que el estimador de Horvitz-Thompson para el total de la tercera ocasión no tiene en cuenta la información de las ocasiones anteriores, perdiendo de vista el diseño muestral bajo el cual se llega a la información recolectada para Y_3 .

El estimador para el total de tres ocasiones \hat{t}_Y , necesita el conocimiento de varianzas y covarianzas poblacionales lo que limita su uso en la práctica.

Las rutinas desarrolladas para la aplicación realizada en el capítulo 4 o en la simulación presentada en el capítulo 5, se presentan de forma general y pueden ser aplicadas a otro conjunto de datos.

Expuesto lo anterior, para trabajos futuros queda:

- Realizar la aplicación de la metodología desarrollada a otro conjunto de datos diferente al utilizado en este trabajo. Lo anterior, debido que al comparar los resultados del año base de la ELCA con lo expuesto por la Gran Encuesta Integrada de Hogares (GEIH) para la población ocupada en el año 2010, se evidenciaron diferencias desde la proporción de ocupados hasta en características demográficas, como el estado civil. No obstante, en caso de utilizar la misma base de la ELCA, buscar calibrar la información con los datos poblacionales de la GEIH.
- Estudiar el efecto de obtener la aproximación de la variable de interés mediante una combinación de lo propuesto por Särndal en los casos donde solo se tiene información de una ocasión previa, y aplicar EEG para los casos en que se cuente con más de una ocasión anterior.
- Analizar el comportamiento del estimador del total para un muestreo de tres ocasiones, considerando ajustar EEG con diferentes respuestas y estructuras de correlación, dado que tanto en el ejemplo práctico como en la simulación, se consideró el ajuste con una respuesta cuasi gamma y estructura de correlación no estructurada.
- Considerar desde la teoría y dentro de la simulación diseños muestrales más complejos al utilizado, ya que en el desarrollo del trabajo solo se utilizó MAS. No obstante, bajo este diseño muestral fue posible observar desde las expresiones teóricas de las varianzas, la diferencia existente entre la varianza del estimador \hat{t}_Y con respecto a \hat{t}_π , teniendo este último menor varianza.
- Dentro del planteamiento hecho para el trabajo de simulación, estudiar el comportamiento del coeficiente de variación para el estimador \hat{t}_Y y $\hat{\hat{t}}_Y$ al aumentar la proporción de muestra emparejada y con errores muestrales bajos. Nuevamente compararlos con \hat{t}_π .
- Evaluar si trabajar con variables menos volátiles y sensibles, como fue salario, conllevan a las mismas conclusiones presentadas en este trabajo, en cuanto al comportamientos de los tres estimadores, \hat{t}_π , \hat{t}_Y y $\hat{\hat{t}}_Y$.

APÉNDICE A

Demostraciones

Retomando lo definido en el capítulo 3, en la Tabla A.1 se muestran las probabilidades de inclusión de primer y segundo orden, así como lo Δ_{kl} para cada muestra.

TABLA A.1. Probabilidades de inclusión de primer orden, segundo orden y Δ_{kl} para cada muestra.

Muestra	Prob. inclusión Primer orden	Prob. inclusión Segundo orden	Δ_{kl}
s_a	π_{ak}	π_{akl}	Δ_{akl}
s_u	$\pi_{k s_a^c}$	$\pi_{kl s_a^c}$	$\Delta_{kl s_a^c}$
s_{m_2}	$\pi_{k s_{m_2}}$	$\pi_{kl s_{m_2}}$	$\Delta_{kl s_{m_2}}$
s_{m_3}	$\pi_{k s_u}$	$\pi_{kl s_u}$	$\Delta_{kl s_u}$
s_n	$\pi_{k s^c}$	$\pi_{kl s^c}$	$\Delta_{kl s^c}$

Además, $(\pi_{ak})^c = 1 - \pi_{ak}$ y $(\pi_{sk})^c = 1 - \pi_{sk}$ con $s = s_a \cup s_u$ y $s^c = (s_a \cup s_u)^c$. Sean

$$I_1(k) = \begin{cases} 1 & \text{si } k \in s_{m_2}, k \in s_a \\ 0 & \text{e.o.c} \end{cases}, \quad I_2(k) = \begin{cases} 1 & \text{si } k \in s_{m_3}, k \in s_u \\ 0 & \text{e.o.c} \end{cases}$$

y

$$I_3(k) = \begin{cases} 1 & \text{si } k \in s_n, k \in s^c \\ 0 & \text{e.o.c} \end{cases}$$

Se tiene que $E(I_1(k)) = \pi_{k|s_{m_2}}$, $E(I_2(k)) = \pi_{k|s_u}$ y $E(I_3(k)) = \pi_{k|s^c}$.

A continuación se presentan las demostraciones de esperanza, varianza y covarianza correspondientes al capítulo 3.

A.1. Esperanza

- Para \hat{t}_1 se tiene que,

$$\begin{aligned}
 E(\hat{t}_1) &= E_{s_a} [E(\hat{t}_1 | s_a)] \\
 &= E_{s_a} \left[E \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_{m_2}} \frac{D_k}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &= E_{s_a} \left[E \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \middle| s_a \right) + E \left(\sum_{s_a} \frac{D_k I_1(k)}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &= E_{s_a} \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_a} \frac{D_k}{\pi_{ak}} \right) \\
 &= E_{s_a} \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_a} \frac{Y_k}{\pi_{ak}} - \sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \right) \\
 &= E_{s_a} \left(\sum_{s_a} \frac{Y_k}{\pi_{ak}} \right) \\
 &= \sum_U Y_k \\
 &= t
 \end{aligned}$$

- La esperanza de \hat{t}_2 está dada por,

$$\begin{aligned}
 E(\hat{t}_2) &= E_{s_u} [E(\hat{t}_2 | s_u)] \\
 &= E_{s_u} \left[E \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_{m_3}} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &= E_{s_u} \left[E \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \middle| s_u \right) + E \left(\sum_{s_u} \frac{D_k I_2(k)}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &= E_{s_u} \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_u} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \right) \\
 &= E_{s_u} \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_u} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} - \sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \right) \\
 &= E_{s_u} \left(\sum_{s_u} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \right) \\
 &= \sum_U Y_k \\
 &= t
 \end{aligned}$$

- Para \widehat{t}_3 la esperanza es,

$$\begin{aligned}
 E(\widehat{t}_3) &= E_{s^c} [E(\widehat{t}_3 | s^c)] \\
 &= E_{s^c} \left[E \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{k|s^c}} \middle| s^c \right) \right] \\
 &= E_{s^c} \left[E \left(\sum_{s^c} \frac{Y_k I_3(k)}{(\pi_{sk})^c \pi_{k|s^c}} \middle| s^c \right) \right] \\
 &= E_{s^c} \left(\sum_{s^c} \frac{Y_k}{(\pi_{sk})^c} \right) \\
 &= \sum_U Y_k \\
 &= t
 \end{aligned}$$

A.2. Varianza

- Reescribiendo la varianza de \widehat{t}_1 como

$$V(\widehat{t}_1) = \underbrace{V_{s_a} [E(\widehat{t}_1 | s_a)]}_{V1.1} + \underbrace{E_{s_a} [V(\widehat{t}_1 | s_a)]}_{V1.2}$$

Se tiene que,

$$\begin{aligned}
 V1.1 &= V_{s_a} [E(\widehat{t}_1 | s_a)] \\
 &= V_{s_a} \left[E \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_{m_2}} \frac{D_k}{\pi_{ak} \pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &= V_{s_a} \left[E \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \middle| s_a \right) + E \left(\sum_{s_a} \frac{D_k I_1(k)}{\pi_{ak} \pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &= V_{s_a} \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_a} \frac{D_k}{\pi_{ak}} \right) \\
 &= V_{s_a} \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_a} \frac{Y_k}{\pi_{ak}} - \sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \right) \\
 &= V_{s_a} \left(\sum_{s_a} \frac{Y_k}{\pi_{ak}} \right) \\
 &= \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{\pi_{al}}
 \end{aligned} \tag{A.1}$$

$$\begin{aligned}
 V1.2 &= E_{s_a} [V(\hat{t}_1 | s_a)] \\
 &= E_{s_a} \left[V \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_{m_2}} \frac{D_k}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &= E_{s_a} \left[V \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_a} \frac{D_k I_1(k)}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &= E_{s_a} \left[\underbrace{V \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \middle| s_a \right)}_1 + V \left(\sum_{s_a} \frac{D_k I_1(k)}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &\quad + E_{s_a} \left[\underbrace{2C \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}}, \sum_{s_a} \frac{D_k I_1(k)}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right)}_2 \right] \\
 &= E_{s_a} \left[V \left(\sum_{s_a} \frac{D_k I_1(k)}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right) \right] \\
 &= E_{s_a} \left(\sum \sum_{s_a} \Delta^{kl|s_{m_2}} \frac{D_k}{\pi_{ak}\pi_{k|s_{m_2}}} \frac{D_l}{\pi_{al}\pi_{l|s_{m_2}}} \right) \tag{A.2}
 \end{aligned}$$

Luego, por (A.1) y (A.2)

$$V(\hat{t}_1) = \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{\pi_{al}} + E_{s_a} \left(\sum \sum_{s_a} \Delta^{kl|s_{m_2}} \frac{D_k}{\pi_{ak}\pi_{k|s_{m_2}}} \frac{D_l}{\pi_{al}\pi_{l|s_{m_2}}} \right)$$

- Reescribiendo la varianza de \hat{t}_2 como

$$V(\hat{t}_2) = \underbrace{V_{s_u} [E(\hat{t}_2 | s_u)]}_{V2.1} + \underbrace{E_{s_u} [V(\hat{t}_2 | s_u)]}_{V2.2}$$

1

$$V \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \middle| s_a \right) = C \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}}, \sum_{s_a} \frac{Y_l^0}{\pi_{al}} \middle| s_a \right) = \sum \sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \frac{Y_l^0}{\pi_{al}} C(1, 1 | s_a) = 0$$

2

$$C \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}}, \sum_{s_a} \frac{D_k I_1(k)}{\pi_{ak}\pi_{k|s_{m_2}}} \middle| s_a \right) = \sum \sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \frac{D_l}{\pi_{al}\pi_{l|s_{m_2}}} C(1, I_1(l) | s_a) = 0$$

Se tiene que,

$$\begin{aligned}
 V2.1 &= V_{s_u} [E(\hat{t}_2 | s_u)] \\
 &= V_{s_u} \left[E \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_{m_3}} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &= V_{s_u} \left[E \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \middle| s_u \right) + E \left(\sum_{s_u} \frac{D_k I_2(k)}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &= V_{s_u} \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_u} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \right) \\
 &= V_{s_u} \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_u} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} - \sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \right) \\
 &= V_{s_u} \left(\sum_{s_u} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \right) \\
 &= \sum \sum_U \Delta_{kl|s_a^c} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l}{(\pi_{al})^c \pi_{l|s_a^c}} \tag{A.3}
 \end{aligned}$$

$$\begin{aligned}
 V2.2 &= E_{s_u} [V(\hat{t}_2 | s_u)] \\
 &= E_{s_u} \left[V \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_{m_3}} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &= E_{s_u} \left[V \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} + \sum_{s_u} \frac{D_k I_2(k)}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &= E_{s_u} \left[\underbrace{V \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \middle| s_u \right)}_3 + V \left(\sum_{s_u} \frac{D_k I_2(k)}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &+ E_{s_u} \left[\underbrace{2C \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}}, \sum_{s_u} \frac{D_k I_2(k)}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right)}_4 \right] \\
 &= E_{s_u} \left[V \left(\sum_{s_u} \frac{D_k I_2(k)}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \right] \\
 &= E_{s_u} \left(\sum \sum_{s_u} \Delta_{kl|s_u} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \frac{D_l}{(\pi_{al})^c \pi_{l|s_a^c} \pi_{l|s_u}} \right) \tag{A.4}
 \end{aligned}$$

Luego, por (A.3) y (A.4)

$$V(\hat{t}_2) = \sum \sum_U \Delta_{kl|s_a^c} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l}{(\pi_{al})^c \pi_{l|s_a^c}} \\ + E_{s_u} \left(\sum \sum_{s_u} \Delta_{kl|s_u} \frac{D_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \frac{D_l}{(\pi_{al})^c \pi_{l|s_a^c} \pi_{l|s_u}} \right)$$

- Reescribiendo la varianza de \hat{t}_3 como

$$V(\hat{t}_3) = \underbrace{V_{s^c} [E(\hat{t}_3|s^c)]}_{V3.1} + \underbrace{E_{s^c} [V(\hat{t}_3|s^c)]}_{V3.2}$$

Se tiene que,

$$\begin{aligned} V3.1 &= V_{s^c} [E(\hat{t}_3|s^c)] \\ &= V_{s^c} \left[E \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{k|s^c}} \middle| s^c \right) \right] \\ &= V_{s^c} \left[E \left(\sum_{s^c} \frac{Y_k I_3(k)}{(\pi_{sk})^c \pi_{k|s^c}} \middle| s^c \right) \right] \\ &= V_{s^c} \left(\sum_{s^c} \frac{Y_k}{(\pi_{sk})^c} \right) \\ &= \sum \sum_U \Delta_{skl}^c \frac{Y_k}{(\pi_{sk})^c} \frac{Y_l}{(\pi_{sl})^c} \end{aligned} \tag{A.5}$$

$$\begin{aligned} V3.2 &= E_{s^c} [V(\hat{t}_3|s^c)] \\ &= E_{s^c} \left[V \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{k|s^c}} \middle| s^c \right) \right] \\ &= E_{s^c} \left[V \left(\sum_{s^c} \frac{Y_k I_3(k)}{(\pi_{sk})^c \pi_{k|s^c}} \middle| s^c \right) \right] \\ &= E_{s^c} \left(\sum \sum_{s^c} \Delta_{kl|s^c} \frac{Y_k}{(\pi_{sk})^c \pi_{k|s^c}} \frac{Y_l}{(\pi_{sl})^c \pi_{l|s^c}} \right) \end{aligned} \tag{A.6}$$

3

$$\begin{aligned} V \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \middle| s_u \right) &= C \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}}, \sum_{s_u} \frac{Y_l^0}{(\pi_{al})^c \pi_{l|s_a^c}} \middle| s_u \right) \\ &= \sum \sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l^0}{(\pi_{al})^c \pi_{l|s_a^c}} C(1, 1 | s_u) = 0 \end{aligned}$$

4

$$\begin{aligned} &C \left(\sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}}, \sum_{s_u} \frac{D_k I_2(k)}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right) \\ &= \sum \sum_{s_u} \frac{Y_k^0}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{D_l}{(\pi_{al})^c \pi_{l|s_a^c} \pi_{l|s_u}} C(1, I_2(l) | s_u) = 0 \end{aligned}$$

Luego, por (A.5) y (A.6)

$$V(\hat{t}_3) = \sum \sum_U \Delta_{skl}^c \frac{Y_k}{(\pi_{sk})^c} \frac{Y_l}{(\pi_{sl})^c} + E_{s^c} \left(\sum \sum_{s^c} \Delta_{kl|s^c} \frac{Y_k}{(\pi_{sk})^c \pi_{k|s^c}} \frac{Y_l}{(\pi_{sl})^c \pi_{l|s^c}} \right)$$

A.3. Covarianza

Sean

$$I_{s_a}(k) = \begin{cases} 1 & \text{si } k \in s_a \\ 0 & \text{e.o.c} \end{cases} \text{ y } I_{s_u}(k) = \begin{cases} 1 & \text{si } k \in s_u \\ 0 & \text{e.o.c} \end{cases}$$

- Considerando la covarianza entre \hat{t}_1 y \hat{t}_3 como

$$C(\hat{t}_1, \hat{t}_3) = \underbrace{E_{s_a} [C(\hat{t}_1, \hat{t}_3 | s_a)]}_{C1.1} + \underbrace{C [E_{s_a}(\hat{t}_1 | s_a), E_{s_a}(\hat{t}_3 | s_a)]}_{C1.2}$$

$$\begin{aligned} C1.1 &= E_{s_a} [C(\hat{t}_1, \hat{t}_3 | s_a)] \\ &= E_{s_a} \left[C \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_{m_2}} \frac{D_k}{\pi_{ak} \pi_{k|s_{m_2}}}, \sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_a \right) \right] \\ &= 0 \end{aligned} \tag{A.7}$$

Lo anterior puesto que $s_a \cap s_n = \emptyset$. Ahora,

$$\begin{aligned} C1.2 &= C [E_{s_a}(\hat{t}_1 | s_a), E_{s_a}(\hat{t}_3 | s_a)] \\ &= C \left[E_{s_a} \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} + \sum_{s_{m_2}} \frac{D_k}{\pi_{ak} \pi_{k|s_{m_2}}} \middle| s_a \right), E_{s_a} \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_a \right) \right] \\ &= C \left[E_{s_a} \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \middle| s_a \right), E_{s_a} \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_a \right) \right] \\ &+ C \left[E_{s_a} \left(\sum_{s_{m_2}} \frac{D_k}{\pi_{ak} \pi_{k|s_{m_2}}} \middle| s_a \right), E_{s_a} \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_a \right) \right] \\ &= C \left[E_{s_a} \left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}} \middle| s_a \right), E_{s_a} \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_a \right) \right] \\ &+ C \left[E_{s_a} \left(\sum_{s_{m_2}} \frac{Y_k}{\pi_{ak} \pi_{k|s_{m_2}}} \middle| s_a \right), E_{s_a} \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_a \right) \right] \\ &- C \left[E_{s_a} \left(\sum_{s_{m_2}} \frac{Y_k^0}{\pi_{ak} \pi_{k|s_{m_2}}} \middle| s_a \right), E_{s_a} \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_a \right) \right] \end{aligned}$$

$$\begin{aligned}
 &= C\left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}}, \sum_{s_a^c} \frac{Y_k}{(\pi_{ak})^c}\right) + C\left(\sum_{s_a} \frac{Y_k}{\pi_{ak}}, \sum_{s_a^c} \frac{Y_k}{(\pi_{ak})^c}\right) - C\left(\sum_{s_a} \frac{Y_k^0}{\pi_{ak}}, \sum_{s_a^c} \frac{Y_k}{(\pi_{ak})^c}\right) \\
 &= C\left(\sum_{s_a} \frac{Y_k}{\pi_{ak}}, \sum_{s_a^c} \frac{Y_k}{(\pi_{ak})^c}\right) \\
 &= \sum \sum_U \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c} C(I_{s_a}(k), 1 - I_{s_a}(l)) \\
 &= - \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c}
 \end{aligned} \tag{A.8}$$

Por tanto, haciendo uso de (A.7) y (A.8) se tiene que

$$C(\hat{t}_1, \hat{t}_3) = - \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c}$$

- Análogo a lo hecho anteriormente, y considerando la covarianza entre \hat{t}_1 y \hat{t}_2 como

$$C(\hat{t}_1, \hat{t}_2) = \underbrace{E_{s_a} [C(\hat{t}_1, \hat{t}_2 | s_a)]}_{C2.1} + \underbrace{C [E_{s_a}(\hat{t}_1 | s_a), E_{s_a}(\hat{t}_2 | s_a)]}_{C2.2}$$

se tiene que

$$C2.1 = E_{s_a} [C(\hat{t}_1, \hat{t}_2 | s_a)] = 0, \tag{A.9}$$

puesto que $s_a \cap s_u = \emptyset$. Ahora,

$$\begin{aligned}
 C2.2 &= C [E_{s_a}(\hat{t}_1 | s_a), E_{s_a}(\hat{t}_3 | s_a)] \\
 &= C \left[E_{s_a} \left(\sum_{s_{m_2}} \frac{Y_k}{\pi_{ak} \pi_{k|s_{m_2}}} \middle| s_a \right), E_{s_a} \left(\sum_{s_{m_3}} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_a \right) \right] \\
 &= C \left(\sum_{s_a} \frac{Y_k}{\pi_{ak}}, \sum_{s_a^c} \frac{Y_k}{(\pi_{ak})^c} \right) \\
 &= \sum \sum_U \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c} C(I_{s_a}(k), 1 - I_{s_a}(l)) \\
 &= - \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c}
 \end{aligned} \tag{A.10}$$

Así, haciendo uso de (A.9) y (A.10) se tiene que

$$C(\hat{t}_1, \hat{t}_2) = - \sum \sum_U \Delta_{akl} \frac{Y_k}{\pi_{ak}} \frac{Y_l}{(\pi_{al})^c}$$

- Ahora, considerando

$$C(\hat{t}_2, \hat{t}_3) = \underbrace{E_{s_u} [C(\hat{t}_2, \hat{t}_3)]}_{C3.1} + \underbrace{C [E_{s_u}(\hat{t}_2 | s_u), E_{s_u}(\hat{t}_3 | s_u)]}_{C3.2}$$

y siguiendo el razonamiento anterior, se tiene que al ser $s_u \cap s_n = \emptyset$, entonces

$$C3.1 = E_{s_u} [C(\hat{t}_2, \hat{t}_3)] = 0, \tag{A.11}$$

$$\begin{aligned}
 C3.2 &= C [E_{s_u}(\hat{t}_2 | s_u), E_{s_u}(\hat{t}_3 | s_u)] \\
 &= C \left[E_{s_u} \left(\sum_{s_{m_3}} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c} \pi_{k|s_u}} \middle| s_u \right), E_{s_u} \left(\sum_{s_n} \frac{Y_k}{(\pi_{sk})^c \pi_{l|s^c}} \middle| s_u \right) \right] \\
 &= C \left(\sum_{s_u} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}}, \sum_{s^c} \frac{Y_k}{(\pi_{ak})^c (\pi_{k|s_a^c})^c} \right) \\
 &= \sum \sum_U \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l}{(\pi_{al})^c (\pi_{l|s_a^c})^c} C(I_{s_u}(k), 1 - I_{s_u}(k)) \\
 &= - \sum \sum_U \Delta_{kl|s_a^c} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l}{(\pi_{al})^c (\pi_{l|s_a^c})^c}
 \end{aligned} \tag{A.12}$$

Luego, haciendo uso de (A.11) y (A.12) se tiene que

$$C(\hat{t}_2, \hat{t}_3) = - \sum \sum_U \Delta_{kl|s_a^c} \frac{Y_k}{(\pi_{ak})^c \pi_{k|s_a^c}} \frac{Y_l}{(\pi_{al})^c (\pi_{l|s_a^c})^c}$$

Tamaños de muestra para la tercera ocasión

Con el objetivo de asegurar dentro de la simulación una muestra de individuos para s_{m_2} , s_{m_3} y s_n , y buscando mantener una proporción de tamaños se planteó lo siguiente.

Como se observa en la Figura B.1, se identificaron los subconjuntos obtenidos en la tercera ocasión y su relación con las ocasiones anteriores. Así, a^* hace referencia al emparejamiento entre la primera ocasión y tercera, b^* identifica la proporción de muestra de la primera, segunda y tercera ocasión y c^* el emparejamiento existente entre la segunda y tercera ocasión.

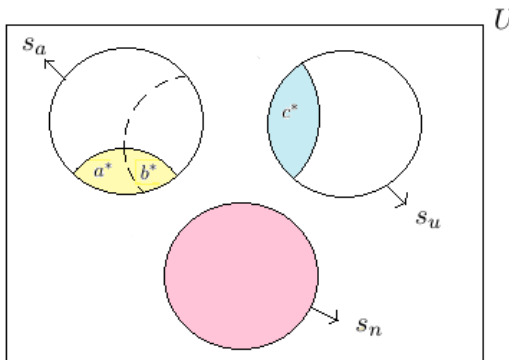


FIGURA B.1. Muestras de la tercera ocasión relacionadas con las ocasiones anteriores.

Teniendo en cuenta que μ es la proporción de muestra emparejada, se tiene que

$$a^* + b^* + c^* = \mu n_3.$$

Ahora, con el fin de mantener la proporción entre b^* y c^* como la existente entre s_{m_1} y s_u se plantea,

$$\begin{aligned} b^* &= \mu(b^* + c^*) \\ b^* &= \mu b^* + \mu c^* \\ (1 - \mu)b^* &= \mu c^* \\ b^* &= \frac{\mu c^*}{1 - \mu} \end{aligned}$$

Manteniendo la proporción entre a^* y $a^* + b^*$, como la que existe entre $s_a - s_{m_1}$ y s_a se tiene,

$$\begin{aligned} \frac{a^*}{a^* + b^*} &= \frac{n_1 - \mu n_2}{n_1} \\ a^* &= \frac{n_1 - \mu n_2}{n_1} a^* + \frac{n_1 - \mu n_2}{n_1} b^* \\ a^* &= \frac{n_1 - \mu n_2}{\mu n_2} b^* \end{aligned}$$

Y finalmente,

$$\begin{aligned} a^* + b^* + c^* &= \mu n_3 \\ \frac{c^*}{a^* + b^* + c^*} &= \frac{c^*}{\mu n_3} = \frac{(1 - \mu)n_2}{n_1 + (1 - \mu)n_2} \\ c^* &= \frac{\mu n_3 (1 - \mu)n_2}{n_1 + (1 - \mu)n_2}. \end{aligned}$$

Nota: Los valores utilizados en la sección 3.1 son:

$$a = a^* + b^*$$

y

$$b = c^*.$$

Código desarrollado

El código presentado a continuación, fue el utilizado en el software R para el cálculo del estimador del total en un muestreo de tres ocasiones, con los datos de la ELCA y la simulación.

Sea PO la muestra notada como s_a y obtenida en la primera ocasión. Sean S012 y S02 las muestras obtenidas de la segunda ocasión y denotadas por s_{m_1} y s_u , respectivamente. Sea T013 la muestra de la tercera ocasión seleccionada de s_a y T0123 la muestra de la tercera ocasión seleccionada de s_{m_1} , la unión de las dos muestras forman s_{m_2} . La muestra T023 es la denotada por s_{m_3} y T03 es s_n . Estas dos últimas también son obtenidas en la tercera ocasión.

```
MuestraAux <- list ()
MuestraAux[[1]] <- Muestra[Tiempo == 2 & Id %n% SO2]
MuestraAux[[2]] <- Muestra[Tiempo == 2 & Id %n% SO]
MuestraAux[[3]] <- Muestra[Tiempo == 3 & Id %n% TO]
MuestraAux[[4]] <- Muestra[Tiempo == 3 & Id %n% TO3]
MuestraAux[[5]] <- Muestra[(Tiempo == 1 & Id %n% PO) |
                             (Tiempo == 2 & Id %n% SO12)]
MuestraAux[[6]] <- Muestra[(Tiempo == 1 & Id %n% PO) |
                             (Tiempo %n% c(2, 3) & Id %n% TO123) |
                             (Tiempo == 3 & Id %n% TO13)]
MuestraAux[[7]] <- Muestra[(Tiempo == 2 & Id %n% SO2) |
                             (Tiempo == 3 & Id %n% TO23)]

mediaP <- c() # predicción
mediaD <- c() # diferencia
mediaS <- c() # salario observado
varP <- c()
varD <- c()
varS <- c()

for (i in c(1:4)) {
  mediaS[i] <- mean(MuestraAux[[i]]$Salario)
  varS[i] <- var(MuestraAux[[i]]$Salario)
}

for (i in c(5:7)) {
  tFit <- as.integer(ifelse(i == 5, 2, 3))
  IdsGEE <- MuestraAux[[i]][Tiempo == tFit, Id]
  datosAjuste <- MuestraAux[[i]][Id %n% IdsGEE]
  datosDiferencia <- datosAjuste[Tiempo == tFit]
  datosPrediccion <- unique(MuestraAux[[i]][, list (Id,
                                                    Tiempo = tFit,
                                                    Edad = Edad + (tFit - Tiempo)*3,
```

```

                                Género)))
modelo <- gee(formula = Salario ~ Tiempo + Edad + Género,
              family = Gamma(link = "log"),
              data = datosAjuste,
              id = Id,
              corstr = "unstructured")
predichos <- proyecta(datos = datosPrediccion,
                    coeficientes = modelo$coefficients)

mediaP[i-4] <- mean(predichos)
mediaD[i-4] <- mean(datosDiferencia[, Salario] -
                  predichos[datosPrediccion[, Id] %n% datosDiferencia[, Id]])
mediaS[i] <- datosAjuste[, mean(Salario)]
varP[i-4] <- var(predichos)
varD[i-4] <- var(datosDiferencia[, Salario] -
                predichos[datosPrediccion[, Id] %n% datosDiferencia[, Id]])
varS[i] <- datosAjuste[, var(Salario)]
}

resumenAux <- data.table(t(c(mediaS, varS, mediaP, varP, mediaD, varD)))
names(resumenAux) <- c("mediaSu2", "mediaSO", "mediaTO",
                    "mediaSn", "mediaSa2", "mediaSa3", "mediaSu3",
                    "varSu2", "varSO", "varTO",
                    "varSn", "varSa2", "varSa3", "varSu3",
                    "mediaPSa2", "mediaPSa3", "mediaPSu3",
                    "varPSa2", "varPSa3", "varPSu3",
                    "mediaDSa2", "mediaDSa3", "mediaDSu3",
                    "varDSa2", "varDSa3", "varDSu3")
resumen <- rbind(resumen, resumenAux)

#### Cálculo t's ####
t13 <- N*(resumen[, mediaPSa3]+resumen[, mediaDSa3])
t23 <- N*(resumen[, mediaPSu3]+resumen[, mediaDSu3])
t3 <- N*resumen[, mediaSn]

estimadores <- data.table(t13, t23, t3)
names(estimadores) <- c("t13", "t23", "t3")

#### Cálculo varianzas muestrales ####
V3_Estandar <- resumen[, V3_Estandar]
V13 <- (((1-f1)/n1)*resumen[, varPSa3] + ((1-f13)/((a+b)*n3))*resumen[, varDSa3])*(N^2)
V23 <- (((1-f2)/(n2*(1-mu)))*resumen[, varPSu3] + ((1-f23)/(c*n3))*resumen[, varDSu3])*(N^2)
V3 <- ((f1 + f2)/((N-n1)^2 + N*n2*(1-mu)) + (1-f3)/(n3*(1-mu)))*(N^2)*resumen[, varSn]

#### Cálculo covarianzas muestrales ####
C13 <- -resumen[, varSa3]*N
C23 <- -N^2/(N-n1)*resumen[, varSu3]

#### Cálculo w's ####
tresOcasiones <- data.table()
nTresOcasiones <- c("t_tresOca", "t_tresOca_pob", "var_tresOca", "w1_est", "w2_est", "w3_est")
escala <- 1e18
varianza <- function(w) {
  matrizVar <- matrix(data = c(V13, C13, C13,
                              C13, V23, C23,
                              C13, C23, V3),
                      nrow = 3,
                      ncol = 3)
  w <- matrix(data = w,
             nrow = 1,
             ncol = length(w))
  varianza <- w %*%matrizVar %*%t(w)
  return(varianza/escala)
}

Aeq <- matrix(rep(1, 3), nrow = 1)
Beq <- 1

lb <- rep(0, 3)
ub <- rep(1, 3)

```

```
w0 <- c(0.3, 0.3, 0.4)

wOpt <- solnl(X = w0,
             objfun = varianza,
             lb = lb,
             ub = ub,
             Aeq = Aeq,
             Beq = Beq,
             tolX = 1e-6,
             tolFun = 1,
             tolCon = 1e-6,
             maxIter = 1000,
             maxnFun = 1e28)

wMin <- wOpt$par
varMin <- wOpt$fn
t_tresOca <- c(t13, t23, t3) %%wMin
t_tresOca_pob <- c(t13, t23, t3) %%wPob
tresOcasionesAux <- data.table(t(c(t_tresOca, t_tresOca_pob, escala * varMin, wMin)))
names(tresOcasionesAux) <- nTresOcasiones
tresOcasiones <- rbind(tresOcasiones, tresOcasionesAux)
```

Bibliografía

- Bell, P. (2001). Comparison of Alternative Labour Force Survey Estimators, *Survey Methodology* **27**(1): 53–63.
- Bernal, R., Cadena, X., Camacho, A., J.C., C., Fergusson, L., Ibáñez, A., Rodríguez, C., Peña, X. & ELCA, C. (2014). Encuesta Longitudinal Colombiana de la Universidad de los Andes - ELCA 2013 , *Serie Documentos Cede* **42**: 1–23.
- Carpenter, J. & Kenward, M. (2013). *Multiple Imputation and its Application*, second edn, John Wiley and Sons.
- CEDE & U.Andes, F. E. (2010). Encuesta longitudinal sobre la dinámica de los hogares colombianos. panel de hogares: Diseño de la muestra, <https://encuestalongitudinal.uniandes.edu.co>.
- CEDE & U.Andes, F. E. (2011). *Colombia en Movimiento. Un Análisis Descriptivo Basado en la Encuesta Longitudinal Colombiana de la Universidad de los Andes ELCA*, Ediciones Uniandes.
- CEDE & U.Andes, F. E. (2014). *Colombia en Movimiento 2010-2013. Los Cambios en la Vida de los Hogares a través de la Encuesta Longitudinal Colombiana de la Universidad de los Andes ELCA*, Ediciones Uniandes.
- CEDE & U.Andes, F. E. (2017). *Colombia en Movimiento 2010-2013-2016. Los Cambios en la Vida de los Hogares a través de la Encuesta Longitudinal Colombiana de la Universidad de los Andes ELCA*, Ediciones Uniandes.
- Cheng, Y., Huang, B. & Yu, Z. (2017). A note on iterative AK composite estimator for Current Population Survey, *Journal of nonparametric statistics* **29**(2): 381–390.
- DANE (2016). Ficha metodológica. encuesta longitudinal de protección social elps, <https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/encuesta-longitudinal-de-proteccion-social-elps>.
- Fuertes, N., Tibavisco, M., Galeano, M. & Castaño, L. (2016). Seis años realizando seguimiento a los hogares de la ELCA , *Boletín de Divulgación ELCA* (6): 1–6.
- Fuertes, N., Tibavisco, M., Galeano, M. & Castaño, L. (2017). Una primera mirada a la ELCA 2016, *Boletín de Divulgación ELCA* (8): 1–6.
- Liang, K. & Zeger, S. (1986). Longitudinal Analysis Using Generalized Linear Models, *Biometrika* **73**: 13–22.

-
- Marsden, J. & Tromba, A. (1991). *Cálculo Vectorial*, 3rd edn, Addison-Wesley Iberoamericana, S.A.
- Rubin, D. (1977). The design of a general and flexible system for handling non-response in sample surveys, Manuscript prepared for the U.S. Social Security Administration.
- Rubin, D. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse, *Proceedings of the Survey Research Methods Section of the American Statistical Association* pp. 20–34.
- Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons.
- Särndal, C., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*, second edn, Springer.
- Strang, G. (2007). *Computational Science and Engineering*, Wellesley-Cambridge Press.
- van Buuren, S. (2012). *Flexible Imputation of Missing Data*, Chapman & Hall/CRC. Interdisciplinary Statistics Series.
- van Buuren, S. (2017). *Package 'mice'*, R package version 2.46.
URL: <https://cran.r-project.org/web/packages/mice/mice.pdf>
- van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R, *Journal of Statistical Software* **45**(3): 1–67.