UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

TRATAMIENTO DE VALORES FALTANTES EN ENCUESTAS DE MUESTREO

TESIS

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS

PRESENTA

OMAR DE LA RIVA TORRES

DIRECTOR DE TESIS: DR. IGNACIO MÉNDEZ RAMÍREZ

MÉXICO, D.F. ENERO, 2010

Agradecimientos

Agradezco al Dr. Ignacio Méndez Rámirez por la dirección de este trabajo así como a la Dra. Guillermina Eslava Gómez, a la Dra. Silvia Ruiz-Velasco Acosta, a la Mtra. Leticia Gracia-Medrano Valdelamar y a la Mtra. Patricia Romero Mares por la revisión y comentarios de la tesis.

Contenido

Prefacio

En este trabajo se presenta una evaluación de algunas alternativas para el tratamiento de los datos faltantes en encuestas de muestreo en su diseño, durante el levantamiento de la información y al final, cuando los datos han sido procesados. La intención del documento es presentar la perspectiva en un análisis de información cuando se enfrenta a los datos faltantes u omisión de respuestas con un tamaño de muestra que permite eliminar, ignorar o emplear un método para predecir la información faltante pero se requiere tener control de sus consecuencias en los resultados finales. Durante el desarrollo del trabajo se indica de manera indistinta los datos faltantes como no respuesta, omisión de respuesta, datos omitidos, etcétera.

El estudio está dividido en cinco capítulos:

Capítulo 1 se describen las características y las causas de la no respuesta así como recomendaciones para evitarla o disminuirla. Se presenta también una clasificación de las omisiones de respuesta que permite emplear métodos para la predicción de la información faltante.

Capítulo 2 es una explicación del concepto de la no respuesta por unidad o total y métodos para su tratamiento tanto durante el levantamiento de datos como al concluir el procesamiento de la información.

Capítulo 3 en este apartado se expone el concepto de la no respuesta por elemento o no respuesta parcial y opciones para reducirla cuando se

está captando la información y métodos para estimar las respuestas faltantes al final del procesamiento de los datos.

Capítulo 4 se presentan los detalles de los procedimientos para la generación de las bases de datos empleadas en el estudio y un mecanismo de omisión para evaluar los métodos del tratamiento de la no respuesta.

Capítulo 5 muestra las evaluaciones de los métodos analizados en este trabajo para el tratamiento de los datos faltantes total y parcial que se utilizan cuando los información está procesada.

La manera de evaluar el desempeño fue a través de la media de los sesgos relativos de las estimaciones y la varianza relativa de los sesgos. Se empleó el programa S-PLUS para la compilación de los programas escritos en R para llevar a cabo los cálculos correspondientes. En los Apéndices se especifican resultados auxiliares en los que se basan algunos de los métodos y las rutinas de cómputo para la instrumentación de los métodos presentados para el tratamiento de la no respuesta.

CAPÍTULO 1

La no respuesta en las encuestas por muestreo

1.1. Introducción

La deficiencia para obtener una medición completa en una encuesta se conoce como la no respuesta (Groves, 1989). Después de realizar el diseño muestral y llevar a cabo la recolección de la información para una encuesta por muestreo, la posibilidad de que no se complete el tamaño de muestra n, previamente calculado y requerido, es casi segura. La no respuesta es el motivo por el cual no se alcanza el tamaño de muestra n, la cual puede ser ocasionada por la ausencia de un informante adecuado, la negación de contestar el cuestionario o la incapacidad para responder todas las preguntas.

Alguna de las razones por la cual no es posible ignorar los efectos de la no respuesta puede ser mostrada con el siguiente planteamiento: suponiendo un muestreo aleatorio simple (mas) con una población de tamaño N, un tamaño de muestra n y un parámetro de la población a estimar T, se obtiene que la varianza del estimador $\hat{t}$ es $Var_{mas}(\hat{t}) = N^2(1 - n/N)s^2/n$, donde s^2 es la varianza muestral. Entonces, a medida de que n se reduce del tamaño originalmente requerido la varianza del estimador $\hat{t}$ se incrementa. Este problema siempre se tiene contemplado en casi todos los esquemas de encuestas por muestreo (en el de cuotas no ocurre así). Por experiencias anteriores se busca un parámetro informativo del porcentaje de no respuesta que se podría presentar en la encuesta a realizar. Teniendo un conocimiento

previo del probable porcentaje de no respuesta, se incrementa el tamaño de muestra para cubrir la no respuesta, con este procedimiento se reduce la varianza del estimador t pero en el caso de una población homogénea, en la variable de interés, tanto de quienes responden como de quienes no responden el problema emergente es el sesgo de los estimadores. Para ilustrar la situación, se presenta un planteamiento hecho por Lehtonen y Pahkinen (1996). Supóngase que se tiene una población de tamaño N dividida en dos estratos, el A de tamaño N_A, donde todos sus elementos responderán a la encuesta y el estrato B de tamaño N_B en el cual ninguno de sus elementos responderán, con $N = N_A + N_B$. Para estimar el total T solamente se podrá obtener a partir de $\hat{t} = N \cdot \overline{y}_A$ en donde $\overline{y}_A$ corresponde a la media del estrato A. Ahora, si $E(\overline{y}_A) = \overline{Y}_A$ y las medias en cada estrato son distintas, es decir $\overline{Y}_A \neq \overline{Y}_B$, calculando el sesgo del estimador $\hat{t}$ se tiene que

$$\begin{aligned} Sesgo(\hat{t}) &= E(\hat{t}) - T = E(N\overline{y}_A) - (N_A\overline{Y}_A + N_B\overline{Y}_B) \\ &= N\overline{Y}_A - (N_A\overline{Y}_A + N_B\overline{Y}_B) \\ &= (N_A + N_B)\overline{Y}_A - (N_A\overline{Y}_A + N_B\overline{Y}_B) \\ &= N_B(\overline{Y}_A - \overline{Y}_B) \neq 0. \end{aligned} \tag{1.1}$$

Con lo que se observa que la no respuesta produce sesgo en los estimadores.

Este tipo de error, ajeno al muestreo, de acuerdo a Sande (1982) puede manejarse de cuatro maneras distintas:

1. Ignorar toda la información perdida, opción que puede producir el caso del ejemplo anterior, estimadores sesgados (a menos que se suponga que el comportamiento de los casos que no respondieron es igual a los que si respondieron).

2. Indicar los casos perdidos para que los usuarios de la información la traten de acuerdo a su elección.

3. Extraer una submuestra representativa de quienes no responden para posteriormente usarla para establecer inferencias acerca de quienes no contestaron.

4. Hacer imputaciones a los datos que se carecen con valores consistentes y realísticos.

Algunos métodos que se emplean para realizar los ajustes del apartado cuatro, serán el tema central de este estudio, los cuales se caracterizarán y evaluarán para conocer sus efectos en las estimaciones de la información de interés.

1.2. Características de la no respuesta

El ejemplo de la sección anterior sirvió para mostrar *la no respuesta por unidad o total*, no obstante, existe también *la ausencia de respuesta por elemento o parcial*. Para la definición de ambas se usará la notación empleada por Särndal et al. (1992): sea y_{ir} la i-ésima unidad de una muestra de r preguntas relacionadas con las respuestas al cuestionario esperadas. Cuando la totalidad de los r preguntas del y_i-ésimo cuestionario no son respondidas, o un determinado porcentaje de ellas (esto depende de las características mismas de la investigación) se denomina *unidad de no respuesta o no respuesta total* y se genera por tres causas principales: la persona que se desea entrevistar no puede ser contactada, rechaza responder, o es incapaz para responder todas las preguntas. Por otra parte, en algunas ocasiones existen preguntas en los cuestionarios que los entrevistados prefieren no responder porque trata de temas delicados como ingresos económicos, violencia familiar o uso de drogas.

Cuando en el y_i-ésimo cuestionario en la k-ésima de las r preguntas, $1 \leqslant k \leqslant r$, no se obtiene respuesta, se denomina que existe un *elemento de no respuesta o no respuesta parcial*. Estas definiciones son aplicables a encuestas con cualquier método de recolección de información; entrevistas personales, por correo, via telefónica y por internet.

Para medir la extensión de la unidad y elemento de no respuesta se utilizará la siguiente notación: sea s el conjunto que compone la muestra de cuestionarios que se desea obtener; el conjunto de los cuestionarios que tienen al menos una respuesta se denominará como r_u. Su complemento $s - r_u$ determinará el conjunto de la no respuesta total. El conjunto de cuestionario con todas las respuestas completas será r_c. Entonces, $r_u - r_c$, el conjunto de no respuesta por elemento esta compuesto por los cuestionarios que al menos tienen una respuesta pero no todas las que comprende el cuestionario. También se definen n_s, n_{r_u} y n_{r_c} como los tamaños de los conjuntos s, r_u y r_c respectivamente. La primera definición corresponde a la medición de la unidad de respuesta, $p_{r_u} = n_{r_u}/n_s$ y por consiguiente la medición de la no respuesta por unidad se obtiene de la siguiente expresión $1 - p_{r_u}$. Análogamente, la no respuesta por elemento se obtiene de la expresión $1 - p_{r_c}$. Cada clase de no respuesta tiene que ser abordada usando métodos distintos. Las posibles causas por las que surge la no respuesta se indican en la sección siguiente.

1.3. Factores probables para la generación de la no respuesta

Las tasas de respuesta son afectadas por la calidad del diseño de la encuesta, por esta razón, en esta fase inicial deben reducirse las posibles deficiencias que pudieran afectar en forma negativa la obtención de los datos requeridos. Los factores que repercuten en el resultado final en levantamiento de la información se presentan dos agrupaciones presentadas por Lohr (2000) y de Heer y Israëls (1992) :

En el diseño

a) *Población Objetivo.* Su definición debe ser clara para evitar ambigüedades, las cuales pueden causar exclusiones de segmentos de la población de interés.

b) *Reglas para la selección de la muestra o unidad de muestreo.* Situaciones totalmente distintas ocurren cuando tienen que ser entrevistados todos los miembros de un hogar o solamente uno de ellos.

c) *Objetivo de la encuesta.* Entre más interesado se halle el entrevistado en el objetivo, será mayor la disponibilidad para suministrar la información solicitada.

d) *Método de la encuesta.* En una encuesta transversal es más fácil obtener repuesta en comparación con una de panel en donde el desgaste de la muestra disminuye la tasa de respuesta.

e) *Técnicas de recolección de los datos.* Las encuestas por correo convencional o electrónico y telefónicas tienen una menor tasa de respuesta que las realizadas por entrevistas personales, sin embargo, éstas tienen un mayor costo.

f) *Temporada y horarios de recolección.* La elección de un horario en el cual no se pueda hallar al entrevistado (en horas laborables) o en una época inadecuada, por ejemplo periodos vacacionales o de migración de trabajadores agrícolas.

g) *Diseño del cuestionario.* Este aspecto es uno de los que posiblemente pueden afectar más la no respuesta por elemento, ya que la formulación incorrecta de una pregunta o su delicadeza tendrá un efecto en la tasa de no respuesta; más adelante se discutirá más ampliamente los factores influyentes en la no respuesta por elemento.

En la organización y los entrevistadores

a) *Características de los entrevistadores.* La edad o el género de los entrevistadores influyen en la posibilidad de obtener la información de los entrevistados.

b) *Selección y capacitación de los entrevistadores.* El tiempo empleado y los métodos para capacitar a los entrevistadores pueden afectar los resultados finales de las entrevistas.

c) *Control de calidad y evaluación.* Para los encuestadores es importante suministrarles información acerca de su desempeño para que ellos mismos vean sus deficiencias y puedan corregirlas.

d) *Experiencia.* Es obvio que un encuestador experimentado tendrá un mejor desempeño respecto a uno que no lo es, debido a que sabe cómo debe conducirse ante diversas situaciones en el levantamiento de datos.

e) *Carga de trabajo y duración de la entrevista.* Una carga de trabajo excesiva puede afectar el desempeño, y por consiguiente influirá en las tasas de respuesta de manera negativa.

f) *Salarios.* Una carga de trabajo pesada y un salario bajo jamás estimulará a los encuestadores para que generen resultados satisfactorios.

g) *Incentivos.* Los incentivos financieros o de otro tipo pueden incrementar las tasas de respuesta.

h) *Número de contactos.* Debe determinarse cual será el número óptimo de intentos para contactar a personas que no han sido encontradas.

i) *Revisitas.* Algunos entrevistadores experimentados pueden convencer a personas a participar en la encuesta, que originalmente se habían negado a hacerlo.

j) *Agobio de las personas que responden.* Quienes responden una encuesta hacen un favor inmenso por lo que la encuesta debe ser lo menos entrometida posible. Un cuestionario breve que requiere pocos detalles puede reducir el agobio, una preocupación fundamental para las encuesta de panel.

Con este listado de causas se intenta mostrar que la parte más importante en una encuesta se corresponde al diseño y levantamiento de los datos.

1.4. Supuestos

En muchas ocasiones se supone que las respuestas de quienes no respondieron no difieren significativamente de aquellas de quienes sí respondieron a las preguntas de la encuesta, es decir que la omisión de datos se generó de manera aleatoria. Pero antes de considerar cierto un supuesto tan importante es necesario definir los conceptos que describe Allison (2001):

1.4.1. Omisión completamente al azar

Supongase que existen datos faltantes en una variable Y de un conjunto de datos, se dice que sus omisiones son completamente al azar (en inglés *missing completely at random*, MCAR) si la probabilidad de la ocurrencia de un dato faltante no esta relacionado con el valor de Y ni con cualquiera otra variable del conjunto de datos. Cuando este supuesto se cumple para todas las variables, tanto el conjunto de casos con información completa como el conjunto con información incompleta pueden ser considerados como submuestras aleatoria de la muestra original.

1.4.2. Omisión al azar

Un supuesto menos fuerte es cuando los datos son omitidos al azar (en inglés *missing at random*, MAR). Supóngase que existen en un cuestionario las variables X y Y, donde X siempre se observa y Y algunas ocasiones es omitida. MAR significa que

$$P(Y_{omitida}|Y, X) = P(Y_{omitida}|X). \tag{1.2}$$

Es decir, la probabilidad condicional de una omisión en la variable Y, dadas las variables Y y X, es igual a la probabilidad de una omisión en la variable Y dada sólo la variable X. Esto significa que la omisión depende de los valores observados en una variable específica. Por ejemplo, el supuesto de MAR se cumple si la probabilidad de omisión en el ingreso monetario depende del estado civil pero en cada categoría del estado civil la probabilidad de omisión en el ingreso no esta relacionado con el mismo ingreso.

1.4.3. Ignorable

El mecanismo de omisión se dice que es ignorable si los datos son MAR y los parámetros que determinan la omisión no están relacionados con los

parámetros que se desea estimar. La característica de ignorabilidad indica que no hay necesidad de modelar el mecanismo de omisión de datos como parte del proceso de estimación. Algunos ejemplos donde se cumple el supuesto de ignorabilidad se presentan a continuación:

En un esquema de muestreo doble, donde se registran k variables para todos los casos en una muestra y l variables adicionales se registran para una submuestra de la muestra original. Si las k primeras variables suministran la información que determinará una submuestra, ya sea empleando un mecanismo de selección sistemática o determinística, las posibles respuestas para las l variables de los casos que no fueron seleccionados para la submuestra son MAR. En censos y encuestas a gran escala en los primeros intentos no se obtiene el número de entrevistas determinado por lo que se hace un seguimiento a una muestra aleatoria de no informantes. Los restantes casos de los que no se obtuvo información se considera ignorables.

En los dos ejemplos descritos en el párrafo anterior los datos faltantes fueron omitidos por diseño, debido a que no se buscaba tener la información de todas las variables para todos los casos; entonces cuando los datos se omiten por diseño pueden ser considerados como MAR porque se conocen las variables que determinaron su omisión.

1.4.4. No ignorable

Si los datos no son MAR, el mecanismo de omisión se denomina como no ignorable. Por ejemplo, en estudios observacionales donde los datos históricos, económicos o de otro tipo se recolectan para su análisis pero por razones ajenas a la información de las variables, no está disponible para algunos casos; la información acerca de los salarios tiende a omitirse en los niveles de ingreso más altos, es decir, el mecanismo de omisión está relacionado con la variable omitida. Cuando los datos se omitieron por causas en las que no se tiene control, no se tiene la certeza que se mantiene la propiedad MAR.

CAPÍTULO 2

Métodos para el tratamiento de la no respuesta por unidad o total

El método general para tratar la no respuesta por unidad es asignar un peso de ponderación a quienes sí respondieron para reducir el sesgo. En el muestreo probabilístico, los factores de ponderación corresponden al inverso de las probabilidades de inclusión o selección de una unidad de muestreo. Si la no respuesta se considera como otra etapa en el esquema de inclusión, entonces el inverso del producto de la probabilidad de selección por la probabilidad de observar una respuesta puede ser usada como factor de ponderación. Las probabilidades de selección son conocidas pero las probabilidades de respuesta son desconocidas y tienen que ser obtenidas de los datos recabados.

La separación de la población en un estrato de quienes responden y otro de los que no responden servirá para ajustar el sesgo por ausencia de respuesta haciendo que la observación de una respuesta o no en la unidad i sea una variable aleatoria. Se define la variable aleatoria

$$R_i = \begin{cases} 1, & \text{si la unidad } i \text{ responde;} \\ 0, & \text{en otro caso.} \end{cases}$$

La probabilidad de que la unidad seleccionada responda, $\phi_i = P(R_i = 1)$, se conoce como calificación de propensión. Con estas definiciones se analizará un método para calcular ϕ.

Los pesos de muestreo son los recíprocos de las probabilidades de selección, una estimación del total de la población es $\sum_{i \in S} w_i y_i$, en el muestreo aleatorio simple estratificado los pesos son $w_i = N_h/n_h$ si la unidad i está en el estrato h. Para esquemas de selección con probabilidades diferentes, $w_i = 1/\pi_i$ con $\pi_i = P(\text{la unidad } i \text{ está en la muestra})$. Se define la variable Z_i indicadora de la presencia en la muestra seleccionada, donde $P(Z_i) = \pi_i$. Si R_i es independiente de Z_i, entonces la probabilidad que la unidad i sea observada es

$$P(\text{seleccionar la unidad } i \text{ en la muestra y que responda}) = \phi_i \pi_i.$$

El peso final para la persona i responda es $1/\phi_i \pi_i$. Los métodos de ponderación suponen que se pueden estimar las probabilidades de respuesta a partir de variables conocidas para todas las unidades suponiendo datos MAR.

En las siguientes secciones se describirán algunos métodos para el análisis de la no respuesta por unidad presentados en los trabajos de Chapman (1976 y 1983), Cochran (1980), Ford (1976) y Lohr (2000).

2.1. Métodos usados durante el levantamiento de datos

En esta clase de métodos más que realizar imputación lo que se busca es que la ausencia de respuesta sea mitigada. En los siguientes apartados se explican dos de los métodos más conocidos de reducción de la no respuesta durante el levantamiento de datos.

2.1.1. Revisitas

Por razones distintas, el primer contacto con un informante potencial puede ser infructuoso, es por esto que en varias encuestas se realizan una o más visitas. En la Sección 2.2.2 se muestra un método que también puede utilizarse para hacer ajustes de la no respuesta mediante revisitas, el submuestreo de los no informantes en la primera visita.

Para el método de las revisitas también se dispone del modelo de Deming (Cochran, 1980) para la estimación de la media $\bar{y}$ y su varianza σ_y^2. La implementación de este método se basa en el conocimiento del costo promedio

de la revisitas. En primer lugar se debe establecer un número mínimo de revisitas antes de definir a un posible entrevistado como contacto imposible. La población se divide en r clases de acuerdo a la probabilidad de que se pueda establecer contacto con el entrevistado. Se definen:
w_{ij} = probabilidad de que un entrevistado de la clase j se encuentre en i visitas, $j = 1, \ldots, r$;
p_j = proporción de la población de la clase j;
μ_j = media poblacional de la clase j;
σ_j^2 = varianza poblacional de la clase j.

Se supone que $w_{ij} > 0$ para todas las clases (el modelo también puede se adaptado para casos imposibles de hallar). Si $\bar{y}_{ij}$ es la media de la clase j de quienes respondieron en i visitas o menos, también se tiene el supuesto que $E(\bar{y}_{ij}) = \mu_j$.

La media verdadera de la población se obtiene de

$$\bar{\mu} = \sum_{j=1}^{r} p_j \mu_j \tag{2.1}$$

Ahora se considerará la composición de la muestra después de i visitas. Los casos de la muestra se clasifican en $r+1$ clases. En cada una de las r clases de probabilidad de respuesta en i visitas se clasifican los casos, la clase $r+1$ la conforman todos aquellos que no respondieron en i visitas. Ignorando la corrección por finitud se tiene que el número n_{ij} de casos en las $r+1$ clases después de i visitas tiene una distribución multinomial

$$M\left(n_0, w_{i1}p_1, w_{i2}p_2, \ldots, w_{ir}p_r, (1 - \sum_{j=1}^{r} w_{ij}p_j)\right) \tag{2.2}$$

donde n_0 es el tamaño inicial de la muestra.

n_i es el número de total de entrevistados después de i visitas y se distribuye binomial $Bin(n_0, \sum_{j=1}^{r} w_{ij}p_j)$. Por lo que

$$E(n_i) = n_0 \sum_{j=1}^{r} w_{ij}p_j \text{ es el número esperado de entrevistas en } i \text{ visitas.} \tag{2.3}$$

Para n_i fijo, el número de entrevistas n_{ij} en la clase j obtenidas en cada una de las r clases se distribuye multinomial $M(n_i, w_{ij}p_j / \sum_{j=1}^{r} w_{ij}p_j)$, se tiene

entonces que

$$E(n_{ij}|n_i) = \frac{n_i w_{ij} p_j}{\sum_{j=1}^{r} w_{ij} p_j} \tag{2.4}$$

Entonces si $\bar{y}_i$ es la media de la muestra obtenida después de i visitas,

$$E(\bar{y}_i|n_i) = E\left(\frac{\sum_{j=1}^{r} n_{ij}\bar{y}_{ij}}{n_i}\right) = \frac{\sum_{j=1}^{r} n_i w_{ij} p_j \mu_j}{n_i \sum_{j=1}^{r} w_{ij} p_j} = \frac{\sum_{j=1}^{r} w_{ij} p_j \mu_j}{\sum_{j=1}^{r} w_{ij} p_j} = \bar{\mu}_i. \tag{2.5}$$

Debido a que el resultado no depende de n_i, la media de y_i es también $\bar{\mu}_i$, por lo que el sesgo de la estimación de $\bar{y}$ es $(\bar{\mu}_i - \bar{\mu})$.
La varianza condicional de $\bar{y}_i$ dado n_i es

$$Var(\bar{y}_i|n_i) = \frac{\sum_{j=1}^{r} w_{ij} p_j \left[\sigma_j^2 + (\mu_j - \bar{\mu}_i)^2\right]}{n_i \sum_{j=1}^{r} w_{ij} p_j} \tag{2.6}$$

La varianza de $\bar{y}_i$, ignorando los términos de orden $1/n_i^2$, se obtiene al reemplazar n_i en la Ecuación 2.6 por su valor esperado indicado en 2.3, es decir

$$Var(\bar{y}_i) = \frac{\sum_{j=1}^{r} w_{ij} p_j \left[\sigma_j^2 + (\mu_j - \bar{\mu}_i)^2\right]}{n_0 \left(\sum_{j=1}^{r} w_{ij} p_j\right)^2}. \tag{2.7}$$

El error cuadrático medio de $\bar{y}_i$ dadas i visitas es

$$ECM(\bar{y}_i|i) = Var(\bar{y}_i|i) + (\mu_j - \bar{\mu}_i)^2. \tag{2.8}$$

También es necesario considerar el costo de hacer i visitas. El número esperado de nuevas entrevistas obtenidas en la $k-$ésima visita es

$$\sum_{j=1}^{r} (w_{kj} - w_{k-1,j}) p_j.$$

Si el costo por una entrevista en la $k-$ésima visita es c_k, entonces el costo esperado de hacer i visitas es $n_0 C(i)$ donde

$$C(i) = c_1 \sum_{j=1}^{r} w_{1j} p_j + c_2 \sum_{j=1}^{r} (w_{2j} - w_{1j}) p_j + \cdots + c_i \sum_{j=1}^{r} (w_{ij} - w_{ij}) p_j \tag{2.9}$$

La importancia del método es que, conforme la información se acumula con relación a costos y sesgos relativos, se puede plantear una estrategia económica para cualquier encuesta.

2.1.2. Sustitución

Para encuestas donde es posible, las imputaciones se realizan seleccionando sustitutos de la población para tomar el lugar de quienes no se tienen sus respuestas. Para tales casos, se obtiene un sustituto con características similares al elemento sin respuesta. En general, dos tipos básicos de sustitución son usados (Chapman, 1983):

a) Selección de un sustituto aleatorio

b) Selección de un sustituto especialmente designado.

Con el procedimiento de sustitución aleatoria, un elemento de la población se selecciona de acuerdo a una probabilidad definida para reemplazar al elemento sin respuesta. Usualmente el sustituto se elige de un subgrupo de la población. En varios procedimientos de selección aleatoria los sustitutos potenciales se eligen previamente a la fase del levantamiento de la información, con esto se evita retrasos en la selección de los sustitutos cuando el levantamiento ha iniciado. Además, más de un sustituto se selecciona para cada elemento en la muestra inicial. Pero en algunas ocasiones no es posible contar con un sustituto para cada elemento en la muestra. Una alternativa para evitar este inconveniente es la construcción de grupos de sustitución. Por ejemplo, en el caso de la edad se indica que el sustituto puede elegirse del grupo quinquenal de edad de quién no respondió. Las clases de sustitución pueden ser compuestas de una o más variables que se consideran en el encuesta.

Una de las ventajas del procedimiento de sustitución es que la muestra siempre se mantendrá balanceada por cada clase de sustitución, lo que es útil para muestras autoponderadas ya que se conservarán las mismas probabilidades de selección. Sin embargo, en algunas ocasiones los esfuerzos para tener contacto con los elementos seleccionados podría no ser tan exhaustivos como cuando no se tiene la opción de la sustitución, además en algunos casos los sustitutos son tratados como si hubieran formado parte de la primera muestra y las tasas de no respuesta son subestimadas.
Una crítica al procedimiento de sustitución es que a menudo no es de ayuda en la reducción del sesgo por la no respuesta puesto que la información faltante se obtiene por respuestas que presumiblemente son similares a las que

ya se captaron en la muestra. Aunque en cualquier método de imputación la información se logra a partir de los informantes.

2.2. Métodos usados después del levantamiento de datos

Cuando el levantamiento de datos ha concluido existen métodos para el tratamiento de la no respuesta, a continuación se explican algunos de ellos.

2.2.1. Ponderación simple

El tipo más simple de ajuste de no respuesta es hacer una ajuste general de ponderación (Chapman, 1976). Este ajuste podría ser igual a la suma de los pesos iniciales. Supóngase un muestreo aleatorio simple donde el peso muestral es N/n, n_r es el número de quienes respondieron en la muestra, el peso de los informantes en la muestra sería n/n_r. Asumiendo que no hubiera otras ponderaciones el peso final será $(N/n)\cdot(n/n_r)$. Para estimar la media poblacional para datos ponderados se utilizará la expresión siguiente:

$$\hat{\overline{y}} = \frac{\sum_{i=1}^{n_r} w_i y_i}{\sum_{i=1}^{n_r} w_i} \tag{2.10}$$

donde:
n_r = Número de quienes respondieron o informantes.
$w_i = \frac{n}{n_r}$ Pesos final asignado el $i-$ésimo informante.
y_i = El valor de la variable de interés para el $i-$ésimo informante.

En este caso debido a que el peso de ponderación de todos los informantes es el mismo, la media estimada en la Ecuación 2.10 se reduce a una media sin ponderar de los n_r informantes

$$\hat{\overline{y}} = \frac{\sum_{i=1}^{n_r} w_i y_i}{\sum_{i=1}^{n_r} w_i} = \frac{\sum_{i=1}^{n_r} (\frac{n}{n_r}) y_i}{\sum_{i=1}^{n_r} (\frac{n}{n_r})} = \frac{\frac{n}{n_r}\sum_{i=1}^{n_r} y_i}{n} = \frac{\sum_{i=1}^{n_r} y_i}{n_r} = \overline{y}_r.$$

También se tiene que $E(\hat{\overline{y}}) = \overline{Y}_r$, la media de la variable para todas aquellas unidades de la población que si son seleccionadas responderán.

El sesgo de $\hat{\overline{y}}$ puede escribirse como

$$Sesgo(\hat{\overline{y}}) = E(\hat{\overline{y}}) - \overline{Y} = \overline{Y}_r - [T_r\overline{Y}_r + (1 - T_r)\overline{Y}_{nr}] \tag{2.11}$$

donde:

T_r = Tasa de respuesta de la población, es decir, la proporción de las N unidades de la población que responderán si son seleccionados para recabar su información.

$\hat{\overline{Y}}_{nr}$ = La media de la variable para todas aquellas unidades de la población que si son seleccionadas no responderán.

El sesgo de $\hat{\overline{y}}_r$ depende de la tasa de respuesta $1 - T_r$ y de la diferencia entre las medias de los informantes y quienes no responden $\overline{Y}_r - \overline{Y}_{nr}$. En la práctica no es posible calcular el sesgo de $\hat{\overline{y}}_r$.

2.2.2. Submuestreo de los no informantes

Éste método fue propuesto por Hansen y Hurwitz en 1946 (citado en Lohr, 2000) y también considerado como un método de muestreo doble para determinar la media o total de la población, la cual se divide en dos estratos, los que contestaron y los que no contestaron. Primero se selecciona una muestra aleatoria de tamaño n de una población con N unidades, de ésta n_r responden y n_{nr} no responden y $n = n_r + n_{nr}$ con n_r y n_{nr} variables aleatorias. Posteriormente, se elige una submuestra aleatoria del $100\nu\,\%$ de los n_{nr}, donde $0 < \nu \leqslant 1$ y ν no depende de n_r. Se hace el mayor esfuerzo para obtener respuesta de cada elemento de n_{nr}. Sea $\overline{y}_r$ el promedio muestral de quienes respondieron originalmente y $\overline{y}_{nr}$ el promedio de quienes respondieron en la submuestra que posteriormente se seleccionó y se estimó de los $m_{nr} = n_{nr} \cdot \nu$ casos. De esta forma las estimaciones de la media y el total de la población son:

$$\hat{\overline{y}} = \frac{n_r}{n}\overline{y}_r + \frac{n_{nr}}{n}\overline{y}_{nr} \tag{2.12}$$

y

$$\hat{t} = N\hat{\overline{y}} = \frac{N}{n}\sum_{i\in E_r} y_i + \frac{N}{n}\frac{1}{\nu}\sum_{j\in E_{nr}} y_j, \tag{2.13}$$

donde:

E_r ={Las unidades de la muestra que están en el estrato de quienes respondieron.}

E_{nr} ={Las unidades de la muestra que están en el estrato de quienes no respondieron pero sí lo hicieron en el submuestreo.}

Nótese que $\hat{t}$ es una suma ponderada de las unidades observadas; los pesos son N/n para quienes responden y $N/n\nu$ para quienes respondieron en el submuestreo. Debido a que sólo se selecciona una submuestra en el estrato de quienes no respondieron, cada unidad de la submuestra representa más unidades en la población de los que representa una unidad en el estrato de quienes no responden. La estimación de la varianza de $\hat{\bar{y}}$, ignorando el factor de finitud, queda definido como

$$\hat{V}(\hat{\bar{y}}) = \frac{n_r - 1}{n-1}\frac{s_r^2}{n} + \frac{n_{nr} - 1}{n-1}\frac{s_{nr}^2}{\nu n} + \frac{1}{n-1}\left[\frac{n_r}{n}(\bar{y}_r - \hat{\bar{y}})^2 + \frac{n_{nr}}{n}(\bar{y}_{nr} - \hat{\bar{y}})^2\right]. \quad (2.14)$$

Cuando se obtienen todas las respuestas de la submuestra, el muestreo en dos fases elimina el sesgo por la falta de respuesta; por otra lado también considera la ausencia de respuesta en la estimación de la varianza de acuerdo a la ecuación 2.14.

2.2.3. Ajuste de clases de ponderación

Chapman (1976) describe el método de la siguiente manera: supóngase que la población se divide en $p_1, p_2, \ldots, p_c$ clases de acuerdo a cierta variable auxiliar de la encuesta. Dicha variable auxiliar puede ser alguna empleada para estratificar la muestra u otra variable adicional que es posible obtener para toda la muestra. Sea también $T_{r_1}, T_{r_2}, \ldots, T_{r_c}$ las proporciones de las unidades en cada una de las c clases que responderían si son seleccionadas para la muestra.

Ahora, considérese el caso de una muestra de tamaño n de una población con N unidades con un esquema de muestreo aleatorio simple. Sean $n_1, n_2, \ldots, n_c$ el número de unidades de la muestra en cada clase , los n_i son variables aleatorias y $\sum_{i=1}^{c} n_i = n$, y $n_{1r}, n_{2r}, \ldots, n_{cr}$ es el número de casos observados en las c clases. La ponderación básica (el inverso de las probabilidades de selección) es N/n para cada unidad muestral. No obstante, el ajuste por no respuesta puede variar de clase en clase. Para cada caso en la $i-$ésima clase este ajuste es (n_i/n_{ir}) que es la suma de los pesos muestrales de todas las unidades muestrales que caen en la celda $i-$ésima dividido por las suma

de los pesos muestrales de todos los informantes que cae en la $i-$ésima celda.

La estimación $\hat{\overline{y}}_r$ de la media se calcula como

$$\hat{\overline{y}}_r = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n_{ir}}\left(\frac{N}{n}\right)\left(\frac{n_i}{n_{ir}}\right)y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_{ir}}\left(\frac{N}{n}\right)\left(\frac{n_i}{n_{ir}}\right)} = \frac{\sum_{i=1}^{c}\frac{n_i}{n_{ir}}\sum_{j=1}^{n_{ir}}y_{ij}}{\sum_{i=1}^{c}\frac{n_i}{n_{ir}}\cdot n_{ir}} = \sum_{i=1}^{c} p_i\overline{y}_{ir} \tag{2.15}$$

donde

$\overline{y}_{ir}$ = La media muestral entre los informantes en la $i-$ésima clase de ponderación.

p_i = La proporción de la muestra que cae en la $i-$ésima clase de ponderación.

La esperanza de $\hat{\overline{y}}_r$ es

$$E(\hat{\overline{y}}_r) = \sum_{i=1}^{c} p_i\overline{Y}_{ir} \tag{2.16}$$

donde

$\overline{Y}_{ir}$ = La media de la variable para todos aquellos en la población contenidos en la $i-$ésima clase de ponderación que responderían si fueran seleccionados en la muestra.

El sesgo de $\hat{\overline{y}}_r$ se calculó como

$$Sesgo(\hat{\overline{y}}_r) = \sum_{i=1}^{c} p_i(1-T_{r_i})(\overline{Y}_{ir} - \overline{Y}_{inr}) \tag{2.17}$$

donde

$\overline{Y}_{inr}$ = La media de la variable para todos aquellos en la población contenidos en la $i-$ésima clase de ponderación que no responderían si fueran seleccionados en la muestra.

Es útil comparar el sesgo de $\hat{\overline{y}}_r$ de la ecuación 2.17 con la de $\hat{\overline{y}}$ en la ecuación

2.11. Si para una de las c clases de ponderación la diferencia $(\overline{Y}_{ir} - \overline{Y}_{inr})$ iguala a la diferencia $(\overline{Y}_r - \overline{Y}_{nr})$, el sesgo de $\hat{\overline{y}}$ y $\hat{\overline{y}}_r$ son idénticos pero también el sesgo de $\hat{\overline{y}}_r$ iguala al de $\hat{\overline{y}}$ si todas las tasas de respuesta son iguales a la tasa de respuesta total T_r.

Sin embargo si la diferencia $(\overline{Y}_{ir} - \overline{Y}_{inr})$ tiende a ser menor, en valor absoluto, que la diferencia $(\overline{Y}_r - \overline{Y}_{nr})$ y las T_{r_i} tasas de respuestas varían de clase en clase, el sesgo por no respuesta se reduce por el uso de clases de ponderación. Por lo tanto, la aplicación exitosa de éste procedimiento requiere de la identificación de las características de la encuesta que definirán las clases de ponderación que variarán tanto en las tasas de respuesta como en las estimaciones de la encuesta; además las características deben estar disponibles tanto para quienes responden la encuesta así como para quienes no lo hacen.

2.2.4. Postestratificación

La postestratificación es similar al ajuste de clases de ponderación, excepto que los datos de la población se usan para ajustar los pesos (Lohr, 2000). Suponga que se extrae una muestra aleatoria simple que será agrupada en H post-estratos, por ejemplo, el género o el grupo de edad. La población tiene N_h unidades en el estrato posterior h, donde n_h se seleccionaron para la muestra y n_{hr} respondieron. El estimador postestratificado para la media de la población $\overline{Y}$ es

$$\overline{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_{hr};$$

el estimador de clases de ponderación para $\overline{Y}$ si las clases de ponderación son los estratos posteriores, es:

$$\overline{y}_{wc} = \sum_{h=1}^{H} \frac{n_h}{n} \overline{y}_{hr}.$$

Los dos estimadores tienen una forma similar. La única diferencia es que en la postestratificación se conocen los N_h, mientras que en los ajustes de clases de ponderación no se conocen y se estiman mediante $N\frac{n_h}{n}$.

Para el estimador estratificado posteriormente, a veces se utiliza la varianza

condicional dados los n_{hr}. Para una muestra aleatoria simple,

$$V(\overline{y}_{post}|n_{hr}, h = 1, \ldots, H) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_{hr}}{N_h}\right) \left(\frac{S_h^2}{n_{hr}}\right) \qquad (2.18)$$

La varianza de $\overline{y}_{post}$ es ligeramente mayor, con algunos términos del orden de $1/n_{hr}^2$ (Oh y Scheuren, 1983).

También es posible hacer una postestratificación utilizando pesos donde se utiliza el estimador de proporción dentro de cada subgrupo debido a las cifras reales de la población. Sea $x_{hi} = 1$ si la unidad i es alguien que responde posteriormente en el estrato h, y 0 en otro caso. Se define

$$w_i^* = \sum_{h=1}^{H} w_i x_{hi} \frac{N_h}{\sum\limits_{j \in S} w_j x_{hj}}$$

donde $w_i = \frac{N_h}{n_h}$ si la unidad i está en el estrato h. Usando estos pesos modificados,

$$\sum_{i \in S} w_i^* x_{hi} = N_h,$$

el estimador estratificado posterior del total de la población es

$$\hat{t}_{post} = \sum_{i \in S} w_i^* y_i.$$

En la postestratificación se puede ajustar la subcobertura y la ausencia de respuesta si las cifras de la población N_h incluyen en las encuestas a personas que no están en su marco de muestreo.

Los supuestos en la estratificación posterior son:

1) Dentro de cada estrato posterior, cada unidad seleccionada para estar en la muestra tiene la misma probabilidad de ser alguien que responda.

2) La respuesta o la falta de respuesta de una unidad es independiente del comportamiento de las demás unidades.

3) Las personas que responden en un estrato posterior son como las que si lo hacen y los datos son MCAR.

Para que se cumplan los supuestos, generalmente se utilizan muchos estratos posteriores, sin embargo, si hay pocas personas que respondan en algún estrato posterior se pueden obtener estimaciones inestables y problemas para aplicar el teorema central del límite. Una solución consiste en integrar a los estratos posteriores con pocas observaciones en otros que tengan medias similares en variables de referencia hasta reunir una cantidad de observaciones de cada estrato posterior.

2.2.5. Procedimiento basado en estimadores de razón

A continuación se describirán dos procedimientos para el tratamiento de la no respuesta presentados por Ford (1976); antes se darán las siguientes definiciones que se emplearán en los métodos basados en estimadores de razón y de regresión.

Sean

p = porcentaje de la población que responde;
q = porcentaje de la población que no responde;
μ = media poblacional;
μ_1 = media de la población que responde;
μ_2 = media de la población que no responde.

Se tiene que $\mu = p \cdot \mu_1 + q \cdot \mu_2$ y $D = \mu_1 - \mu_2$.

Entonces la diferencia relativa entre los datos que se observan y los datos no observados es $D' = \frac{\mu_1 - \mu_2}{\mu}$. El sesgo usando sólo los datos observados para estimar μ es $B = \mu_1 - \mu = q(\mu_1 - \mu_2) = q \cdot D$.

Como consecuencia, la relación entre B, q y D es lineal. De la misma forma la relación entre B' (sesgo relativo), q y D' también es lineal, $B' = \frac{\mu_1 - \mu}{\mu} = q \cdot \frac{\mu_1 - \mu_2}{\mu} = q \cdot D'$

Generalmente existe una variable auxiliar asociada a cada unidad muestral. Esta variable auxiliar puede ser la variable utilizada para estratificar a la población, una variable observada o cualquier otra variable adicional que puede ser obtenida para la muestra total. Debe existir una correlación razonable entre la variable primaria y la variable auxiliar. En el diseño de doble muestreo la primera muestra consiste de los datos observados y los datos faltantes, en la segunda muestra corresponde sólo a los datos observados. El

estimador de proporción y su varianza aproximada son

$$\overline{y}_{razón} = \frac{\bar{y}}{\bar{x}}\overline{x}' \tag{2.19}$$

$$Var(\overline{y}_{razón}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)(2R\rho S_y S_x - R^2 S_x^2) \tag{2.20}$$

donde

$\overline{x}'$ = media de la variable auxiliar en toda la muestra;
$\overline{x}$ = media de la variable auxiliar en la parte de muestra en la que se observaron datos;
$\overline{y}$ = media de la variable principal de la parte de la muestra en la que se observaron datos;
$\overline{X}$ = media de la variable auxiliar en toda la población;
$\overline{Y}$ = media de la variable primaria en toda la población;
$R = \overline{Y}/\overline{X}$;
S_x^2 = varianza de la variable auxiliar;
S_y^2 = varianza de la variable principal;
ρ = correlación entre x y y.
n' = tamaño de la muestra completa;
n = tamaño de quienes respondieron en la muestra;
N = tamaño de la población.

Aunque la estimación de razón es casi siempre un estimador sesgado, es fácil calcularlo aún para encuestas complejas. Usualmente S_y^2, S_x^2, ρ y R son desconocidos, sus estimaciones se pueden sustituir en las Ecuación 2.20.

A pesar de que la estimación de la varianza no es insesgada puede ser tomada como una aproximación aceptable. El estimador de razón asume dos supuestos: 1) la muestra inicial es una muestra aleatoria y 2) los datos faltantes conforman una submuestra aleatoria de la muestra inicial. Este segundo supuesto probablemente no se cumpla en la mayoría de las encuestas. Esencialmente el estimador de razón es un estimador de regresón lineal asumiendo el intercepto como 0. Si en la población analizada no se cumple el supuesto de un modelo lineal, entonces el estimador de razón (o cualquier de regresión) se convierte en un estimador sesgado. En el caso de un diseño muestral estratificado existe un estimador de razón combinado, tal estimador se utiliza cuando la proporción $R = \overline{Y}/\overline{X}$ es igual en todos los estratos.

2.2.6. Procedimiento basado en estimadores de regresión

De la misma manera que en el procedimiento de la razón, se debe contar con una variable auxiliar además de la variable principal. Los estimadores son

$$\overline{y}_{regr} = \overline{y} + b(\overline{x}' - \overline{x}) \tag{2.21}$$

$$Var(\overline{y}_{regr}) = \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} \tag{2.22}$$

La estimación de la varianza $Var(\overline{y}_{regr})$ se obtiene de la siguiente ecuación

$$\widehat{Var}(\overline{y}_{regr}) = \frac{s_{yx}^2}{n} + \frac{s_y^2 - s_{yx}^2}{n'} \tag{2.23}$$

$Var(\overline{y}_{regr})$ se ajusta por un factor de corrección por finitud de $1 - \frac{n}{N}$, por lo que se obtiene

$$\widehat{Var'}(\overline{y}_{regr}) = \left(1 - \frac{n}{N}\right)\left[\frac{s_{yx}^2}{n} + \frac{s_y^2 - s_{yx}^2}{n'}\right] \tag{2.24}$$

como una estimación de la varianza de $\overline{y}_{regr}$ en una población finita donde $\overline{x}', S_y^2, \rho^2, n', n$ y N son las mismas que en el procedimiento de estimadores de razón y

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \qquad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}; \qquad b = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2};$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}; \qquad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}; \quad \rho = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$s_{yx}^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2 - b^2 \sum_{i=1}^{n}(x_i - \overline{x})^2\right].$$

Bajo un esquema muestreo aleatorio simple.

CAPÍTULO 3

Métodos para el tratamiento de la no respuesta parcial o no respuesta por elemento

La no respuesta por elemento surge cuando el informante inicia el cuestionario asociado a la encuesta pero por distintas razones no lo concluye. En otros casos por tratarse de preguntas delicadas (nivel de ingresos, consumo de drogas, etcétera) no se obtiene la información solicitada. Posteriormente puede ocurrir que en el proceso de depuración de las bases de datos algunas respuestas son descartadas por su inconsistencia. Los métodos para el tratamiento de la no respuesta por elemento son conocidos como de imputación, y consisten en asignar valores a los elementos o respuestas faltantes usando información de una o más personas que sí respondieron y la información de otras variables con información completa. Los métodos de imputación no sólo se usan para eliminar el sesgo por no respuesta sino también para contar con un conjunto rectangular de datos completo.

3.1. Método usado durante el levantamiento de datos

El método siguiente se emplea para disminuir las omisiones en preguntas difíciles de responder durante la recolección de los datos.

3.1.1. Respuesta aleatorizada

El método de respuesta aleatorizada se utiliza para reducir la posible no respuesta ante temas sensibles durante el levantamiento de los datos. Warner (1965) fue el primero en proponer este método. Se supone que la mayor cooperación del informante se obtendrá utilizando un procedimiento que selecciona aleatoriamente una de dos preguntas a la que responderá confiablemente, sin indicar al entrevistador cuál de las dos preguntas respondió. Las dos preguntas son del tipo: "Tengo el atributo A (sensible)" y "No tengo el atributo A". Al entrevistado se le indica que sólo tiene que responder "si" o "no" a la pregunta que generó el procedimiento aleatorio. El entrevistador no sabe a qué pregunta está respondiendo el informante por lo que la confidencialidad se conserva (Emrich, 1986).

Asumiendo que cada entrevistado responde verazmente. Un estimador insesgado de la proporción de la población con el atributo sensible A se obtiene de la ecuación siguiente:

$$\theta = P\pi_A + (1-P)(1-\pi_A) \tag{3.1}$$

donde

θ = la proporción de la población que respondió "sí".
π_A = la proporción verdadera de la población con el atributo A.
P = es la probabilidad conocida que el procedimiento de selección elija la pregunta sensible ($P \neq \frac{1}{2}$).

De la Ecuación 3.1 se obtiene el estimador insesgado

$$\tilde{\pi}_A = \frac{\tilde{\theta} - (1-P)}{2P-1} \tag{3.2}$$

donde $\tilde{\theta}$ es la proporción de la población que respondió "sí" de una muestra aleatoria simple (con reemplazo) de tamaño n. La varianza de $\tilde{\pi}_A$ es

$$Var(\tilde{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{3.3}$$

El segundo término de la Ecuación 3.3 corresponde a la varianza adicional debida al mecanismo de aleatorización. Horvitz, Shah and Simmons (1967) sugirieron que la cooperación del entrevistado podría incrementarse si se utilizaban dos preguntas no relacionadas del siguiente tipo: 1) "Tengo el

atributo A" y 2) "Tengo el atributo (no sensible) B". Si π_B, la proporción de la población con el atributo B es conocida, la estimación de π_A es

$$\tilde{\pi}_A = \frac{\tilde{\theta} - (1-P)\pi_B}{P} \tag{3.4}$$

donde $\tilde{\theta}$ y P ya habían sido definidas.

Si π_B se desconoce, dos muestras independientes se requieren para estimar π_A y π_B.

En este caso sean:
θ_i = la proporción de la muestra que respondió "sí" en la muestra i;

P_i = la probabilidad (conocida) de que el mecanismo de aleatorización seleccione la pregunta sensible en la muestra i ($P_1 \neq P_2$);

n_i = el tamaño de la muestra i. Asumiendo que todas las respuestas son confiables, un estimador insesgado de π_A es

$$\pi_A = \frac{(1-P_2)\tilde{\theta}_1 - (1-P_1)\tilde{\theta}_2}{P_1 - P_2} \tag{3.5}$$

El incremento de las tasas de respuestas, las disminución de su sesgo en preguntas sensibles y además la protección de la confidencialidad del entrevistado y al entrevistador de situaciones legales son las principales ventajas de los métodos de respuesta aleatorizada. No obstante, también deben mencionarse que con el uso de estos métodos se incrementa el tiempo de la entrevista lo que podría incrementar la tasa de no respuesta en esta u otras preguntas del cuestionario.

3.2. Métodos usados después del levantamiento de datos

3.2.1. Análisis de casos completos

También conocido como exclusión por lista (Allison, 2001), el método consiste en borrar todos los casos que contengan un dato faltante en cualquier variable o respuesta y después realizar el análisis de los datos con los casos completos, gracias a su simpleza el método puede ser aplicado a cualquier conjunto de datos. Si el patrón de datos faltantes se considera

que es MCAR entonces la muestra reducida puede considerarse como una submuestra de la original. Como consecuencia, si para cualquier parámetro de interés la estimación sería insesgada para la muestra original (sin datos faltantes) entonces también lo será con la exclusión por lista. Además los errores estándares y las pruebas estadísticas serán consistentes con la estimación para la muestra completa. Es claro que los errores estándares serán mayores debido a que es menor la cantidad de información utilizada. En general, parece ser que la exclusión por lista no es robusta a la violación del supuesto de MCAR; sin embargo, es el método más robusto ante violaciones de MAR entre las variables independientes en un análisis de regresión, especialmente cuando las probabilidades de ocurrencia de datos faltantes no obedecen a los valores de la variable dependiente. Tal característica no sólo aplica a la regresión lineal sino también a la regresión logística, de Cox y de Poisson. Hay una importante salvedad acerca de las propiedades de la exclusión por lista en el análisis de regresión: se asume que los coeficientes de la regresión son los mismos para todos los casos en la muestra. Si los coeficientes varían en algunos subconjuntos de la muestra, pueden existir ponderaciones hacia un subconjunto u otro, pero pueden ser corregidas ajustando regresiones para cada subconjunto.

3.2.2. Análisis de datos disponibles

También conocido como exclusión por pareja, el análisis de datos disponibles es una alternativa que puede utilizarse para regresiones lineales, análisis de factores y modelos de ecuaciones estructurales (Allison, 2001). Se sabe que un modelo de regresión lineal puede ser estimado usando únicamente medias muestrales y matrices de covarianza o medias, desviaciones estándares y la matriz de correlaciones. La idea del análisis de casos completos es calcular cada una de estas estadísticas de resumen utilizando todos los casos disponibles. Por ejemplo, para calcular la covarianza entre las variables A y B, todos los casos que tengan presencia en las variable A y en la B serán utilizados. Si los datos son MCAR, el análisis de datos disponibles produce estimaciones de parámetros que son consistentes (y aproximadamente insesgados en muestras grandes). Por otro lado, si los datos son MAR, pero no observados aleatoriamente, las estimaciones pueden estar seriamente sesgadas.

Debido a que se utiliza una cantidad mayor de información en el análisis de datos disponibles, cuando los datos son MCAR las estimaciones de los parámetros tendrán menor variabilidad muestral (errores estándares meno-

res) que el análisis de datos completos. En estudios analíticos y de simulación de modelos lineales se ha observado que la exclusión por lista produce estimaciones más eficientes cuando las correlaciones entre las variables son bajas, mientras que el análisis de datos completos no funciona mejor cuando las correlaciones son altas. El principal problema con el análisis de datos disponibles es que para el cálculo de los errores estándares se involucra el tamaño de muestra, que se modifica entre las diferentes combinaciones de parejas de variables y por consiguiente genera estimaciones sesgadas. Un problema adicional que surge ocasionalmente con muestras pequeñas es que la matriz de correlaciones o de covarianzas puede no ser positiva definida y por lo tanto, no se pueden llevar a cabo los cálculos de las estimadores de regresión. Debido a estas dificultades el análisis de datos disponibles no puede ser recomendado como una alternativa al análisis de datos completos.

3.2.3. Imputación *hot-deck*

De manera general el procedimiento puede ser definido cuando un valor imputado se selecciona de una distribución estimada para cada dato faltante (Ford, 1976). En la mayoría de las aplicaciones la distribución empírica se obtiene de los casos con respuestas por lo que la imputación con el método *hot-deck* involucra la sustitución de valores individuales extraídos de casos similares con respuestas válidas.

Una breve explicación del procedimiento *hot-deck* es:

1) Separar en I clases basadas en k variables.

2) Si una variable está omitida en cierta clase, entonces, aleatoriamente se selecciona una variable reportada en la misma clase.

3) Se sustituye la variable seleccionada por el valor elegido.

4) Se llevan a cabo las estimaciones de los parámetros de interés en la muestra como si no hubiera datos faltantes.

La consecuencia del uso de este método es que las varianzas estimadas de la media muestral están sesgadas en comparación con su valor sin imputar. El paso 4 permite hacer uso del tamaño muestral que incluyen a los datos faltantes, de esta forma la pérdida muestral debida a estos valores no se reflejan en los valores muestrales. También debe notarse que los valores muestrales ya no serán independientes. El procedimiento *hot-deck* es esencialmente un proceso de duplicación con valores reportados que sustituyen

a valores faltantes. La covarianza se ignora con el procedimiento *hot-deck* lo cual puede ser un grave error.

El principal atractivo del procedimiento *hot-deck* es su simplicidad operacional pero la libertad en su método de clasificación ha impedido lograr una comparación teórica con otros métodos para el tratamiento de los datos faltantes.

El método *hot-deck* tiene algunas cualidades simples para recomendarlo. Por ejemplo, si $E(\overline{x} - \mu) = B$ es el sesgo asociado con la no respuesta cuando se estima la media poblacional μ con la media de una muestra aleatoria simple. Para estimar μ usando el procedimiento *hot-deck* se divide la muestra en I clases. Sean $E(\overline{x}_i - \mu_i) = B_i$ los sesgos en las clases i con $i = 1, \ldots, I$. Si p_i^* es la proporción de la población en la clase i entonces el sesgo B_{hd} asociado con la media estimada $\overline{x}_{hd}$ de los datos muestrales después de aplicar el procedimiento *hot-deck* es $B_{hd} = E(\overline{x}_{hd} - \mu) = \sum_{i=1}^{I} p_i^* B_i$. Para probar esta ecuación se tiene que

$$E[\overline{x}_{hd}] = E\left[\sum_{i=1}^{I} p_i \overline{x}_i\right] = E_{n_i}\left[E\left(\sum_{i=1}^{I} p_i \overline{x}_i | n_i\right)\right]$$

donde

n_i = Número de las unidades muestrales que caen en la clase i;
n = Tamaño de la muestra;
$p_i = \frac{n_i}{n}$;
$\overline{x}_i$ = Media muestral para la clase i.

El valor estimado dentro de las llaves esta basado en una n_i fija y el valor esperado fuera de las llaves es sobre todos los posibles valores de n_i. Entonces

$$E_{n_i}\left[E\left(\sum_{i=1}^{I} p_i \overline{x}_i | n_i\right)\right] = E\left[\sum_{i=1}^{I} p_i \mu_i\right] = \sum_{i=1}^{I} p_i^* \mu_i. \tag{3.6}$$

Nótese que $\overline{x}_i$ y n_i están correlacionados. Ahora, si $|B_i| < |B|$ para cada i, entonces:

$$|B_{hd}| = |\sum_{i=1}^{I} p_i^* B_i| < \sum_{i=1}^{I} p_i^* |B_i| < \sum_{i=1}^{I} p_i^* |B| = |B|$$

Por lo que se puede notar que el sesgo usando el procedimiento *hot-deck* es menor que el sesgo causado por la omisión de los datos bajo la condición de que $|B_i| < |B|$ para toda i. Tal condición se mantiene en la mayoría de los casos pero no está garantizada. Un buen método de clasificación debería disminuir el valor absoluto de los sesgos bajo $|B|$ en cada una de las I clases. No obstante, el procedimiento *hot-deck* permite cualquier clasificación; la bondad del proceso es dejado al criterio de quién empleara el método.

Una de las posibles alternativas a la sustitución aleatoria es la *sustitución a el vecino más cercano* reportada para el dato faltante. Con una variable auxiliar, el valor más cercano a un dato faltante es aquel que minimiza la diferencia absoluta entre el valor auxiliar del dato reportado y el dato faltante. En el caso de empates para el valor más cercano de la variable auxiliar se selecciona aleatoriamente.

Este procedimiento debería tener el mismo efecto cómo si se asignara la población a muchos estratos y seleccionando algunas unidades de cada estrato (puesto que la estratificación está basada en la variable auxiliar). Así, suposiciones de que el método *hot-deck* mejora con estratos definidos más estrechamente puede ser examinado con el resultado del procedimiento de el *vecino más cercano*.

Con un conjunto grande de valores completos para imputar el procedimiento es razonablemente robusto con correlaciones altas entre las variables primarias y las auxiliares.

Otra variación del procedimiento *hot-deck* es de *los dos vecinos más cercanos*. En lugar de sustituir el vecino más cercano reportado para cada variable faltante se sustituye la media de las dos más cercanos cuyo valor de la variable auxiliar es más pequeño que en la variable reportada y si el valor del vecino *más cercano* cuyo valor es mayor que la variable reportada.

Otra variación del método *hot-deck* que utiliza información de encuestas previas realizadas a la misma población (Chapman, 1976) o de datos históricos se denomina *cold-deck*, que consiste en clasificar las respuestas en una variable o en un conjunto de variables. El primer intento se hace definiendo categorías cruzadas o celdas de tal forma que las respuestas serán relativamente homogéneas dentro de la celda y heterogéneas entre las celdas. Debe haber al menos una respuesta en cada celda disponible para la imputación. El informante se asocia con la celda correspondiente a los valores de la res-

puesta. A continuación, se selecciona una respuesta de los valores disponibles en el *cold-deck* incluidos en la misma celda. Este valor generalmente se selecciona aleatoriamente o sistemáticamente. Cada respuesta faltante que se sustituye por un valor seleccionado se marca para diferenciarla de las respuestas observadas. Por ejemplo, se cuenta con información de dos encuestas que se aplicaron a una misma muestra, en la primera que se aplicó se tiene información completa de la escolaridad de los entrevistados agrupada por grupos quinquenales de edad, en la segunda encuesta se tienen omisiones en la información individual de escolaridad; los datos faltantes de esta variable son reemplazados de acuerdo a la distribución de los niveles de escolaridad correspondiente a la edad del entrevistado con omisión en la variable escolaridad de la segunda encuesta.

3.2.4. Imputación con la media

El método de sustitución por la media puede ser considerado como una variación del *hot-deck*. Es el método más barato pero en términos de estimación de la media verdadera equivale a ignorar los datos faltantes, es decir, $\bar{x} = \bar{x}_{obs}$ (Kalton y Kasprzyk, 1982). Esta afirmación se probará con el siguiente cálculo. Se tiene un conjunto de $n-k$ observaciones y k datos faltantes que se sustituyen con la media de los datos observados. Se estima la media de los n casos observados.

$$\overline{x}_{obs} = \frac{x_1 + \cdots + x_{n-k}}{n-k}$$

ahora se estimará media de los n datos

$$\overline{x} = \frac{x_1 + \cdots + x_{n-k} + k \cdot \overline{x}_{obs}}{n} = (x_1 + \cdots + x_{n-k}) \cdot \left[\frac{1}{n} + \frac{k}{n(n-k)}\right]$$

$$= (x_1 + \cdots + x_{n-k}) \cdot \left[\frac{1}{n-k}\right] = \overline{x}_{obs}.$$

En el caso de la varianza muestral se tiene que

$$S^2_{obs} = \frac{(x_1 - \overline{x}_{obs})^2 + \cdots + (x_{n-k} - \overline{x}_{obs})^2}{(n-k)-1}$$

para la varianza de los n casos

$$S^2 = \frac{1}{n-1}\left[(x_1 - \overline{x}_{obs})^2 + \cdots + (x_{n-k} - \overline{x}_{obs})^2 + k \cdot (\overline{x}_{obs} - \overline{x}_{obs})^2\right]$$

$$= \frac{1}{n-1}\left[(x_1-\overline{x})^2+\cdots+(x_{n-k}-\overline{x})^2)\right]\cdot\frac{n-k-1}{n-k-1}$$

$$= \frac{n-k-1}{n-1}S^2_{obs}.$$

Estas dos pruebas muestran que la sustitución de datos faltantes por la media $\bar{x}$ no cambia la estimación de la media de los datos observados y la varianza de los datos muestrales se modifica por un factor de (n casos observados+k casos faltantes-1)$/(n-1)$ que siempre es menor a uno por lo que la varianza estimada con sustitución de la media en los datos faltantes será menor a S^2_{obs}, lo cual es obvio ya que a medida que se sustituyen datos faltantes con un constante la varianza tiende a disminuir.

3.2.5. Ajuste por la variable ficticia

Este método se utiliza en presencia de datos faltantes en un análisis de regresión (Allison, 2001). Supóngase que hay datos faltantes en la variable X, que es una de la varias variables que componen el modelo de regresión. Se crea una variable ficticia D que es igual a 1 si hay una omisión en X y 0 en otro caso. También se crea una variable X^* tal que

$$X^* = \begin{cases} X, & \text{cuando no hay un dato faltante;} \\ c, & \text{cuando hay un dato faltante.} \end{cases}$$

donde c puede ser cualquier constante. Posteriormente se ajusta la regresión de la variable dependiente Y con la variables X^*, D y las restantes variables del modelo que se intenta obtener.

La ventaja aparente de este método es que utiliza toda la información que está disponible para la información faltante. La sustitución de datos omitidos por el valor c no se considera como una imputación debido a que el coeficiente X^* es invariante a la selección de c. De hecho, el único aspecto del modelo que depende de la elección de c es el coeficiente D, el indicador del dato faltante. Por facilidad de interpretación, una elección conveniente de c es la media de X de los casos sin respuestas omitidas. Entonces, el coeficiente de D puede ser interpretado como el valor predicho de Y para los casos con datos faltantes en X menos el valor predicho de Y con la media de los casos completos de X, controlando las otras variables en el modelo. El coeficiente de X^* puede ser visto como una estimación del efecto de X entre el subgrupo de aquellos casos que cuentan con información en X.

Este método también se ha propuesto para el caso de variables categóricas independientes en el análisis de regresión. El procedimiento consiste en la creación de un conjunto de variables ficticias, una variable por cada categoría excepto para la variable de referencia. La propuesta es simplemente crear una categoría adicional —y una variable ficticia adicional— para aquellos casos con datos faltantes en las variables categóricas.

3.2.6. Métodos de regresión

La imputación por regresión emplea las relaciones entre variables. Una aplicación de esta idea es emplear la información de los informantes para ajustar una regresión a una variable para la cual una o más imputaciones son necesarias empleando como variables, las que se asume tienen un alto valor de predicción. Las variables predictoras pueden ser las que forman parte del cuestionario o variables auxiliares(Särndal et al., 1992).

Para mostrar brevemente el procedimiento, se toma un conjunto de s variable, $y_1, \ldots, y_s$ con y_{jk} el valor de y_j para el $k-$ésimo elemento. Supóngase que para cierto elemento k no se cuenta con las respuestas y_{1k} y y_{2k}, el registro de sus datos podría ser representado de la forma siguiente $(-, -, y_{3k}, \ldots, y_{sk})$. La imputación de los dos datos faltantes se obtendría de la manera siguiente: sea $\tilde{y}_1 = \tilde{f}_1(y_3, \ldots, y_s)$ la ecuación de regresión de y_1 con $y_3, \ldots, y_s$ utilizando los elementos que contienen respuestas completas en las variables independientes. Sea $\tilde{y}_2 = \tilde{f}_1(y_3, \ldots, y_s)$ la estimación de la regresión de y_2 con $y_3, \ldots, y_s$. Estas dos ecuaciones y los $s-2$ valores registrados para el elemento k producen los datos imputados $\tilde{y}_{1k} = \tilde{f}_1(y_{3k}, \ldots, y_{sk})$ y $\tilde{y}_{2k} = \tilde{f}_2(y_{3k}, \ldots, y_{sk})$.

En algunas ocasiones un residual producido aleatoriamente se agrega para reflejar la incertidumbre en el valor imputado. El procedimiento puede ser empleado con regresiones lineales multivariadas, regresiones logísticas o multinomiales.

3.2.7. Imputación multiple

La imputación múltiple es un técnica que busca que las distribuciones de las variables y las relaciones entre ellas se mantengan y además se tenga en consideración la incertidumbre de los datos faltantes. Si se ignora la incertidumbre en la predicción se generarán errores estándares demasiado pequeños, valores de significancia artificialmente bajos y tasas de errores

Tipo I que serán mayores que los valores nominales.

El supuesto inicial del que parte la imputación múltiple es que en un conjunto de datos incompletos, los datos observados proporcionarán evidencia indirecta de los valores probables de los datos no observados. Dicha evidencia, cuando se combina con ciertos supuestos implica una distribución de probabilidad para los datos que serán considerados para los análisis estadísticos posteriores.

En la imputación múltiple (IM) cada dato faltante se reemplaza por un conjunto de $m > 1$ valores admisibles extraídos de su distribución predictiva. Las variaciones entre las m imputaciones del conjunto de datos refleja la incertidumbre con la cual los datos faltantes pueden ser estimados a partir de los datos observados. Después de hacer una IM hay m conjuntos completos de datos, en los cuales se puede llevar cualquier análisis estadístico pertinente. Después de hacer análisis con los m conjuntos, las estimaciones de los parámetros y errores estándares se combinan de acuerdo a las Reglas para la inferencia de imputaciones dadas por Rubin (1987), para producir estimaciones generales y errores estándares que reflejan la incertidumbre de la omisión de datos. La IM tiene como atractivos que las inferencias (valores de significancia, errores estándares) que se obtienen son generalmente válidas debido a que incorporan la incertidumbre ocasionada por los datos faltantes y además de que, entre tres a cinco imputaciones, se pueden obtener excelentes resultados. Rubin (1987) mostró que la eficiencia de una estimación basada en m imputaciones es aproximadamente

$$\left(1 + \frac{\delta}{m}\right)^{-1},$$

donde δ es la *fracción de información faltante* del parámetro que será estimado que mide la precisión de la estimación en el caso de que no hubiera habido omisiones. La eficiencia lograda por varios valores de m y proporciones de información omitida se muestra en la Tabla 3.1. En la mencionada tabla se observa que la ganancia en eficiencia disminuye rápidamente después de las primeras imputaciones. Por ejemplo, la columna de 30 % de información faltante ($\delta = 0.30$), una tasa alta para algunas aplicaciones, con $m = 5$ imputaciones se alcanza el 94 % de eficiencia, con un incremento a $m = 10$ se llega al 97 %, considerando que se duplica el número de imputaciones es una mejora bastante modesta. En la mayoría de la situaciones la mejora en la eficiencia es baja cuando se incrementa a un número grande de conjuntos y análisis de datos.

Tabla 3.1: Porcentaje de eficiencia de las estimaciones de IM por el número de imputaciones m y la fracción de información faltante.

	δ				
m	10 %	30 %	50 %	70 %	90 %
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

La afirmación de que un número pequeño de m imputaciones puede ser sorprendente debido a que en ejercicios de simulación son necesarias cientos o miles de iteraciones para lograr un nivel aceptable de precisión. En la IM un pequeño número m usualmente será suficiente. Esto se explica por dos razones: primera, la IM se basa en la simulación únicamente para resolver el aspecto de la información faltante. En cualquier método de simulación, el error de Monte Carlo se elimina seleccionando un valor muy grande de m, pero en la IM la ganancia en eficiencia no será importante debido a que el error de Monte Carlo es un pequeña proporción de la incertidumbre inferencial; la segunda razón es que la Reglas de Rubin para combinar los m conjuntos completos de datos explícitamente toman en cuenta el error de Monte Carlo. Un intervalo para la estimación basada en IM toma en cuenta el hecho que tanto los estimadores puntuales como la varianza contiene un error de simulación predecible debido a la finitud de m y la longitud del intervalo se ajusta para mantener una probabilidad de cobertura adecuada.

El uso de la IM había sido limitada desde los primeros años que fue propuesta por Rubin (1978) principalmente por la falta de herramientas computacionales. En condiciones triviales, las distribuciones de probabilidad de las que se obtendrán las IM tienden a ser demasiado complicadas. Esta situación cambió cuando se desarrollaron métodos de simulación conocidos conjuntamente como Cadenas de Markov Monte Carlo (MCMC por las siglas en inglés de *Markov chain Monte Carlo*).

De la misma forma que cualquier otro método estadístico, IM se basa en ciertos supuestos relacionados con tres aspectos:

a) Datos. Para generar imputaciones para los datos faltantes, se debe indicar un modelo de probabilidad para los datos completos (tanto los observados como los faltantes). La mayoría de los modelos asumen

que los datos pueden ajustarse a una distribución Normal univariada o multivariada, sin embargo, en la vida real en pocas ocasiones el supuesto se cumple. En la mayoría de las aplicaciones de la IM, el modelo usado para generar las imputaciones será en el mejor de los casos aproximado. Afortunadamente, la experiencia muestra que la IM tiende a comportarse aceptablemente. Por ejemplo, cuando se trabaja con variables categóricas (binarias u ordinales), a menudo se hacen imputaciones asumiendo normalidad y después se redondean los valores imputados continuos a la categoría más cercana. Las variables cuyas distribuciones son muy sesgadas pueden transformarse para que se parezcan a una normal y después de la imputación, regresarse a su escala original. No obstante, la elección del modelo de imputación no es un hecho trivial. Un modelo de imputación debe ser seleccionado para que al menos sea aproximadamente compatible con los análisis que se aplicarán a los datos imputados. El modelo debe ser suficiente para preservar las asociaciones o relaciones entre las variables que serán el centro de interés en investigaciones posteriores. Supóngase, que una variable Y se imputa asumiendo un modelo normal que incluye la variable X_1. Después de la imputación, se emplea una regresión lineal para predecir Y a partir de X_1 y otra variable X_2 que no se empleó en el modelo de imputación. El coeficiente estimado para X_2 tendería a cero debido a que Y se imputó sin considerar su posible relación con X_2. Como regla general, cualquier asociación que pueda ser importante para análisis posteriores debe tomarse en cuenta para el modelo de imputación. El recíproco de esta regla no es del todo necesaria. Si Y ha sido imputada con un modelo que incluye a X_2, no es necesario incluir a X_2 en análisis posteriores que involucren a Y, a menos que su relación con Y sea de interés sustancial. Los resultados relativos a Y no deben ser afectados por la inclusión de variables extras que no fueron consideradas en el proceso de imputación. En conclusión, un modelo de imputación que preserva un gran número de asociaciones es deseable debido a que puede ser empleado para una gran variedad de análisis posteriores a la imputación.

b) Distribución previa de los parámetros del modelo. La teoría estadística que subyace la IM involucra la ley fundamental de probabilidad conocida como el Teorema de Bayes (puede consultarse en el Apéndice A.1). La naturaleza bayesiana de la IM requiere la especificación de una distribución previa de los parámetros del modelo de imputación. En la Teoría Bayesiana, la distribución inicial cuantifica la creencia

o la cantidad de información que se tiene acerca de los parámetros del modelo antes que cualquier dato haya sido observado. Debido a que cada distribución inicial lleva a resultados previos, algunos métodos bayesianos han sido considerados subjetivos y poco científicos. Sin embargo, en la práctica, los resultados de un procedimiento bayesiano son más sensibles a la elección del modelo de los datos que a la elección de la distribución inicial. En muchos casos —especialmente cuando el tamaño de la muestra es moderadamente grande— cualquier distribución inicial razonable debería llevar a los mismos resultados

c) Mecanismo de la omisión de respuestas. En la Sección 1.4 se describen los supuestos que se asumen en la omisión de respuestas.

3.2.7.1. Herramientas para la generación de imputaciones multiples

Para crear imputaciones múltiples es necesario emplear algunas técnicas computacionales.

Método EM

El método EM (por el inglés *Expectation Maximization*) es una técnica para ajustar modelos a datos incompletos. EM aprovecha la relación entre los datos faltantes y los parámetros desconocidos del modelo de datos observados. Si se conocen los datos omitidos, entonces la estimación de los parámetros del modelo será directa. Entonces es posible obtener predicciones insesgadas para los datos faltantes. La interdependencia entre los parámetros del modelo y los datos faltantes sugiere un método iterativo donde primero se predicen los datos faltantes asumiendo parámetros iniciales, toma las predicciones para actualizar los parámetros y repite el proceso. La secuencia de parámetros converge a estimaciones de máxima verosimilitud que implícitamente promedian sobre la distribución de los datos faltantes.

El algoritmo EM formaliza una idea antigua para el tratamiento de los datos faltantes:

1.) Los datos faltantes se reemplazan por estimaciones.

2.) Estimar los parámetros.

3.) Volver a estimar los datos faltantes asumiendo que los nuevos parámetros son los correctos.

4.) Volver a estimar los parámetros hasta lograr la convergencia

Los métodos EM se emplean en modelos donde la log-verosimilitud $l(\theta|Y_{obs}, Y_{fal}) = \ln L(\theta|Y_{obs}, Y_{fal})$ es lineal en Y_{fal}, donde θ es el parámetro que se desea estimar, Y_{obs} y Y_{fal} indican los datos observados y los faltantes, respectivamente. En términos generales, para cada iteración es necesario estimar la log-verosimilitud $l(\theta|Y)$.

Cada iteración de EM consiste en una etapa E (esperanza) y otra etapa M (maximización). Una ventaja del algoritmo es que se puede confiar que convergerá, en el sentido de que bajo condiciones generales cada iteración incrementa la log-verosimilitud $l(\theta|Y_{obs})$, y si $l(\theta|Y_{obs})$ está acotada, la secuencia $l(\theta^{(t)}|Y_{obs})$ converge a un valor estacionario de $l(\theta|Y_{obs})$. De manera general, si la secuencia $\theta^{(t)}$ converge, entonces su convergencia la alcanza a un máximo local o punto silla de $l(\theta|Y_{obs})$. Una desventaja de EM es que converge muy lentamente cuando existe una gran cantidad de datos faltantes.

Little y Rubin (1987) describen el algoritmo EM de la siguiente manera: en el paso M simplemente se hace una estimación de máxima verosimilitud como si no hubiera datos faltantes, como consecuencia el paso M del método EM emplea los mismos métodos de máxima verosimilitud para estimar $l(\theta|Y)$. En el paso E se busca la esperanza condicional de los datos faltantes dados los datos observados y los parámetros estimados actuales, para después substituir los valores esperados por los datos faltantes. En este caso los datos faltantes no necesariamente son sustituidos por EM. La idea central del método EM, que se delinea como un proceso de sustitución de datos faltantes e iteraciones, es que los datos faltantes no son Y_{fal} sino la función de Y_{fal} que figuran en la log-verosimilitud de los datos completos, esto es, $l(\theta|Y)$. Específicamente, sea $\theta^{(t)}$ la estimación actual del parámetro θ. El paso E de EM busca el valor esperado de la log-verosimilitud suponiendo que θ fuera $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y) f(Y_{fal}|Y_{obs}, \theta = \theta^{(t)}) dY_{fal}.$$

El paso M del método EM determina $\theta^{(t+1)}$ maximizando el valor esperado de la log-verosimilitud:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geqslant Q(\theta|\theta^{(t)}), \quad \text{para toda } \theta$$

La tasa de convergencia está determinada por la proporción de la datos faltantes. Si no hay omisión de datos la convergencia es inmediata; si existe una

gran cantidad de información faltante relacionada con uno o más parámetros entonces la convergencia requerirá bastantes iteraciones. Una forma de revisar la convergencia de EM es examinar la función de log-verosimilitud y confirmar que se incrementa en cada iteración. También es útil correr con diferentes valores iniciales para asegurar que la log-verosimilitud tiene un máximo único. En algunas ocasiones el máximo único no existe cuando, por ejemplo, la verosimilitud tiene múltiples modas. Tales situaciones pueden ocurrir con muestras pequeñas, en muestras con altas tasas de omisión y con modelos que tienen demasiados parámetros en relación con los datos observados. Otras situaciones que afectan la convergencia ocurren cuando hay punto sillas, es decir, cuando en las derivadas direccionales de la función log-verosimilitud del parámetro de interés son cero pero no corresponde a un mínimo o máximo local; puede suceder también que el máximo valor de la función de verosimilitud se logra en varios valores del parámetros a estimar. Tal situación ocurre cuando una o más componentes de la función de log-verosimilitud son inestimables en el sentido de que no aparecen en la verosimilitud y por lo tanto la log-verosimilitud es la misma para cualquier valor de las componentes.

Aumento de datos

El aumento de datos (DA por el inglés *data augmentation*) es un algoritmo de cadenas de Markov generadas con métodos de Monte Carlo (MCMC), un método para encontrar distribuciones posteriores empleada en la estadística Bayesiana. El procedimiento iterativo es muy parecido al EM. Antes de iniciar DA, es necesario elegir un conjunto de variables para el proceso de imputación, que deben incluir a todas las variables con datos faltantes además de otras variables en el modelo que serán estimadas. También es importante incluir variables que estén altamente correlacionadas con las variables con datos faltantes o asociados con la probabilidad de que aquellas variables tengan omisiones.

Cuando las variables han sido seleccionadas, DA consiste en los siguientes pasos:

1.) Se eligen valores iniciales para los parámetros. Para el caso de un modelo normal multivariado los parámetros son las medias y la matriz de covarianza. Los valores iniciales pueden ser tomados de las fórmulas estándares usando los datos disponibles o los datos completos. Es preferible utilizar las estimaciones con el método EM.

2.) Se usan los valores actuales de medias y covarianzas para obtener estimaciones de coeficientes de regresión para las ecuaciones en las que cada variable con omisiones se ajusta sobre las variables observadas. El procedimiento se repite para cada patrón de omisiones de respuesta.

3.) Con el modelo de regresión se obtienen estimaciones que permitirán predecir valores para los datos faltantes. Cada valor obtenido por la predicción, se agrega un residual de la distribución normal a esa variable elegida de manera aleatoria.

4.) Con el conjunto de datos *completo*, que incluye las variables imputadas y observadas se recalculan las medias y covarianzas empleando las fórmulas estándares.

5.) Basado en el nuevo cálculo de las medias y covarianza, se hace una selección aleatoria de la distribución posterior de las medias y covarianzas.

6.) Con las medias y covarianza elegidas aleatoriamente, se regresa al paso 2 y se continua con el ciclo de los cinco pasos hasta que se logra la convergencia. Las imputaciones que se producen durante la iteración final se emplean para integrar la base de datos completa.

El paso 5 debe explicarse de una manera más detallada. Para obtener la distribución posterior de los parámetros, es necesario contar con una distribución *previa*. Aunque esto se puede lograr basándose en conocimientos previos acerca de los parámetros, la práctica usual es usar una distribución previa *no informativa*, es decir, una distribución inicial que contiene muy poca o ninguna información acerca de los parámetros. Por ejemplo, este procedimiento funcionaría de la siguiente manera: se tiene una muestra de tamaño n de mediciones de una variable Y que se distribuye normal. La media muestral es $\bar{y}$ y la varianza muestral es s^2. Se requieren calcular μ y σ para muestras aleatorias de distribución posterior. Con una distribución no informativa inicial[1] se puede obtener $\tilde{\sigma}^2$, un estimador para la varianza, con un muestra aleatoria de una distribución χ^2 con $n-1$ grados de libertad, tomando el recíproco del estimador y multiplicando el resultado por ns^2. El estimador de la media se puede obtener de una muestra de una distribución normal con media $\bar{y}$ y varianza $\tilde{\sigma}^2/n$.

[1]Para el procedimiento de aumento de datos, la distribución previa no informativa estándar (Schafer, 1997) se conoce como distribución previa de Jeffreys y se escribe como $|\Sigma|^{-(p+1)/2}$, donde Σ es la matriz de covarianza y p es el número de variables.

Si no hubiera datos faltantes, entonces las selecciones aleatorias corresponderían a la distribución posterior verdadera de los parámetros, pero si se ha imputado un dato faltante se tendría selecciones aleatorias de una distribución posterior suponiendo que los datos imputados fueran los datos verdaderos. De la misma manera, en el paso 3, lo que se tiene son selecciones aleatorias de la distribución posterior de los datos faltantes, dados los valores actuales de los parámetros. Sin embargo, debido a que los valores actuales no son los valores verdaderos, los datos imputados no pueden ser selecciones aleatorias de la distribución posterior verdadera, por esta razón el procedimiento debe ser iterativo. Al alternar entre las selecciones aleatorias de parámetros (condicionados tanto a los datos observados como a los imputados) y la selección aleatoria de los datos faltantes (condicionados a los parámetros actuales), se tiende a hacer selecciones aleatorias de una distribución posterior conjunta tanto de los datos como de los parámetros, condicionada únicamente a los datos observados.

La estimación iterativa por máxima verosimilitud, como en el algoritmo de EM, converge a un conjunto de valores y su convergencia puede ser verificada evaluando los cambios que ocurren entre una estimación y la anterior. En el caso de DA, el algoritmo converge a una distribución de probabilidad, no a un conjunto de valores. Ésta situación dificulta la determinación de que si se ha logrado la convergencia. Se han propuesto algunos diagnóstico estadísticos para evaluar la convergencia (Schafer, 1997), sin embargo, no se han considerado como definitivos. Por ejemplo, la convergencia se puede reinterpretar como la carencia de dependencia serial. Puede afirmarse que DA ha logrado la convergencia en k si cualquier parámetro en el ciclo t es estadísticamente independiente de su valor en el ciclo $t+k$ para $t = 1, 2, \ldots$. Se puede evaluar el grado de dependencia serial almacenando los parámetros en cada ciclo y graficando series de tiempo. El almacenaje de los valores de los parámetros permite calcular y graficar la función de autocorrelación (FAC) muestral de cualquier parámetro. La FAC es simplemente el coeficiente de correlación de Pearson con k retrasos para varios valores de k, es decir, la correlación entre los valores simulados de un parámetro en cualquier ciclo y su valor k ciclos después. El algoritmo DA se puede decir que ha logrado la convergencia en k ciclos si las FAC's muestrales para todos los parámetros se acercan a cero en el $k-$ésimo retraso.

Para determinar el número de iteraciones se deben tener en cuenta dos aspectos: Primero, entre más alta sea la proporción de datos faltantes serán

mayores las iteraciones necesarias para alcanzar la convergencia. Si en solamente el 5 % de los casos hubiera datos faltantes la convergencia se lograría con un número pequeño de iteraciones. Segundo, la tasa de convergencia del algoritmo EM es un indicador útil de la tasa de convergencia para DA. Una regla útil es que el número de iteraciones para el DA deberá ser al menos igual al de las iteraciones requeridas para EM. Por esta razón es mejor correr primero el algoritmo EM y después el DA (la principal razón es que EM suministra adecuados valores iniciales para DA).

3.2.7.2. Reglas para la inferencia de imputaciones múltiples

Cuando se han generado las imputaciones de los datos empleando algún método como el EM (Ver Sección 3.2.7.1), el conjunto de datos puede ser analizado con cualquier método que sería apropiado si los datos estuvieran completos. Por ejemplo, se pueden llevar a cabo regresiones lineales o logísticas en cualquier paquete de análisis estadístico. Cualquier modelo debería ajustarse m ocasiones, una por cada conjunto de datos imputados, y los resultados de estos conjuntos variarían como resultado de la incertidumbre de los datos faltantes. Para obtener un conjunto global de estimaciones de los coeficientes y los errores estándares, se debería almacenarlos y combinarlos de acuerdo a las reglas dadas por Rubin (1987):

Sea $\hat{Q}$ una estimación de una cantidad de interés de una población y U su varianza estimada. $\hat{Q}$ podría ser la estimación de un coeficiente de regresión y U su error estándar al cuadrado. Después de hacer el mismo análisis para cada conjunto de datos imputado, se tienen m posibles estimaciones $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m$ y sus varianzas correspondientes $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_m$. La estimación IM, o estimación global, está dada por

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i.$$

La varianza total de la estimación se conforma de dos componentes que toman en cuenta la variabilidad dentro y a través de cada conjunto de datos. La intra-varianza de la imputación,

$$\overline{U} = \frac{1}{m} \sum_{i=1}^{m} \hat{U}_i,$$

es simplemente el promedio de la varianza estimadas. La inter-varianza,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2,$$

es la varianza muestral de las mismas estimaciones. La varianza total, T, es la suma de las dos componentes con factor de corrección adicional para considerar el error de simulación en $\bar{Q}$,

$$T = \overline{U} + \left(1 + \frac{1}{m}\right) B.$$

La raíz cuadrada de T es el error estándar global asociado con $\bar{Q}$. Si no hubiera datos faltantes, entonces $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m$ serían idénticos, B sería 0 y T sería igual a $\overline{U}$. El tamaño de B con respecto a $\overline{U}$ es un reflejo de cuánta información está contenida en la parte faltante respecto a la parte observada.

Un intervalo de confianza del 95 % puede ser obtenido como $\bar{Q} \pm 2\sqrt{T}$. Sin embargo, es mejor calcular los intervalos usando la aproximación

$$\bar{Q} \pm t_{g.l.}\sqrt{T},$$

donde $t_{g.l.}$ indica un cuantil de una distribución t de Student con grados de libertad igual

$$g.l. = (m-1)\left(1 + \frac{m\overline{U}}{(m+1)B}\right)^2.$$

Los valores críticos (*p-values*) para probar la hipótesis nula $Q = 0$ pueden obtenerse comparando la proporción $\bar{Q}/\sqrt{T}$ con la misma distribución t.

Cuando m tiende a infinito la varianza global se reduce a la suma de las dos componentes de varianza y el intervalo de confianza se basa en una distribución normal con grados de libertad que tienden a infinito. Los grados de libertad están influenciados tanto por el número de imputaciones y la relación entre el tamaño de B y de $\bar{U}$. Cuando B es mayor a $\bar{U}$ los grados de libertad son cercanos al valor mínimo de $m-1$, pero cuando $\bar{U}$ es mayor a B los grados de libertad se aproximan a infinito. Si el valor calculado de $g.l.$ es muy pequeño —es decir, menor a 10— sugiere que la mayor eficiencia (estimaciones más exactas e intervalos más estrechos) se podrían lograr incrementando el número de imputaciones m. Si $g.l.$ es grande, entonces habrá poca ganancia si se incrementa m.

Rubin (1987) mostró que una estimación de la fracción de información faltante relativa a la cantidad poblacional Q es

$$\lambda = \frac{r + \frac{2}{(g.l. + 3)}}{r + 1},$$

donde

$$r = \frac{(1 + \frac{1}{m})B}{\overline{U}}$$

es el incremento relativo en la varianza debido a la no respuesta. Ambas cantidades sirven para diagnosticar y revelar la magnitud que podría influenciar la estimación de Q.

CAPÍTULO 4

Construcción y generación de bases de datos

Cuando se emplea cualquier método para el tratamiento de datos faltantes no se tiene una forma de verificar la precisión de las estimaciones de los parámetros. La única manera de verificar la consistencia de las estimaciones es a través del cálculo de intervalos de confianza y de las varianzas que implican las estimaciones con los métodos de imputación. Por tal razón, en este capítulo se describe el procedimiento que se siguió para simular dos bases de datos correspondientes a dos poblaciones generadas con distribuciones de probabilidad distintas. En cada población se simularon 1000 muestras para evaluar los métodos de tratamiento de no respuesta total y parcial. En las muestras se simuló el mecanismo de omisión MAR con dos tasas de no respuesta de 10 % y 30 %. En el Capítulo 5 se presentan los resultados de las evaluaciones. Este experimento completo de simulación es sólo un ejemplo de la enorme cantidad que se pueden obtener de la combinación de tamaños poblaciones; número y funciones de densidad de las variables; variables con información completa e incompleta; tamaño, precisión y confiabilidad de la muestra así como esquemas de muestreo y tasas y mecanismos generadores de la no respuesta, por lo que los resultados obtenidos pueden ser distintos con la modificación de alguna de las componentes que conformaron el experimento.

4.1. Variables componentes de las dos poblaciones

En las encuestas de muestreo los tipo de datos que se obtienen del cuestionario pueden ser nominales, ordinales, intervalares y escalares. Stevens (1946) describe las propiedades de las cuatro escalas de medición y el tratamiento estadístico que es aplicable. En el presente estudio se contó con dos poblaciones conformadas por variables ordinales y continuas. A continuación se explican las estructuras de las bases de datos que se emplearon en la evaluación de los métodos de tratamiento de la no respuesta.

Se decidió que las dos poblaciones las integrarían 10 variables con las características siguientes:

a) Seis variables discretas ordinales

b) Cuatro variables continuas

c) Las variables discretas ordinales formarían tres grupos (cada uno contenía dos variables) con dos, cuatro y 10 niveles de medición etiquetados como 0 y 1; $0, 1, 2$ y 3; y $0, 1, \ldots, 9$, respectivamente.

d) Las variables continuas formaron dos grupos (cada uno contenía dos variables), el primero con medias igual a cero y varianzas 1; en el otro grupo las medias igual a 50 y varianza 90.

La razón para que las variables ordinales tuvieron los niveles de medición indicados fue que se deseaba tener representadas a variables dicotómicas, politómicas y en el caso de las variables con 10 niveles de medición, un tipo de variables discretas que tienen la dualidad de poder ser consideradas como variables discretas o continuas. En el caso de las variables continuas se consideró tener variables con una distribución normal y normal estándar.

En la Tabla 4.1 se presenta el resumen de las variables y las etiquetas asignadas para identificar las características de las variables; de tal forma que la variable $vodic_1$ indica que se trata de una ordinal dicotómica, $vo4c_1$ una variable ordinal con cuatro categorías; $vo10c_1$ una ordinal con 10 categorías; $vnstd_1$ una variable normal estándar y finalmente vn_1 una variable normal.

Tabla 4.1: Características de las variables simuladas

No. de variables	Etiquetas	Ordinales	Continuas	
		Niveles de medición	Media	Varianza
2	$vodic_1$, $vodic_2$	2	–	–
2	$vo4c_1$, $vo4c_2$	4	–	–
2	$vo10c_1$, $vo10c_2$	10	–	–
2	$vnstd_1$, $vnstd_2$	–	0	1
2	vn_1, vn_2	–	50	90

4.2. Generación de las poblaciones *Pobl* y *PoblM*

La primera población *Pobl* se simuló por medio de una distribución normal multivariada con las variables v_i con medias $\mu_i = 0$ y varianzas $\sigma_i = 1$ para $i = 1, \ldots, 8$ y $\mu_i = 50$ y $\sigma_i = 4.49$ para $i = 9, 10$, la matriz de correlaciones aleatorias que se empleó para generar las variables con distribución normal multivariada con el siguiente procedimiento: se generó la matriz A con $a_{ij} \sim Norm(0, 1)$, $i, j = 1, \ldots, 10$, después se hizo la multiplicación $A^tA = B$, donde A^t indica la transposición de la matriz A. Se calculó la matriz diagonal $C = 1/\sqrt{diag(B)}$ y la matriz de correlaciones aleatorias D se obtuvo del producto de las matrices $D = CBC$. El paso siguiente fue categorizar las variables $v_i, i = 1, \ldots, 6$; de esta forma se aseguró que las variables ordinales tendrían una distribución subyacente normal multivariada. El procedimiento consistió en generar $m_i - 1$ números aleatorios con distribución normal estándar, ordenados de manera ascendente, donde $m_1 = m_2 = 2$, $m_3 = m_4 = 4$ y $m_5 = m_6 = 10$ fueron los niveles de medición para categorizar a v_i, $i = 1, \ldots, 6$. Por ejemplo, la variable v_4 se transformaría en $vo4c_2$ con cuatro niveles de medición, de acuerdo a tres umbrales aleatorios con distribución normal, es decir, se utilizó el proceso inverso descrito en el Anexo A.2. La matriz de correlaciones final de *Pobl*, después de las categorización de la variables indicadas, se presenta en la Tabla 4.2.

La segunda población *PoblM* se creó con la generación por separado de cada una de las v_i variables, $i = 1, \ldots, 6$ de acuerdo a una distribución aleatoria en los niveles de medición correspondientes. Para obtener las probabilidades de seleccionar el m_i-ésimo nivel, dependiendo si la variable era dicotómica,

de cuatro o diez niveles de medición de v_i, se empleó la ecuación:

$$P(F_{im-1} \leqslant U \leqslant F_{im}) = F_{im} - F_{im-1} = \tau_{im}. \tag{4.1}$$

donde τ_{im} correspondíó a la probabilidad de elegir el nivel m en la variable v_i, F_{im}, $i = 1, \ldots, 6$ indicaría la distribución acumulativa hasta τ_{im}. La asignación en los niveles de medición fue aleatoria con reemplazo hasta obtener la población total $PoblM$ con las v_i variables, $i = 1, \ldots, 6$, que fue equivalente a simular una distribución multinomial para cada variable. Con el procedimiento descrito se generaron los patrones de respuestas correspondientes a la variables ordinales. Las variables continuas v_i, $i = 7, \ldots, 10$ se generaron con una distribución normal multivariada de acuerdo con las especificaciones de las medias y varianzas de las Tabla 4.1 y correlaciones generadas aleatoriamente con el mismo procedimiento descrito para la población $Pobl$. La matriz de correlaciones de las variables de $PoblM$ se calcularon después de generarlas. En la Tabla 4.3 se presenta la matriz de correlaciones de $PoblM$. Resalta que las correlaciones fueron casi nulas entre las variables ordinales y, entre las ordinales y las continuas. En resumen, $Pobl$ se generaron con una distribución normal multivariada con las primeras seis variables categorizadas y las últimas cuatro continuas. $PoblM$ se conformó con la unión de seis variables cada una generadas con una distribución multinomial y cuatro variables normales multivariadas.

4.3. Determinación de tamaños de poblaciones y muestra

Para la determinación de los tamaños de muestra se asumió un esquema de muestreo aleatorio simple para la estimación de proporciones. El cálculo

Tabla 4.2: Matriz de correlaciones de la población $Pobl$.

	$vodic_1$	$vodic_2$	$vo4c_1$	$vo4c_2$	$vo10c_1$	$vo10c_2$	$vnstd_1$	$vnstd_2$	vn_1	vn_2
$vodic_1$	1.00									
$vodic_2$	-0.02	1.00								
$vo4c_1$	0.43	-0.22	1.00							
$vo4c_2$	-0.10	-0.48	0.26	1.00						
$vo10c_1$	0.13	0.22	-0.24	-0.51	1.00					
$vo10c_2$	0.06	-0.33	0.15	0.22	-0.01	1.00				
$vnstd_1$	0.03	0.10	-0.06	-0.50	0.40	0.35	1.00			
$vnstd_2$	0.01	0.30	-0.38	-0.17	0.28	-0.50	-0.39	1.00		
vn_1	0.16	0.21	0.14	-0.12	-0.31	-0.29	-0.12	-0.25	1.00	
vn_2	0.18	-0.18	0.20	0.22	-0.29	0.33	-0.28	-0.11	-0.23	1.00

Tabla 4.3: Matriz de correlaciones de la población *PoblM*

	$vodic_1$	$vodic_2$	$vo4c_1$	$vo4c_2$	$vo10c_1$	$vo10c_2$	$vnstd_1$	$vnstd_2$	vn_1	vn_2
$vodic_1$	1.00									
$vodic_2$	0.01	1.00								
$vo4c_1$	0.01	0.01	1.00							
$vo4c_2$	-0.01	0.01	-0.01	1.00						
$vo10c_1$	0.01	0.01	0.01	0.02	1.00					
$vo10c_2$	0.01	0.00	0.01	0.01	0.00	1.00				
$vnstd_1$	0.01	0.00	-0.01	0.01	0.00	-0.01	1.00			
$vnstd_2$	0.01	0.00	-0.02	0.00	0.00	-0.01	0.55	1.00		
vn_1	0.00	0.00	0.01	0.01	0.00	0.00	0.42	-0.51	1.00	
vn_2	-0.01	0.00	0.00	0.00	0.00	0.01	-0.69	-0.06	-0.66	1.00

del tamaño de muestra se determinó de acuerdo a las ecuación

$$n_0 = \frac{z^2_{\alpha/2} p(1-p)}{\delta^2} \tag{4.2}$$

donde:
n_0 =tamaño de muestra inicial.
$1-\alpha$ =nivel de confianza de la muestra.
$z^2_{\alpha/2}$ =valor de una distribución normal estándar que deja una área de $\alpha/2$ a su derecha.
p =proporción de unidades en la muestra con una característica de interés.
δ =error relativo de estimación de la proporción.

Con una corrección de finitud se tiene que el tamaño de muestra n se obtiene de

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \tag{4.3}$$

Cuando $p = 0.5$ la varianza de p se maximiza y también el tamaño de n; con este valor de p y la variación de los componentes de las Ecuaciones 4.2 y 4.3 se generó la Tabla 4.4. Se decidió que para mantener un error de estimación entre $\pm 1\,\%$ y $\pm 5\,\%$ con una confianza de $99\,\%$ el tamaño de muestra fue de $n = 660$ para el tamaño $N = 10000$ de las poblaciones *Pobl* y *PoblM*.

Los valores de las variables que se presentan en las Tablas 4.5 fueron los que se estimaron empleando los métodos de tratamiento de la no respuesta total y no respuesta parcial.

Tabla 4.4: Determinación de los tamaños de las poblaciones y muestra

Conf	$1-\alpha=0.90$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Err. est.	10 %	5 %	1 %	10 %	5 %	1 %	10 %	5 %	1 %
N (Pobl.)	muestra requerida n								
100	41	74	99	49	80	99	63	87	100
1000	64	213	872	88	278	906	143	399	944
10000	68	264	4035	96	370	4899	164	623	6239
100000	68	270	6336	96	383	8763	166	660	14228
1000000	68	271	6719	97	384	9513	166	664	16317
10000000	68	271	6760	97	385	9595	166	664	16560
100000000	68	271	6764	97	385	9603	166	664	16585

4.4. Simulación del mecanismo de no respuesta total y no respuesta parcial

No existen fundamentos teóricos o empíricos para determinar una tasa mínima de respuesta de una encuesta, generalmente depende del conocimiento que se tiene de las características propias de la población a muestrear. Por esta razón, tampoco se puede determinar de manera definitiva cuándo es posible aplicar métodos de tratamiento de la no respuesta dada una perdida de información por esta causa. Para el caso de este estudio se seleccionaron tasas de no respuesta del 10 % y 30 %, considerando que estas tasas mantienen la idea que una proporción mayor de informantes permitirá estimar la información perdida de un grupo de menor magnitud de quienes no respondieron en una encuesta.

En el Capítulo 1 se presentaron las características de los mecanismos de omisión MCAR y MAR, para este análisis no se consideró el mecanismo MCAR debido a que en la práctica pocas veces se verifica y es equivalente a eliminar una submuestra aleatoria simple de la muestra original. El mecanismo de omisión que se emplearía fue el MAR de acuerdo a la función de probabilidad:

$$P_i(\text{el caso } i \text{ está omitido}) = \frac{\exp(vn'_{2,i} - \beta_{nr})}{1 + \exp(vn'_{2,i} - \beta_{nr})} \tag{4.4}$$

donde:
$vn'_{2,i}$ =es el valor estandarizado de la variable vn_2 para el $i-$ésimo caso, $i = 1, \ldots, 660$.
β_{nr} =corresponde al valor para generar la tasa de no respuesta

Tabla 4.5: Valores de las proporciones de las variables ordinales en los niveles de medición y medias de la variables continuas de las poblaciones $Pobl$ y $PoblM$, N =10 000

Variable	Niveles de medición (nm)	$Pobl$	$PoblM$
vod_1	0	0.77	0.74
	1	0.23	0.26
vod_2	0	0.62	0.78
	1	0.38	0.22
$vo4c_1$	0	0.72	0.46
	1	0.10	0.15
	2	0.06	0.05
	3	0.12	0.33
$vo4c_2$	0	0.34	0.37
	1	0.05	0.10
	2	0.08	0.35
	3	0.53	0.17
$vo10c_1$	0	0.09	0.06
	1	0.11	0.16
	2	0.01	0.19
	3	0.04	0.11
	4	0.04	0.14
	5	0.01	0.03
	6	0.06	0.03
	7	0.04	0.04
	8	0.23	0.08
	9	0.36	0.17
$vo10c_2$	0	0.16	0.13
	1	0.15	0.04
	2	0.01	0.10
	3	0.21	0.13
	4	0.03	0.13
	5	0.02	0.11
	6	0.09	0.10
	7	0.08	0.12
	8	0.13	0.04
	9	0.11	0.10
$vnstd_1$	$\bar{y}$	-0.01	0.01
$vnstd_2$	$\bar{y}$	0.03	0.01
vn_1	$\bar{y}$	50.05	50.06
vn_2	$\bar{y}$	50.08	49.92

$nr = \{\{0.10\}, \{0.30\}\}$.
exp() =la función exponencial.

La función de probabilidad 4.4 indica que a medida que se incrementa el valor de vn_2 (esta variable se eligió como auxiliar y por lo tanto estaría disponible para todos los casos), la probabilidad de omisión aumenta. Los valores de β_{nr} se obtuvieron de la relación $\ln(p/(1-p))$, $p = \{0.90, 0.70\}$ que corresponden a las tasas de respuesta para omisiones del 10 % y 30 % (Ver Gráfica 4.1).

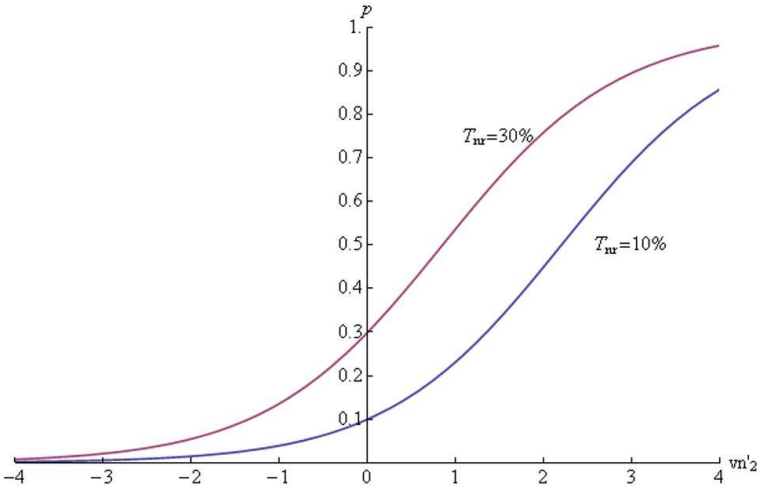

Figura 4.1: Funciones de probabilidad para generar tasas de no respuesta total y parcial del 10 % y 30 %

Para generar la no respuesta total, el caso i, $i = 1, \ldots, 660$ se eliminaron dada la regla:

$$P_i(\text{el caso } i \text{ está omitido}) = \frac{\exp(vn'_{2,i} - \beta_{nr})}{1 + \exp(vn'_{2,i} - \beta_{nr})} > r_{\text{unif}_i} \qquad (4.5)$$

donde r_{unif} es un número aleatorio con distribución uniforme en (0,1).

La generación de la no respuesta parcial también utilizaría la Regla 4.5, pero el caso seleccionado tenía alguno de los $2^9 - 1 = 511$ patrones de omisión, con selección aleatoria con reemplazo.

El mecanismo de omisión MAR se aplicó a las muestras extraídas de las dos poblaciones *Pobl* y *PoblM* para obtener la no respuesta total y no respuesta parcial para las tasas de no respuesta del 10 % y 30 %. Ejemplos de

la no respuesta total y parcial en las muestras simuladas se pueden observar en las Tablas 4.6 y 4.7.

Tabla 4.6: Ejemplos de patrones de no respuesta total

$vodic_1$	$vodic_2$	$vo4c_1$	$vo4c_2$	$vo10c_1$	$vo10c_2$	$vnstd_1$	$vnstd_2$	vn_1	vn_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
–	–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–	–
1	2	1	4	10	6	1.79	1.75	52.26	53.86
2	2	1	3	3	9	0.25	0.41	51.76	48.86
–	–	–	–	–	–	–	–	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	2	3	2	4	10	1.50	-1.31	65.66	47.31
1	1	2	1	10	7	1.46	0.271	52.57	40.08
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabla 4.7: Ejemplos de patrones de no respuesta parcial

$vodic_1$	$vodic_2$	$vo4c_1$	$vo4c_2$	$vo10c_1$	$vo10c_2$	$vnstd_1$	$vnstd_2$	vn_1	vn_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
–	1	1	–	10	10	1.40	–	–	58.49
–	1	3	–	–	10	0.90	–	–	70.46
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
–	1	–	3	10	2	–	–	46.49	43.69
–	2	1	1	10	2	0.70	–	44.87	49.80
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
–	–	–	–	–	1	–	–	39.78	51.97
1	1	–	4	–	4	–	–	–	42.64
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

CAPÍTULO 5

Evaluación de los métodos del análisis de datos faltantes

En este capítulo se instrumentan los métodos para el tratamiento de la no respuesta total y parcial, que se pueden aplicar después de la recolección de los datos y están disponibles en una base de datos, excepto por el submuestreo de los no informantes que es un método aplicado en el levantamiento de datos. Los valores verdaderos de proporciones y medias son los que se presentaron en la Tabla 4.5 del Capítulo 4 y los que se tomaron como referencia en éste capítulo. Se simularon 1000 muestras aleatorias de las dos poblaciones $Pobl$ y $PoblM$ para evaluar cada método. El desempeño de los métodos primero se midió calculando las medias de los sesgos $\bar{B}_k$ y varianzas de los sesgos $Var(B_k)$ con las ecuaciones siguientes:

$$\bar{B}_k = \left[\frac{1}{1000} \sum_{j=1}^{1000} B_{kj} = \frac{1}{1000} \sum_{j=1}^{1000} (\hat{\theta}_{jk} - \tilde{\theta}_{jk}) \right] \tag{5.1}$$

$$Var(B_{kj}) = \left[\frac{1}{1000 - 1} \sum_{j=1}^{1000} (B_{jk} - \bar{B}_k)^2 \right] \tag{5.2}$$

donde:
$\hat{\theta}_{jk}$ = es el valor estimado (proporción en los niveles de medición o media de las variables continua) para la $j-$ésima muestra, $j = 1, \ldots, 1000$ y el $k-$ésimo valor, $k = 1, \ldots, 36$.
$\tilde{\theta}_{jk}$ = es el valor muestral si todos los casos en la muestra hubieran respondido (proporción en los niveles de medición o media de las variables continuas)

para la $j-$ésima muestra, $j = 1, \ldots, 1000$ y el $k-$ésimo valor, $k = 1, \ldots, 36$.

Con las Ecuaciones 5.1 y 5.2 se obtuvieron la diferencia de sesgos relativos a los valores poblacionales

$$\hat{B}_{jk} = \left(\frac{\hat{\theta}_{jk} - \theta_k}{\theta_k} - \frac{\tilde{\theta}_{jk} - \theta_k}{\theta_k} \right) \times 100 = \frac{\hat{\theta}_{jk} - \tilde{\theta}_k}{\theta_k} \times 100 = \frac{B_{jk}}{\theta_k} \times 100 \quad (5.3)$$

y la longitud

$$L_c = \left| \sqrt{\frac{Var(B_{jk})}{1000}} \frac{z_{0.025}}{\theta_k} \right| \times 100, \quad (5.4)$$

respectivamente, con θ_k el $k-$ésimo valor poblacional. Los extremos de un intervalo de confianza de 95 % para $\hat{B}_{jk}$ se pueden calcular con $\hat{B}_{Ijk} = \hat{B}_{jk} - L_c$ y $\hat{B}_{Sjk} = \hat{B}_{jk} + L_c$. En las tablas que presentan los resultados obtenidos por el empleo de cada método, se indican los valores para $\hat{B}_{jk}$ y L_c.

No se evaluó la estimación de la varianza σ^2 de las variables continuas dado que habitualmente se considera un parámetro de *ruido*.

5.1. Métodos para la no respuesta por unidad o total

En esta sección se muestra la evaluación de un método para la no respuesta total para ser usado durante el levantamiento de datos pero que fue posible simular el procedimiento y cinco métodos usados después del levantamiento de datos. En todos los métodos resaltaron los valores desproporcionados del sesgo relativo $\hat{B}$ y de L_c en las variables $vnstd_1$ y $vnstd_2$ debido a que sus valores poblacionales $\overline{Y}_i$ estimados con los valores muestrales $\bar{y}_i$, $i = 7, 8$ eran cercanos a cero tanto en las muestras de $Pobl$ como en las de $PoblM$ por lo que cualquier diferencia aunque pequeña, entre las estimaciones y los valores verdaderos se presentaría extremadamente alta, por esta razón, en estas variables se presentaron los valores de sesgos B_{kj} y sus varianzas $Var(B_k)$, en lugar de $\hat{B}$ y L_c, respectivamente. Las etiquetas de las variables $vnstd_1$ y $vnstd_2$ fueron marcadas con un asterisco para señalar que sus indicadores de estimación no son directamente comparable con los de las restantes variables.

En las siguientes tablas donde se muestran los valores de $\hat{B}$ y L_c se incluyeron los valores poblacionales, de las distribuciones de proporciones y

medias de las variables que conformaron las poblaciones $Pobl$ y $PoblM$, en la columna valor a estimar (vae).

5.1.1. Método usado durante el levantamiento de los datos

5.1.1.1. Submuestreo de los no informantes

El método está concebido para ser empleado durante el levantamiento de datos, sin embargo, es posible simularlo con la extracción de una muestra de los casos que se consideraron como omisiones, para este caso en particular se eligió un 10 % de ellos con un esquema de selección aleatoria simple. Después se procedió a hacer las estimaciones de los valores como se describió en la Sección 2.2.2. Los resultados se presentan en la Tabla 5.1. Los valores negativos de $\hat{B}$ indican subestimaciones de los valores con el uso de este método tanto en las muestras de $Pobl$ como de $PoblM$. Los valores de L_c fueron menores al 1.0 % en los dos conjuntos de muestras, excepto en el nivel 2 en $vo10c_1$ de las muestras de $Pobl$ cuando la tasa de no repuesta (T_{nr}) fue de =10 % y cuando $T_{nr} = 30$ % en las muestras de $Pobl$, el rango observado de L_c se ubicó entre 0.3 % y 5.0 %. Los valores más grandes de L_c se observaron en los niveles de medición con proporciones más bajas sin importar la tasa de no respuesta o población de donde se extrajeron las muestras. Los sesgos de las estimaciones fueron muy homogéneos en $PobM$ ya que los rangos observados fueron, en valor absoluto, de 10.8 % a 13.7 % y de 27.5 % y 32.8 %, para $T_{nr} = 10$ % y $T_{nr} = 30$ %, respectivamente. En las muestras de $Pobl$ se observó que a medida que se incrementaban los niveles de medición el método tendió a mostrar al menos una estimación más imprecisa en comparación con las variables con menos niveles de medición. Es decir, cuando $T_{nr} = 10$ %, el mayor sesgo, en valor absoluto, de $vo10c_1$ y $vo10c_2$ fue 18.1 %, el de $vo4c_1$ y $vo4c_2$ fue 16.5 % y el de vod_1 y vod_2 fue 15.2 %. En el caso de $T_{nr} = 30$ % el mayor sesgo, en valor absoluto, de $vo10c_1$ y $vo10c_2$ fue 39.4 %, el de $vo4c_1$ y $vo4c_2$ fue 37.2 % y el de vod_1 y vod_2 fue 35.2 %. Para el caso de las variables continuas, los valores de $\hat{B}$ y de L_c se mantuvieron dentro de los rangos descritos para las variables ordinales en las dos tasas de no respuesta y poblaciones. Las estimaciones del parámetro de $vnstd_1$ fueron muy imprecisas ya que en las dos tasas de no respuesta y ambas poblaciones los sesgos fueron mayores al valor a estimar, en el caso de $vnstd_2$ los sesgos fueron menores a 0.01; en éstas dos últimas variable mencionadas la varianza de las estimaciones fueron menores a 0.001.

Tabla 5.1: Método de submuestreo de los no informantes. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	-10.98	0.08	-28.04	0.38	0.74	-12.03	0.09	-29.66	0.12
$vod_{1,1}$	0.23	-15.17	0.18	-35.16	0.85	0.26	-11.77	0.16	-29.51	0.24
$vod_{2,0}$	0.62	-13.17	0.10	-31.96	0.44	0.78	-11.94	0.08	-29.59	0.11
$vod_{2,1}$	0.38	-9.95	0.11	-25.98	0.58	0.22	-12.04	0.18	-29.77	0.26
$vo4c_{1,0}$	0.72	-10.76	0.08	-27.59	0.39	0.46	-11.94	0.11	-29.69	0.17
$vo4c_{1,1}$	0.10	-13.72	0.28	-32.67	1.35	0.15	-11.90	0.21	-29.62	0.33
$vo4c_{1,2}$	0.06	-14.68	0.38	-35.22	1.81	0.05	-12.20	0.37	-30.47	0.57
$vo4c_{1,3}$	0.12	-16.35	0.26	-37.16	1.26	0.33	-11.97	0.13	-29.41	0.20
$vo4c_{2,0}$	0.34	-9.38	0.12	-24.81	0.62	0.37	-12.12	0.12	-29.77	0.19
$vo4c_{2,1}$	0.05	-11.45	0.38	-28.65	1.83	0.10	-12.29	0.26	-29.96	0.39
$vo4c_{2,2}$	0.08	-10.79	0.27	-27.80	1.36	0.35	-11.58	0.13	-29.24	0.20
$vo4c_{2,3}$	0.53	-13.80	0.11	-33.16	0.49	0.17	-12.18	0.20	-29.92	0.31
$vo10c_{1,0}$	0.09	-18.07	0.33	-40.39	1.56	0.06	-11.80	0.34	-29.39	0.52
$vo10c_{1,1}$	0.11	-15.24	0.28	-35.63	1.28	0.16	-12.37	0.21	-29.90	0.32
$vo10c_{1,2}$	0.01	-16.03	1.06	-37.15	4.90	0.19	-12.24	0.18	-30.27	0.29
$vo10c_{1,3}$	0.04	-13.72	0.45	-33.05	2.12	0.11	-11.25	0.25	-28.64	0.40
$vo10c_{1,4}$	0.04	-13.98	0.43	-32.86	2.09	0.14	-11.81	0.21	-29.44	0.32
$vo10c_{1,5}$	0.01	-13.00	0.74	-32.07	3.69	0.03	-11.51	0.43	-29.56	0.73
$vo10c_{1,6}$	0.06	-12.96	0.35	-32.14	1.75	0.03	-12.10	0.50	-28.57	0.76
$vo10c_{1,7}$	0.04	-11.62	0.38	-29.20	1.89	0.04	-11.55	0.39	-29.77	0.63
$vo10c_{1,8}$	0.23	-11.47	0.17	-29.25	0.83	0.08	-12.23	0.29	-29.95	0.48
$vo10c_{1,9}$	0.36	-9.01	0.12	-24.09	0.59	0.17	-11.93	0.19	-29.52	0.30
$vo10c_{2,0}$	0.16	-7.50	0.16	-21.01	0.83	0.13	-12.31	0.23	-30.28	0.36
$vo10c_{2,1}$	0.15	-9.61	0.20	-25.56	0.99	0.04	-11.16	0.40	-28.85	0.59
$vo10c_{2,2}$	0.01	-9.47	0.67	-25.24	3.30	0.10	-11.82	0.27	-29.44	0.41
$vo10c_{2,3}$	0.21	-11.11	0.17	-28.67	0.81	0.13	-11.34	0.23	-28.31	0.34
$vo10c_{2,4}$	0.03	-11.43	0.44	-29.04	2.20	0.13	-11.75	0.22	-29.23	0.35
$vo10c_{2,5}$	0.02	-12.86	0.58	-30.76	2.94	0.11	-11.91	0.24	-29.46	0.37
$vo10c_{2,6}$	0.09	-12.94	0.29	-31.21	1.40	0.10	-12.14	0.27	-30.12	0.40
$vo10c_{2,7}$	0.08	-13.36	0.31	-32.69	1.51	0.12	-12.25	0.25	-30.18	0.38
$vo10c_{2,8}$	0.13	-14.74	0.25	-35.62	1.19	0.04	-12.74	0.42	-31.07	0.62
$vo10c_{2,9}$	0.11	-18.05	0.28	-39.42	1.35	0.10	-12.24	0.25	-29.91	0.38
$vnstd^*_{1,Y}$	-0.01	0.027	0.000	0.050	0.000	0.01	0.063	0.000	0.112	0.000
$vnstd^*_{2,Y}$	0.03	0.008	0.000	0.011	0.000	0.01	0.004	0.000	0.007	0.000
$vn_{1,Y}$	50.05	-11.56	0.07	-28.98	0.31	50.06	-10.79	0.07	-27.53	0.10
$vn_{2,Y}$	50.08	-13.69	0.08	-32.83	0.32	49.92	-13.73	0.08	-32.80	0.10

5.1.2. Métodos usados después del levantamiento de los datos

5.1.2.1. Ponderación simple

En el método de ponderación simple, los valores negativos y positivos de $\hat{B}$ de la Tabla 5.2 indican que se sobreestimaron y subestimaron los valores muestrales de los valores de las variables ordinales y continuas de $Pobl$ y $PoblM$. Los valores de $\hat{B}$ en las muestras de $Pobl$ se ubicaron entre $-8.1\,\%$ y $5.8\,\%$, en este mismo conjunto el valor más alto de L_c fue de $1.2\,\%$ en el nivel 2 de $vo10c_2$ cuando $T_{nr} = 10\,\%$, y cuando la tasa de no respuesta fue del $30\,\%$ los valores más grandes, en valor absoluto, de $\hat{B}$ se observaron en las variables $vo10c_1$ en el nivel 0, con $17.7\,\%$ y en el nivel 9 de $vo10c_2$, $16.8\,\%$. Los valores de L_c se ubicaron entre $0.03\,\%$ y $2.0\,\%$. En $PoblM$ cuando $T_{nr} = 10\,\%$ solamente en las variables vn_1 y vn_2, los valores absolutos de $\hat{B}$ fueron mayores al $1.0\,\%$, lo mismo sucedió cuando $T_{nr} = 30\,\%$, los mayores sesgos, en valor absoluto, correspondieron a las dos variables continuas mencionadas, $3.5\,\%$ y $5.3\,\%$, respectivamente. Los valores de L_c fueron menores a $1.0\,\%$ en esta misma población en ambas tasas de no respuesta. La estimación del parámetro de $vnstd_1$ fue muy imprecisa ya que los sesgo fueron mayores al parámetro estimar pero la peor estimación se halló en $PoblM$ cuando $T_{nr} = 30\,\%$. En la variable $vnstd_2$ las estimaciones del parámetros tampoco fueron precisas ya que el menor sesgo fue de 0.01. La varianza de los sesgos fueron mayores a 0.0001 en las dos tasas de no respuesta y poblaciones. (Ver Tabla 5.2).

5.1.2.2. Clases de ponderación

Para la creación de las clases de ponderación, se tomaron los cuartiles de variable auxiliar vn_2 para agrupar las muestras en cuatro submuestras de tamaños aproximadamente iguales. La aplicación de clases de ponderación incrementó los valores de $\hat{B}$ en comparación con los resultados obtenidos con la ponderación simple, sin embargo, la reducción de sesgos hubiera ocurrido si las clases de ponderación hubieran generado agrupaciones de casos más homogéneas que la muestra original, situación que no se verificó con las clases definidas por vn_2. En las muestras de $Pobl$, los valores menores de L_c se encontraron en las variables continuas vn_1 y vn_2 tanto en $T_{nr} = 10\,\%$ como en $T_{nr} = 30\,\%$. Por otra parte, los valores más altos de L_c se hallaron en el nivel 2 de $vo10c_1$, $2.1\,\%$ y $2.9\,\%$, en $T_{nr} = 10\,\%$ y en $T_{nr} = 30\,\%$, respectivamente. El rango de $\hat{B}$ en $T_{nr} = 10\,\%$ se halló entre $-17.2\,\%$ y $23.7\,\%$, en $T_{nr} = 30\,\%$ los sesgos relativos se ubicaron entre $-25.7\,\%$ y $39.7\,\%$.

Tabla 5.2: Método ponderación simple. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	1.21	0.06	2.74	0.09	0.74	-0.10	0.06	-0.16	0.10
$vod_{1,1}$	0.23	-4.05	0.18	-9.16	0.32	0.26	0.27	0.16	0.46	0.28
$vod_{2,0}$	0.62	-1.62	0.07	-3.80	0.13	0.78	0.08	0.05	0.08	0.09
$vod_{2,1}$	0.38	2.62	0.12	6.15	0.22	0.22	-0.30	0.18	-0.30	0.34
$vo4c_{1,0}$	0.72	1.52	0.06	3.40	0.11	0.46	0.09	0.10	-0.08	0.19
$vo4c_{1,1}$	0.1	-2.17	0.31	-5.16	0.52	0.15	-0.13	0.22	0.27	0.39
$vo4c_{1,2}$	0.06	-3.83	0.39	-8.32	0.73	0.05	-0.12	0.43	-0.30	0.73
$vo4c_{1,3}$	0.12	-5.55	0.30	-12.31	0.48	0.33	-0.04	0.14	0.04	0.24
$vo4c_{2,0}$	0.34	3.30	0.13	7.95	0.23	0.37	-0.12	0.12	-0.15	0.22
$vo4c_{2,1}$	0.05	1.14	0.42	2.09	0.72	0.10	-0.05	0.27	-0.96	0.48
$vo4c_{2,2}$	0.08	1.42	0.29	2.89	0.56	0.35	0.33	0.13	0.66	0.23
$vo4c_{2,3}$	0.53	-2.42	0.09	-5.70	0.16	0.17	-0.40	0.22	-0.46	0.38
$vo10c_{1,0}$	0.09	-8.05	0.35	-17.72	0.58	0.06	0.33	0.38	0.62	0.68
$vo10c_{1,1}$	0.11	-4.29	0.27	-10.64	0.50	0.16	-0.15	0.23	-0.79	0.38
$vo10c_{1,2}$	0.01	-3.58	1.19	-10.79	2.03	0.19	-0.49	0.21	-0.93	0.36
$vo10c_{1,3}$	0.04	-2.61	0.51	-6.08	0.86	0.11	0.73	0.27	1.30	0.48
$vo10c_{1,4}$	0.04	-2.47	0.47	-5.23	0.78	0.14	0.03	0.24	0.68	0.39
$vo10c_{1,5}$	0.01	-0.93	0.83	-2.67	1.55	0.03	0.41	0.52	0.88	0.89
$vo10c_{1,6}$	0.06	-1.14	0.39	-3.96	0.68	0.03	0.25	0.57	-0.13	0.97
$vo10c_{1,7}$	0.04	0.78	0.46	0.56	0.79	0.04	0.18	0.44	0.36	0.80
$vo10c_{1,8}$	0.23	0.56	0.17	1.26	0.32	0.08	-0.55	0.33	-0.68	0.60
$vo10c_{1,9}$	0.36	3.76	0.12	9.11	0.21	0.17	0.17	0.20	0.22	0.38
$vo10c_{2,0}$	0.16	5.78	0.19	14.49	0.35	0.13	-0.38	0.26	-0.86	0.44
$vo10c_{2,1}$	0.15	3.06	0.22	6.81	0.39	0.04	0.83	0.46	1.37	0.83
$vo10c_{2,2}$	0.01	3.47	0.72	7.10	1.42	0.10	0.08	0.29	0.60	0.55
$vo10c_{2,3}$	0.21	1.20	0.18	1.89	0.32	0.13	0.72	0.24	1.38	0.43
$vo10c_{2,4}$	0.03	0.37	0.49	0.89	0.88	0.13	0.11	0.24	0.73	0.45
$vo10c_{2,5}$	0.02	-0.76	0.63	-2.96	1.15	0.11	0.30	0.26	0.09	0.48
$vo10c_{2,6}$	0.09	-1.26	0.31	-2.91	0.57	0.10	-0.05	0.29	-0.51	0.52
$vo10c_{2,7}$	0.08	-1.72	0.35	-4.28	0.60	0.12	-0.43	0.26	-0.95	0.47
$vo10c_{2,8}$	0.13	-4.05	0.26	-9.79	0.43	0.04	-0.87	0.47	-1.79	0.81
$vo10c_{2,9}$	0.11	-7.84	0.31	-16.76	0.50	0.10	-0.39	0.28	-0.50	0.50
$vnstd^*_{1,Y}$	-0.01	0.032	0.000	0.077	0.001	0.01	0.083	0.000	0.192	0.001
$vnstd^*_{2,Y}$	0.03	0.014	0.000	0.030	0.001	0.01	0.006	0.000	0.015	0.001
$vn_{1,Y}$	50.05	0.50	0.02	1.20	0.03	50.06	1.53	0.02	3.49	0.03
$vn_{2,Y}$	50.08	-2.25	0.02	-5.27	0.03	49.92	-2.29	0.02	-5.31	0.03

En las muestras de $PoblM$ los sesgos fueron menores ya que se hallaron entre $-7.1\,\%$ y $4.7\,\%$ y entre $-10.5\,\%$ y $10.5\,\%$, para $T_{nr} = 10\,\%$ y $T_{nr} = 30\,\%$, respectivamente. De las misma forma que ocurrió en los dos métodos anteriores, las variables con un número mayor de valores a estimar, tuvieron los valores de L_c más altos: $vo10c_1$, nivel 6, $1.3\,\%$ y $1.8\,\%$, en las tasas de no respuesta del $10\,\%$ y $30\,\%$, respectivamente. Con el uso de este método tampoco se obtuvieron mejores estimaciones del parámetro de $vnstd_1$ ya que fueron mayores a 0.11, cuando el valor poblacional era -0.01 y 0.01 en $Pobl$ y $PoblM$, respectivamente. La estimación del parámetro de $vnstd_2$ fue mayor a 0.03 en $Pobl$ y a 0.02 en $PoblM$. Las varianzas de estimaciones de éstas dos variables fue superior a 0.001 (Ver Tabla 5.3).

5.1.2.3. Postestratificación

El método de postestratificación no fue evaluado en la estimación de proporciones y medias de las muestras de $Pobl$ y $PoblM$ debido a que sólo es una variación de método de clases de ponderación con variables auxiliares obtenidas de la población.

5.1.2.4. Procedimiento de estimadores de razón

Para el método de estimadores de razón se tomó como variable auxiliar vn_2, por lo que no tuvo sesgos relativos ni varianza relativas y sus valores de L_c fueron iguales a cero. En las muestras de $Pobl$ los sesgos se hallaron en un rango de $-6.1\,\%$ y $8.1\,\%$ en $T_{nr} = 10\,\%$ y entre $-12.7\,\%$ y $20.5\,\%$ en $T_{nr} = 30\,\%$. Los valores absolutos mayores de $\hat{B}$ se ubicaron en las variables con el mayor número de valores a estimar, es decir, $vo10c_1$ y $vo10c_2$. Los valores menores de L_c se encontraron en vn_1, en ambas tasas de no respuesta. Por otro lado, los valores mayores de L_c se detectaron en $vo10c_1$ en el nivel 2, $1.9\,\%$, en $T_{nr} = 10\,\%$ y $2.1\,\%$ cuando $T_{nr} = 30\,\%$. En las muestras de $PoblM$ sucedió que los sesgos más altos se verificaron en vn_1, $3.9\,\%$ y $9.3\,\%$, en las dos tasa de no respuesta del $10\,\%$ y $30\,\%$, respectivamente, pero también en esta misma variable se halló que tenía los valores más bajos de L_c, $0.04\,\%$ y $0.06\,\%$. En las variables ordinales todos los valores de L_c fueron menores a $1.0\,\%$, excepto por $vo10c_1$ en el nivel 6 en $T_{nr} = 30\,\%$; en las mismas variables ordinales los valores de $\hat{B}$ fueron sobreestimados dentro de un rango de $1.7\,\%$ a $3.3\,\%$ cuando $T_{nr} = 10\,\%$ y en un rango de $3.9\,\%$ a $7.6\,\%$ cuando $T_{nr} = 30\,\%$. Un hecho notable es que a pesar de aumentar el número de valores de una variable ordinal, en comparación otra ordinal con menos valores, los sesgos observados no aumentaron notoriamente, es decir,

Tabla 5.3: Método clases de ponderación. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	4.38	0.19	9.59	0.26	0.74	0.69	0.20	3.93	0.25
$vod_{1,1}$	0.23	-12.20	0.37	-15.72	0.50	0.26	2.34	0.39	5.91	0.53
$vod_{2,0}$	0.62	-4.60	0.22	-4.73	0.29	0.78	0.43	0.19	3.59	0.24
$vod_{2,1}$	0.38	8.93	0.32	17.54	0.43	0.22	3.63	0.42	7.55	0.57
$vo4c_{1,0}$	0.72	4.49	0.20	10.22	0.27	0.46	1.73	0.27	4.99	0.38
$vo4c_{1,1}$	0.10	-6.96	0.62	-8.60	0.82	0.15	-0.40	0.51	3.69	0.71
$vo4c_{1,2}$	0.06	-7.60	0.74	-10.57	1.06	0.05	0.51	0.94	3.87	1.25
$vo4c_{1,3}$	0.12	-13.11	0.51	-18.31	0.71	0.33	1.06	0.34	4.13	0.45
$vo4c_{2,0}$	0.34	12.60	0.35	22.62	0.46	0.37	2.11	0.31	5.81	0.42
$vo4c_{2,1}$	0.05	-1.48	1.00	3.44	1.33	0.10	-0.93	0.62	2.44	0.87
$vo4c_{2,2}$	0.08	-0.71	0.71	4.20	0.98	0.35	1.06	0.33	4.56	0.43
$vo4c_{2,3}$	0.53	-6.75	0.23	-8.31	0.31	0.17	0.31	0.47	2.42	0.66
$vo10c_{1,0}$	0.09	-15.85	0.57	-24.13	0.78	0.06	0.89	0.87	4.59	1.18
$vo10c_{1,1}$	0.11	-12.20	0.51	-18.03	0.71	0.16	-1.90	0.51	0.30	0.71
$vo10c_{1,2}$	0.01	-3.62	2.05	-6.86	2.87	0.19	0.99	0.46	3.72	0.62
$vo10c_{1,3}$	0.04	-8.69	0.92	-9.64	1.33	0.11	2.65	0.60	7.87	0.87
$vo10c_{1,4}$	0.04	-13.34	0.97	-14.57	1.26	0.14	1.63	0.55	5.72	0.72
$vo10c_{1,5}$	0.01	-9.67	1.72	-8.31	2.21	0.03	-2.54	1.14	-0.45	1.46
$vo10c_{1,6}$	0.06	-6.35	0.76	-7.12	1.09	0.03	4.77	1.28	9.21	1.76
$vo10c_{1,7}$	0.04	5.02	1.00	9.78	1.41	0.04	2.11	1.00	5.39	1.40
$vo10c_{1,8}$	0.23	0.93	0.41	4.39	0.60	0.08	0.91	0.73	4.67	1.04
$vo10c_{1,9}$	0.36	12.11	0.33	22.46	0.46	0.17	2.71	0.49	5.63	0.65
$vo10c_{2,0}$	0.16	23.74	0.56	39.70	0.78	0.13	-1.50	0.56	0.36	0.74
$vo10c_{2,1}$	0.15	11.20	0.58	20.41	0.75	0.04	3.54	1.03	7.44	1.38
$vo10c_{2,2}$	0.01	6.00	2.07	15.59	2.53	0.10	-0.72	0.64	3.11	0.89
$vo10c_{2,3}$	0.21	-0.24	0.44	3.08	0.58	0.13	0.99	0.57	5.26	0.80
$vo10c_{2,4}$	0.03	-2.41	1.10	2.13	1.56	0.13	5.80	0.58	10.52	0.76
$vo10c_{2,5}$	0.02	-7.18	1.33	-6.87	1.79	0.11	0.98	0.62	3.89	0.83
$vo10c_{2,6}$	0.09	-5.75	0.67	-6.64	0.91	0.10	-1.32	0.64	1.81	0.86
$vo10c_{2,7}$	0.08	-7.47	0.66	-8.90	0.95	0.12	0.66	0.61	3.37	0.80
$vo10c_{2,8}$	0.13	-12.49	0.49	-16.62	0.65	0.04	0.36	0.98	3.28	1.33
$vo10c_{2,9}$	0.11	-17.17	0.49	-25.72	0.66	0.10	2.65	0.62	5.62	0.87
$vnstd^*_{1,Y}$	-0.01	0.105	0.001	0.154	0.002	0.01	0.258	0.001	0.380	0.002
$vnstd^*_{2,Y}$	0.03	0.028	0.001	0.046	0.002	0.01	0.018	0.001	0.023	0.002
$vn_{1,Y}$	50.05	1.91	0.04	2.74	0.05	50.06	4.69	0.04	6.97	0.05
$vn_{2,Y}$	50.08	-7.24	0.04	-10.63	0.05	49.92	-7.08	0.03	-10.47	0.05

el método funcionó aproximadamente igual para estimar dos valores de la variable vod_1 como diez valores de la variable $vo10c_1$ considerando sólo los valores de los sesgos. Con respecto a las estimaciones de las variables $vnstd_1$ y $vnstd_2$ se repitió el mismo comportamiento que en los métodos de ponderación y clases de ponderación: los sesgos fueron mayores al parámetro a estimar, y la de mayor imprecisión ocurrió en la variable $vnstd_1$ en la muestra de $PoblM$ en la tasa de no respuesta del 10 % (Ver Tabla 5.4).

5.1.2.5. Procedimiento de estimadores de regresión

Para la instrumentación de este método se tomó como variable de auxiliar vn_2. En las muestras de $Pobl$ los valores de $\hat{B}$ se ubicaron en un rango de -8.6 % a 5.9 % cuando $T_{nr} = 10$ % y entre -17.9 % y 14.4 % con $T_{nr} = 30$ %. En la variables vn_1 se encontró el valor menor de L_c, 0.02 % en $T_{nr} = 10$ % y 0.01 % cuando $T_{nr} = 30$ %. Los valores más altos de L_c de las variables ordinales se hallaron en $vo10c_1$ en el nivel 2, 1.2 % y 1.9 % para $T_{nr} = 10$ % y $T_{nr} = 30$ %, respectivamente. En las muestras de $PoblM$ el método generó sesgos menores al 1.0 %, en valor absoluto, en las variables ordinales cuando $T_{nr} = 10$ %. En el caso de $T_{nr} = 30$ % los sesgos se ubicaron entre -5.3 % y 1.7 %. Los menores valores de L_c se encontraron en la variable continua vn_1 pero además en todas las variables fueron menores a 1.0 %, en las dos tasas de no respuesta. En las comparaciones de los sesgos de las muestras obtenidas de $Pobl$ y $PoblM$ los menores valores se obtuvieron para ésta última. Para este método, debido a la correlación alta, en valor absoluto, de $vnstd_1$ con la variable auxiliar vn_2 los sesgos se ubicaron entre -0.001 y 0.001 en los dos conjuntos de muestras de $Pobl$ y $PblM$ en las dos tasas de no respuesta. Por esta misma razón con una correlación, en valor absoluto, de 0.11 la variable auxiliar entre vn_2 y $vnstd_2$ sus sesgos fueron menores, en valor absoluto, a 0.002 en las dos tasas de no respuesta y en los dos conjuntos de muestras de $Pobl$ y $PblM$. La varianza de las estimaciones fue menor a 0.001 en las dos variables, en ambas tasas de no respuestas y en los dos conjuntos de muestras (Ver Tabla 5.5).

Tabla 5.4: Método procedimiento de estimadores de razón. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	3.56	0.07	8.46	0.11	0.74	2.27	0.06	5.39	0.11
$vod_{1,1}$	0.23	-1.88	0.19	-4.28	0.34	0.26	2.60	0.17	6.16	0.30
$vod_{2,0}$	0.62	0.72	0.08	1.56	0.14	0.78	2.39	0.06	5.76	0.10
$vod_{2,1}$	0.38	4.88	0.13	11.96	0.23	0.22	2.23	0.19	4.99	0.35
$vo4c_{1,0}$	0.72	3.79	0.07	9.18	0.12	0.46	2.39	0.11	5.60	0.19
$vo4c_{1,1}$	0.10	0.29	0.32	-0.34	0.55	0.15	2.37	0.23	5.84	0.43
$vo4c_{1,2}$	0.06	-1.23	0.42	-3.78	0.72	0.05	2.04	0.41	5.07	0.77
$vo4c_{1,3}$	0.12	-3.26	0.30	-7.29	0.51	0.33	2.34	0.14	5.54	0.25
$vo4c_{2,0}$	0.34	5.62	0.13	13.84	0.26	0.37	2.20	0.13	5.54	0.24
$vo4c_{2,1}$	0.05	3.62	0.42	8.14	0.83	0.10	2.30	0.30	5.20	0.52
$vo4c_{2,2}$	0.08	3.93	0.33	8.11	0.58	0.35	2.70	0.13	6.17	0.24
$vo4c_{2,3}$	0.53	-0.17	0.09	-0.40	0.16	0.17	2.00	0.22	4.71	0.40
$vo10c_{1,0}$	0.09	-6.12	0.36	-12.73	0.62	0.06	2.61	0.40	5.67	0.73
$vo10c_{1,1}$	0.11	-2.02	0.31	-5.17	0.51	0.16	2.08	0.23	4.65	0.42
$vo10c_{1,2}$	0.01	-2.90	1.19	-7.42	2.08	0.19	1.72	0.21	4.49	0.36
$vo10c_{1,3}$	0.04	-0.13	0.51	-0.76	0.90	0.11	3.12	0.28	7.60	0.54
$vo10c_{1,4}$	0.04	0.72	0.48	-0.08	0.88	0.14	2.47	0.23	6.10	0.44
$vo10c_{1,5}$	0.01	1.42	0.85	0.87	1.58	0.03	2.69	0.51	6.43	0.95
$vo10c_{1,6}$	0.06	1.34	0.40	2.05	0.72	0.03	2.71	0.57	6.06	1.07
$vo10c_{1,7}$	0.04	2.57	0.44	7.11	0.84	0.04	3.35	0.46	5.81	0.87
$vo10c_{1,8}$	0.23	2.83	0.18	6.45	0.34	0.08	1.88	0.34	5.64	0.64
$vo10c_{1,9}$	0.36	6.15	0.13	14.99	0.25	0.17	2.48	0.21	5.64	0.40
$vo10c_{2,0}$	0.16	8.22	0.20	20.55	0.39	0.13	1.84	0.25	4.20	0.47
$vo10c_{2,1}$	0.15	5.47	0.23	12.57	0.44	0.04	3.00	0.46	7.02	0.88
$vo10c_{2,2}$	0.01	6.26	0.76	10.62	1.47	0.10	2.50	0.30	5.87	0.55
$vo10c_{2,3}$	0.21	3.43	0.18	7.54	0.35	0.13	3.04	0.25	6.97	0.45
$vo10c_{2,4}$	0.03	3.14	0.51	7.13	0.91	0.13	2.65	0.25	6.39	0.46
$vo10c_{2,5}$	0.02	1.64	0.68	2.99	1.23	0.11	2.31	0.28	6.09	0.51
$vo10c_{2,6}$	0.09	0.88	0.32	2.41	0.58	0.10	2.11	0.31	4.87	0.54
$vo10c_{2,7}$	0.08	0.46	0.35	0.85	0.62	0.12	1.97	0.28	4.86	0.48
$vo10c_{2,8}$	0.13	-1.84	0.27	-4.45	0.49	0.04	1.68	0.47	3.91	0.83
$vo10c_{2,9}$	0.11	-5.56	0.32	-11.90	0.53	0.10	2.33	0.29	5.42	0.51
$vnstd^*_{1,Y}$	-0.01	0.034	0.000	0.079	0.001	0.01	0.084	0.000	0.203	0.001
$vnstd^*_{2,Y}$	0.03	0.014	0.000	0.033	0.001	0.01	0.005	0.000	0.016	0.001
$vn_{1,Y}$	50.05	2.82	0.03	6.81	0.05	50.06	3.91	0.04	9.27	0.06
$vn_{2,Y}$	50.08	0.00	0.00	0.00	0.00	49.92	0.00	0.00	0.00	0.00

Tabla 5.5: Método procedimiento de estimadores de regresión. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	1.23	0.05	2.76	0.10	0.74	-0.04	0.06	-0.10	0.10
$vod_{1,1}$	0.23	-4.10	0.18	-9.22	0.33	0.26	0.10	0.16	0.28	0.28
$vod_{2,0}$	0.62	-1.59	0.07	-3.81	0.13	0.78	0.03	0.05	-0.08	0.09
$vod_{2,1}$	0.38	2.57	0.12	6.15	0.21	0.22	-0.12	0.18	0.27	0.31
$vo4c_{1,0}$	0.72	1.54	0.06	3.45	0.11	0.46	-0.10	0.10	0.03	0.18
$vo4c_{1,1}$	0.10	-2.04	0.30	-5.16	0.54	0.15	0.26	0.23	-0.02	0.42
$vo4c_{1,2}$	0.06	-3.76	0.41	-9.19	0.70	0.05	-0.43	0.41	-1.00	0.71
$vo4c_{1,3}$	0.12	-5.69	0.28	-11.84	0.47	0.33	0.09	0.14	0.14	0.24
$vo4c_{2,0}$	0.34	3.20	0.13	7.73	0.23	0.37	-0.05	0.12	-0.11	0.22
$vo4c_{2,1}$	0.05	1.35	0.40	2.53	0.73	0.10	-0.09	0.28	-0.24	0.50
$vo4c_{2,2}$	0.08	1.62	0.30	3.67	0.55	0.35	0.22	0.12	0.51	0.22
$vo4c_{2,3}$	0.53	-2.36	0.09	-5.56	0.16	0.17	-0.29	0.21	-0.67	0.36
$vo10c_{1,0}$	0.09	-8.59	0.34	-17.85	0.59	0.06	-0.03	0.37	0.04	0.69
$vo10c_{1,1}$	0.11	-4.33	0.29	-11.04	0.47	0.16	-0.30	0.22	-0.67	0.39
$vo10c_{1,2}$	0.01	-7.19	1.18	-18.18	1.94	0.19	-0.73	0.21	-1.13	0.36
$vo10c_{1,3}$	0.04	-2.76	0.49	-6.41	0.85	0.11	0.62	0.27	1.43	0.47
$vo10c_{1,4}$	0.04	-2.58	0.48	-7.11	0.80	0.14	0.19	0.23	0.89	0.42
$vo10c_{1,5}$	0.01	-1.87	0.82	-7.07	1.49	0.03	0.80	0.51	0.52	0.93
$vo10c_{1,6}$	0.06	-1.67	0.39	-4.50	0.70	0.03	-0.36	0.57	1.18	0.96
$vo10c_{1,7}$	0.04	-0.15	0.42	-0.10	0.79	0.04	0.45	0.45	0.01	0.82
$vo10c_{1,8}$	0.23	0.66	0.17	0.74	0.33	0.08	-0.34	0.34	-0.49	0.58
$vo10c_{1,9}$	0.36	3.65	0.12	8.78	0.22	0.17	0.48	0.20	0.15	0.37
$vo10c_{2,0}$	0.16	5.86	0.18	14.43	0.37	0.13	-0.47	0.26	-0.69	0.44
$vo10c_{2,1}$	0.15	3.09	0.21	7.39	0.40	0.04	0.24	0.45	1.75	0.80
$vo10c_{2,2}$	0.01	5.49	0.72	12.48	1.40	0.10	0.27	0.30	0.11	0.51
$vo10c_{2,3}$	0.21	1.22	0.17	2.59	0.33	0.13	0.98	0.24	1.42	0.43
$vo10c_{2,4}$	0.03	1.19	0.50	2.97	0.87	0.13	0.34	0.24	1.27	0.42
$vo10c_{2,5}$	0.02	-0.26	0.67	0.73	1.16	0.11	0.06	0.28	0.50	0.48
$vo10c_{2,6}$	0.09	-1.17	0.30	-2.19	0.56	0.10	-0.38	0.29	-0.89	0.53
$vo10c_{2,7}$	0.08	-1.38	0.34	-4.00	0.60	0.12	-0.33	0.26	-0.98	0.46
$vo10c_{2,8}$	0.13	-3.74	0.26	-9.15	0.44	0.04	-0.95	0.45	-1.83	0.79
$vo10c_{2,9}$	0.11	-7.37	0.30	-16.29	0.49	0.10	-0.31	0.28	-0.96	0.51
$vnstd^*_{1,Y}$	-0.01	0.000	0.000	0.001	0.001	0.01	-0.001	0.000	-0.001	0.000
$vnstd^*_{2,Y}$	0.03	0.002	0.000	0.001	0.001	0.01	-0.002	0.000	-0.002	0.001
$vn_{1,Y}$	50.05	-0.03	0.02	-0.06	0.04	50.06	0.04	0.01	0.04	0.03
$vn_{2,Y}$	50.08	0.00	0.00	0.00	0.00	49.92	0.00	0.00	0.00	0.00

5.2. Métodos para la no respuesta parcial o por elemento

En la mayoría de los análisis estadísticos, además de obtener valores de los valores de la muestra, se requiere contar con un conjunto de datos rectangular completo para poder emplearlo en el cálculo de correlaciones entre variables, análisis de factores, ajustes de modelos de ecuaciones estructurales y todo el conjunto de análisis que requieren datos sin faltantes en todas las variables y casos, este objetivo se logra con los métodos para la no respuesta parcial o por elemento.

En las siguientes tablas donde se muestran los valores de $\hat{B}$ y L_c se incluyeron los valores poblacionales, de las distribuciones de proporciones y medias de las variables que conformaron las poblaciones $Pobl$ y $PoblM$, en la columna valor a estimar (vae).

5.2.1. Métodos usados después del levantamiento de los datos

En esta sección se consideró que la variable de referencia o auxiliar para no tener omisiones fuera vn_2 por lo que sus sesgos relativos y varianzas relativas fueron cero. Se repitió la situación de que los valores de sesgos relativos y varianzas relativas de $vnstd_1$ y $vnstd_2$ se muestran desproporcionados debido a que los valores poblacionales de los valores de interés fueron cercanos a cero, sin embargo, en el método de imputación múltiple no ocurrió así. Nuevamente, las etiquetas de las variables $vnstd_1$ y $vnstd_2$ fueron marcadas con un asterisco para señalar que sus indicadores de estimación no son directamente comparable con los de las restantes variables.

5.2.1.1. Análisis de casos completos

El análisis de casos completos es el equivalente al método de ponderación simple para la no respuesta total. Se puede observar en la Tabla 5.6 que en las muestras de $Pobl$, los sesgos se mantuvieron entre $-8.4\,\%$ a $5.9\,\%$ cuando $T_{nr} = 10\,\%$ y entre $0.02\,\%$ y $1.2\,\%$ los valores de L_c, cuyo valor más alto correspondió a la variable $vo10c_1$ en el nivel 2. Cuando $T_{nr} = 30\,\%$, los valores de $\hat{B}$ se ubicaron en el rango de $-17.7\,\%$ y $14.5\,\%$. Los valores más altos de L_c se encontraron en las variables ordinales y los más altos de todos en las variable $vo10c_1$ en el nivel 2, $1.9\,\%$ y en $vo10c_2$ en el nivel 2, $1.3\,\%$, los menores valores de L_c correspondieron a las variables continuas vn_1 y vn_2,

0.03 % en las dos. En $PoblM$, en las muestras todos los valores de L_c fueron menores a 1.0 % y los valores más bajos correspondieron a las variables continuas vn_1 y vn_2 en ambas tasas de no respuesta. En valor absoluto, los mayores sesgos no fueron mayores a 2.3 % cuando $T_{nr} = 10\,\%$ y menores a 5.3 % cuando $T_{nr} = 30\,\%$. En las muestras de $PoblM$ los valores más altos de $\hat{B}$ se encontraron en las variables vn_1 y vn_2 en las dos tasas de no respuesta. Una desventaja inherente a este método es que a pesar de que la variable auxiliar vn_2 no contenía omisiones, se generaron estimaciones sesgadas ya que todos los casos incompletos fueron descartados. En las variables $vnstd_1$ y $vnstd_2$ las varianzas de estimaciones fueron menores a 0.001 en las dos tasas de no respuesta y conjuntos de muestras de $Pobl$ y $PoblM$. Los sesgos de estimación fueron más imprecisas en la variable $vnstd_1$ en comparación de $vnstd_2$, ya que en la primer variable el menor sesgo, 0.032, se encontró en las muestras de $Pobl$ cuando $T_n = 10\,\%$ mientras que en menor sesgo de $vnstd_2$ se halló en las muestra de $PoblM$ cuando $T_n = 10\,\%$ y fue de 0.006.

5.2.1.2. Análisis de casos disponibles

En el análisis de casos disponibles, la variable vn_2 no tuvo omisiones por lo que los valores de $\hat{B}$ y L_c fueron iguales a cero. En las muestras de $Pobl$ los valores de L_c fueron menores a 1.0 % cuando $T_{nr} = 10\,\%$ y en el caso de $T_{nr} = 30\,\%$, sólo en la variable $vo10c1_1$ en el nivel 2 el valor de L_c fue mayor a 1.0 %. En las dos tasas de no respuesta en la variable continua vn_1 se halló que los valores más bajos de L_c de todas las variables de $Pobl$. Los sesgos $\hat{B}$ se ubicaron entre $-3.8\,\%$ y 2.6 % en $T_{nr} = 10\,\%$. En la otra tasa de no respuesta de 30 % los sesgos tuvieron un rango de $-7.3\,\%$ a 6.5 %. En las muestras de $PoblM$ todos los valores de L_c fueron menores a 1.0 % y los sesgos $\hat{B}$ menores a 0.8 %, en valor absoluto, en las dos tasas de no respuesta del 10 % y 30 %. En la variable continua vn_1 se encontraron los valores menores de L_c de todos los valores a estimar de $PoblM$. En este método también sucedió que a medida que se incrementaban los números de valores a estimar, surgían valores más altos de sesgos $\hat{B}$, en valor absoluto, en comparación con otras variables con un número menor de valores a estimar. En este método las varianzas de las estimaciones de $vnstd_1$ y $vnstd_2$ fueron menor a 0.0001 en las dos tasas de no respuestas y en ambos conjuntos de muestras de $Pobl$ y $PoblM$. En las muestras de $PoblM$ se observaron las estimaciones más imprecisas para $vnstd_1$ y las más precisas para $vnstd_2$ en comparación con los sesgos calculados para las muestras de $Pobl$ en las dos tasas de no respuesta (Ver Tabla 5.7).

Tabla 5.6: Método análisis de casos completos. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	1.20	0.06	2.79	0.09	0.74	-0.11	0.06	-0.05	0.10
$vod_{1,1}$	0.23	-4.02	0.19	-9.33	0.31	0.26	0.31	0.16	0.15	0.29
$vod_{2,0}$	0.62	-1.61	0.07	-3.87	0.13	0.78	0.07	0.05	0.03	0.09
$vod_{2,1}$	0.38	2.61	0.12	6.26	0.20	0.22	-0.24	0.18	-0.09	0.33
$vo4c_{1,0}$	0.72	1.53	0.06	3.45	0.11	0.46	-0.13	0.11	0.14	0.18
$vo4c_{1,1}$	0.10	-2.39	0.30	-5.58	0.52	0.15	0.09	0.22	0.03	0.41
$vo4c_{1,2}$	0.06	-3.57	0.40	-8.41	0.67	0.05	-0.22	0.43	-0.50	0.74
$vo4c_{1,3}$	0.12	-5.61	0.29	-12.26	0.50	0.33	0.17	0.14	-0.13	0.23
$vo4c_{2,0}$	0.34	3.34	0.13	7.93	0.23	0.37	-0.08	0.12	-0.13	0.23
$vo4c_{2,1}$	0.05	1.13	0.41	2.39	0.74	0.10	-0.23	0.29	-0.46	0.50
$vo4c_{2,2}$	0.08	1.59	0.30	2.90	0.56	0.35	0.29	0.12	0.66	0.23
$vo4c_{2,3}$	0.53	-2.48	0.09	-5.72	0.16	0.17	-0.28	0.21	-0.80	0.39
$vo10c_{1,0}$	0.09	-8.44	0.37	-17.66	0.57	0.06	0.25	0.38	0.18	0.68
$vo10c_{1,1}$	0.11	-4.12	0.28	-10.08	0.47	0.16	-0.41	0.22	-0.83	0.39
$vo10c_{1,2}$	0.01	-4.77	1.18	-10.81	1.94	0.19	-0.65	0.20	-0.98	0.34
$vo10c_{1,3}$	0.04	-2.37	0.50	-5.92	0.89	0.11	0.89	0.26	1.55	0.46
$vo10c_{1,4}$	0.04	-2.08	0.47	-5.42	0.83	0.14	0.32	0.24	0.57	0.41
$vo10c_{1,5}$	0.01	-0.81	0.83	-2.96	1.47	0.03	0.31	0.52	0.31	0.93
$vo10c_{1,6}$	0.06	-1.25	0.37	-3.27	0.69	0.03	0.58	0.52	0.35	0.94
$vo10c_{1,7}$	0.04	0.66	0.43	1.06	0.79	0.04	0.79	0.45	1.25	0.81
$vo10c_{1,8}$	0.23	0.45	0.18	1.02	0.32	0.08	-0.71	0.33	-0.31	0.58
$vo10c_{1,9}$	0.36	3.85	0.12	8.91	0.22	0.17	0.15	0.20	0.04	0.36
$vo10c_{2,0}$	0.16	5.87	0.18	14.46	0.36	0.13	-0.25	0.24	-1.08	0.46
$vo10c_{2,1}$	0.15	2.83	0.21	6.90	0.41	0.04	0.55	0.43	1.30	0.78
$vo10c_{2,2}$	0.01	3.38	0.75	8.04	1.32	0.10	0.41	0.29	0.47	0.52
$vo10c_{2,3}$	0.21	1.31	0.18	1.96	0.32	0.13	0.76	0.24	1.50	0.44
$vo10c_{2,4}$	0.03	0.43	0.51	1.32	0.87	0.13	0.32	0.25	1.01	0.43
$vo10c_{2,5}$	0.02	-0.75	0.65	-2.42	1.19	0.11	0.05	0.27	0.87	0.47
$vo10c_{2,6}$	0.09	-0.99	0.31	-2.99	0.56	0.10	-0.25	0.29	-1.00	0.51
$vo10c_{2,7}$	0.08	-1.55	0.35	-4.98	0.59	0.12	-0.47	0.27	-1.23	0.49
$vo10c_{2,8}$	0.13	-4.20	0.28	-9.55	0.44	0.04	-1.12	0.45	-1.56	0.80
$vo10c_{2,9}$	0.11	-8.06	0.31	-17.02	0.52	0.10	-0.46	0.27	-0.72	0.50
$vnstd^*_{1,Y}$	-0.01	0.032	0.000	0.077	0.001	0.01	0.082	0.000	0.192	0.001
$vnstd^*_{2,Y}$	0.03	0.015	0.000	0.030	0.001	0.01	0.006	0.000	0.016	0.001
$vn_{1,Y}$	50.05	0.50	0.02	1.20	0.03	50.06	1.51	0.02	3.48	0.03
$vn_{2,Y}$	50.08	-2.25	0.02	-5.26	0.03	49.92	-2.29	0.02	-5.29	0.03

Tabla 5.7: Método análisis de casos disponibles. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	0.60	0.04	1.10	0.06	0.74	-0.02	0.04	-0.08	0.07
$vod_{1,1}$	0.23	-2.02	0.13	-3.70	0.20	0.26	0.07	0.11	0.23	0.19
$vod_{2,0}$	0.62	-0.78	0.05	-1.56	0.08	0.78	0.03	0.03	0.01	0.06
$vod_{2,1}$	0.38	1.26	0.08	2.52	0.14	0.22	-0.09	0.12	-0.04	0.21
$vo4c_{1,0}$	0.72	0.71	0.04	1.37	0.07	0.46	0.01	0.07	0.10	0.12
$vo4c_{1,1}$	0.10	-0.94	0.21	-2.16	0.34	0.15	0.02	0.16	-0.03	0.26
$vo4c_{1,2}$	0.06	-1.83	0.28	-3.46	0.45	0.05	0.22	0.27	-0.34	0.48
$vo4c_{1,3}$	0.12	-2.66	0.20	-4.87	0.31	0.33	-0.05	0.09	-0.08	0.15
$vo4c_{2,0}$	0.34	1.49	0.09	3.15	0.14	0.37	-0.02	0.08	-0.13	0.14
$vo4c_{2,1}$	0.05	0.43	0.28	0.35	0.46	0.10	-0.17	0.20	-0.26	0.32
$vo4c_{2,2}$	0.08	0.63	0.21	0.83	0.37	0.35	0.13	0.09	0.31	0.14
$vo4c_{2,3}$	0.53	-1.09	0.06	-2.17	0.10	0.17	-0.14	0.14	-0.22	0.24
$vo10c_{1,0}$	0.09	-3.71	0.25	-7.27	0.39	0.06	-0.04	0.25	0.01	0.44
$vo10c_{1,1}$	0.11	-2.03	0.20	-3.93	0.32	0.16	-0.06	0.15	-0.27	0.24
$vo10c_{1,2}$	0.01	-1.40	0.77	-4.41	1.31	0.19	-0.15	0.14	-0.44	0.22
$vo10c_{1,3}$	0.04	-1.15	0.34	-2.62	0.56	0.11	0.18	0.18	0.62	0.31
$vo10c_{1,4}$	0.04	-0.94	0.34	-2.30	0.53	0.14	0.20	0.15	0.17	0.26
$vo10c_{1,5}$	0.01	-0.41	0.58	-1.18	0.95	0.03	0.09	0.36	0.62	0.54
$vo10c_{1,6}$	0.06	-0.38	0.26	-1.10	0.43	0.03	-0.23	0.38	0.46	0.62
$vo10c_{1,7}$	0.04	0.11	0.30	-0.01	0.49	0.04	0.14	0.30	0.18	0.51
$vo10c_{1,8}$	0.23	0.22	0.12	0.50	0.20	0.08	-0.26	0.23	-0.08	0.35
$vo10c_{1,9}$	0.36	1.75	0.08	3.59	0.14	0.17	0.07	0.14	-0.01	0.24
$vo10c_{2,0}$	0.16	2.63	0.12	5.80	0.22	0.13	-0.12	0.17	-0.29	0.27
$vo10c_{2,1}$	0.15	1.55	0.14	2.90	0.25	0.04	0.48	0.30	0.31	0.52
$vo10c_{2,2}$	0.01	1.30	0.48	3.10	0.90	0.10	-0.18	0.19	0.16	0.32
$vo10c_{2,3}$	0.21	0.64	0.12	0.61	0.21	0.13	0.31	0.16	0.79	0.27
$vo10c_{2,4}$	0.03	0.08	0.32	0.44	0.56	0.13	0.20	0.16	-0.05	0.27
$vo10c_{2,5}$	0.02	-0.73	0.45	-0.50	0.76	0.11	0.14	0.18	0.43	0.30
$vo10c_{2,6}$	0.09	-0.80	0.22	-1.43	0.35	0.10	0.10	0.19	-0.32	0.33
$vo10c_{2,7}$	0.08	-0.91	0.23	-1.83	0.41	0.12	-0.29	0.19	-0.47	0.30
$vo10c_{2,8}$	0.13	-1.66	0.18	-3.93	0.29	0.04	-0.33	0.30	-0.59	0.50
$vo10c_{2,9}$	0.11	-3.78	0.21	-6.52	0.34	0.10	-0.30	0.19	-0.22	0.31
$vnstd^*_{1,Y}$	-0.01	0.015	0.000	0.031	0.000	0.01	0.038	0.000	0.077	0.000
$vnstd^*_{2,Y}$	0.03	0.007	0.000	0.012	0.000	0.01	0.002	0.000	0.005	0.000
$vn_{1,Y}$	50.05	0.23	0.01	0.48	0.02	50.06	0.70	0.01	1.39	0.02
$vn_{2,Y}$	50.08	0.00	0.00	0.00	0.00	49.92	0.00	0.00	0.00	0.00

5.2.1.3. Imputación hot-deck

El método *hot-deck* en las muestra de $Pobl$ cuando $T_{nr} = 10\,\%$ mantuvo los valores de $\hat{B}$ en el rango de $-4.0\,\%$ a $2.7\,\%$ donde el mayor valor de L_c se encontró en $vo10c_1$ en el nivel 2 con $1.1\,\%$ y el menor valor de L_c correspondió a la variable vn_1, $0.02\,\%$; en el mismo conjunto de muestras cuando $T_{nr} = 30\,\%$ los valores de $\hat{B}$ tuvieron un rango de variación de $-7.0\,\%$ a $5.4\,\%$ y en el caso de L_c ocurrió el valor máximo en la variable $vo10c_1$ en el nivel de 2 con $1.6\,\%$ y el valor mínimo de L_c en la variable vn_1 con $0.03\,\%$. En las muestras de $PoblM$ los sesgos se mantuvieron en un rango de $-0.6\,\%$ a $0.7\,\%$, éste último valor correspondió a la variable continua vn_1; en este conjunto de muestras los valores de L_c fueron muy homogéneos ya que el máximo valor de L_c fue $0.5\,\%$ y el menor fue de $0.02\,\%$. A pesar de que la tasa de no respuesta aumentó, los sesgos no tuvieron incrementos notables como se puede ver en el rango obtenido: de $0.5\,\%$ a $1.4\,\%$, en el mismo sentido los valores máximo de L_c, $0.7\,\%$ y mínimo, $0.03\,\%$. Los casos con mayores sesgos se encontraron en las variables continuas pero con los valores mínimos de L_c. Para el caso de las variables $vnstd_1$ y $vnstd_2$ tuvo un desempeño similar al método de casos disponibles ya que las varianzas de las estimaciones fueron menores a 0.001 y los sesgos de estimaciones fueron mayores en la variable $vnstd_1$ en las muestras de $PoblM$, en comparación con las de $Pobl$ y para el caso de $vnstd_2$ los sesgos menores se encontraron en las muestras dep $PoblM$. Sin embargo, en $vnstd_1$ los sesgos fueron mayores al parámetro a estimar (Ver Tabla 5.8).

5.2.1.4. Imputación con la media

De acuerdo a lo descrito en la sección 3.2.4, la imputación con la media de los datos observados $\bar{y}_{obs}$ no cambia la estimación de la media del total de la muestra, es decir, equivale a ignorar los datos faltantes debido a que la media que se obtiene después de la imputación es igual a la de los datos observados; por otra parte, la estimación de la varianza total S^2 es igual a la de los datos observados S^2_{obs} reducida por un factor de $(n-k-1)/(n-1)$. Por lo que las estimaciones de valores con el uso de la imputación de la media, dependerá totalmente de los datos observados y además se obtendrá una estimación sesgada de la varianza muestral verdadera de los datos, una variante al método es considerar sólo los casos completos o los casos disponibles para estimar $\bar{y}_{obs}$.

Tabla 5.8: Método *hot-deck*. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	0.54	0.05	1.07	0.08	0.74	-0.06	0.05	-0.01	0.09
$vod_{1,1}$	0.23	-1.80	0.17	-3.57	0.28	0.26	0.16	0.15	0.02	0.25
$vod_{2,0}$	0.62	-0.73	0.07	-1.46	0.12	0.78	0.05	0.05	0.00	0.08
$vod_{2,1}$	0.38	1.19	0.11	2.37	0.19	0.22	-0.20	0.17	0.01	0.28
$vo4c_{1,0}$	0.72	0.71	0.06	1.33	0.09	0.46	-0.02	0.10	0.07	0.15
$vo4c_{1,1}$	0.10	-0.89	0.27	-1.67	0.47	0.15	0.13	0.21	-0.05	0.35
$vo4c_{1,2}$	0.06	-1.54	0.37	-3.68	0.59	0.05	0.31	0.37	0.23	0.62
$vo4c_{1,3}$	0.12	-2.80	0.26	-4.89	0.42	0.33	-0.08	0.13	-0.11	0.20
$vo4c_{2,0}$	0.34	1.45	0.12	3.31	0.20	0.37	0.04	0.11	0.03	0.19
$vo4c_{2,1}$	0.05	0.54	0.38	0.81	0.64	0.10	-0.15	0.27	-0.51	0.43
$vo4c_{2,2}$	0.08	0.81	0.31	1.00	0.49	0.35	0.16	0.12	0.31	0.21
$vo4c_{2,3}$	0.53	-1.10	0.08	-2.34	0.14	0.17	-0.34	0.20	-0.40	0.33
$vo10c_{1,0}$	0.09	-3.97	0.32	-6.96	0.48	0.06	0.11	0.37	0.06	0.60
$vo10c_{1,1}$	0.11	-1.94	0.25	-4.26	0.41	0.16	0.03	0.21	-0.45	0.33
$vo10c_{1,2}$	0.01	-2.40	1.08	-5.47	1.65	0.19	-0.40	0.19	-0.43	0.31
$vo10c_{1,3}$	0.04	-1.23	0.46	-2.74	0.73	0.11	0.24	0.26	0.61	0.42
$vo10c_{1,4}$	0.04	-1.31	0.44	-1.64	0.71	0.14	-0.02	0.22	0.35	0.37
$vo10c_{1,5}$	0.01	-0.40	0.76	-1.62	1.25	0.03	0.22	0.50	0.37	0.79
$vo10c_{1,6}$	0.06	-0.50	0.36	-0.94	0.56	0.03	0.29	0.50	0.03	0.82
$vo10c_{1,7}$	0.04	0.41	0.42	0.15	0.67	0.04	0.14	0.42	0.48	0.70
$vo10c_{1,8}$	0.23	0.39	0.17	0.43	0.26	0.08	-0.01	0.31	-0.31	0.50
$vo10c_{1,9}$	0.36	1.73	0.12	3.59	0.19	0.17	0.12	0.20	0.14	0.32
$vo10c_{2,0}$	0.16	2.66	0.19	5.44	0.31	0.13	-0.13	0.23	-0.53	0.38
$vo10c_{2,1}$	0.15	1.44	0.21	3.10	0.35	0.04	0.46	0.42	0.31	0.69
$vo10c_{2,2}$	0.01	1.90	0.74	3.13	1.20	0.10	0.12	0.27	0.02	0.44
$vo10c_{2,3}$	0.21	0.44	0.17	0.79	0.28	0.13	0.48	0.23	0.52	0.39
$vo10c_{2,4}$	0.03	0.04	0.49	-0.21	0.76	0.13	0.11	0.24	0.64	0.39
$vo10c_{2,5}$	0.02	-0.68	0.66	-0.50	1.02	0.11	0.26	0.25	0.01	0.41
$vo10c_{2,6}$	0.09	-0.74	0.31	-1.33	0.48	0.10	-0.19	0.28	-0.31	0.44
$vo10c_{2,7}$	0.08	-0.80	0.30	-1.69	0.51	0.12	-0.30	0.26	-0.38	0.40
$vo10c_{2,8}$	0.13	-1.72	0.24	-4.12	0.38	0.04	-0.60	0.42	-0.85	0.66
$vo10c_{2,9}$	0.11	-3.42	0.28	-6.36	0.41	0.10	-0.37	0.26	0.12	0.43
$vnstd^*_{1,Y}$	-0.01	0.015	0.000	0.029	0.001	0.01	0.038	0.000	0.078	0.001
$vnstd^*_{2,Y}$	0.03	0.006	0.000	0.012	0.001	0.01	0.002	0.000	0.006	0.001
$vn_{1,Y}$	50.05	0.25	0.02	0.48	0.03	50.06	0.69	0.02	1.39	0.03
$vn_{2,Y}$	50.08	0.00	0.00	0.00	0.00	49.92	0.00	0.00	0.00	0.00

5.2.1.5. Ajuste por la variable ficticia

El ajuste por la variable ficticia mas que ser un método de imputación de datos faltantes es una alternativa para que en el análisis de regresión, empleando como sustituto de los datos faltantes la media de los datos observados $\bar{y}_{obs}$, la variable ficticia D —que es igual a 1 si hay una omisión en la variable de interés y 0 en otro caso— resuma la diferencia de predicciones de la variable Y entre los datos observados y omitidos y además se pueda estimar el efecto de los datos observados en la predicción de Y. El objetivo de este estudio fue evaluar los métodos para llenar los vacíos en una matriz de datos para poder estimar los valores muestrales verdaderos por tal razón este método no fue evaluado.

5.2.1.6. Método de regresión

La utilización de este método de imputación requería contar con un conjunto de variables predictoras. En la Tabla 4.3 se puede observar que las correlaciones eran cercanas a cero en las variables ordinales de la población $PoblM$ por lo que se decidió no evaluar el método de regresión para este conjunto de datos, debido a que el desempeño del método con las variables continuas se observaría en las muestras de $Pobl$ donde sus correlaciones permitieron evaluar el método en todas sus variables. Para la aplicación de este método se empleó la librería MICE de SPLUS (van Buuren y Oudshoorn, 2000). En la tasa de no respuesta del 10 % se observó que el método funcionó de manera deficiente ya que en la variable $vo10c_1$ se encontró un sesgo del 166.2 % en el nivel 2, de 37.7 % en el nivel 5 de la misma variable y en $vo10c_2$ en el nivel 2, de 92.5 % y de 68.8 % en el nivel 4. En las variables ordinales vod_1 y vod_2 se hallaron sesgos de -16.8 % y -11.1 %, pero los sesgos $\hat{B}$ de las variables $vo4c_1$ y $v04c_2$ se ubicaron en el rango de -0.52 % a 3.0 %. Respecto a los valores de L_c, su rango de variación fue de 0.02 % a 2.5 %. Con el incremento de la tasa de no respuesta a 30 %, los sesgos también se incrementaron y otra vez las variables $vo10c_1$ y $vo10c_2$ mostraron valores con valores notablemente altos, 504.4 % y 300.1 %, respectivamente. El rango de L_c fue de 0.03 % a 4.9 %. Debido a la alta correlación que tenía la variable auxiliar vn_2 con vn_1, sus valores de $\hat{B}$ y L_c fueron los menores en comparación con las variables ordinales. Para el caso de las variables $vnstd_1$ y $vnstd_1$ los sesgos de estimación fueron de 0.001 y de -0.002, respectivamente, en las dos tasas de no respuesta y la varianza de los sesgos de estimación fue menor a 0.001 (Ver Tabla 5.9).

Tabla 5.9: Método de regresión. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	3.87	0.06	10.13	0.10	–	–	–	–	–
$vod_{1,1}$	0.23	-16.76	0.19	-41.85	0.38	–	–	–	–	–
$vod_{2,0}$	0.62	5.58	0.09	14.65	0.14	–	–	–	–	–
$vod_{2,1}$	0.38	-11.13	0.13	-28.79	0.21	–	–	–	–	–
$vo4c_{1,0}$	0.72	-0.07	0.07	-0.09	0.12	–	–	–	–	–
$vo4c_{1,1}$	0.10	0.19	0.42	-1.21	0.72	–	–	–	–	–
$vo4c_{1,2}$	0.06	1.15	0.57	2.25	0.99	–	–	–	–	–
$vo4c_{1,3}$	0.12	-0.34	0.33	0.41	0.56	–	–	–	–	–
$vo4c_{2,0}$	0.34	-0.05	0.09	-0.28	0.16	–	–	–	–	–
$vo4c_{2,1}$	0.05	1.11	0.56	-1.15	1.01	–	–	–	–	–
$vo4c_{2,2}$	0.08	3.00	1.05	0.44	0.72	–	–	–	–	–
$vo4c_{2,3}$	0.53	-0.52	0.17	0.22	0.11	–	–	–	–	–
$vo10c_{1,0}$	0.09	-8.55	0.21	-11.12	0.70	–	–	–	–	–
$vo10c_{1,1}$	0.11	19.34	0.09	27.17	0.37	–	–	–	–	–
$vo10c_{1,2}$	0.01	166.22	2.48	504.41	3.39	–	–	–	–	–
$vo10c_{1,3}$	0.04	-2.55	0.90	20.47	0.98	–	–	–	–	–
$vo10c_{1,4}$	0.04	1.18	0.25	-1.59	0.41	–	–	–	–	–
$vo10c_{1,5}$	0.01	37.69	1.76	88.20	0.87	–	–	–	–	–
$vo10c_{1,6}$	0.06	4.24	0.36	-45.25	0.47	–	–	–	–	–
$vo10c_{1,7}$	0.04	2.25	0.67	-11.28	0.35	–	–	–	–	–
$vo10c_{1,8}$	0.23	-4.48	0.32	-16.37	0.15	–	–	–	–	–
$vo10c_{1,9}$	0.36	-7.10	0.34	-3.18	0.13	–	–	–	–	–
$vo10c_{2,0}$	0.16	-1.25	0.21	-0.65	0.31	–	–	–	–	–
$vo10c_{2,1}$	0.15	4.17	0.40	3.70	0.42	–	–	–	–	–
$vo10c_{2,2}$	0.01	92.53	2.34	300.14	4.94	–	–	–	–	–
$vo10c_{2,3}$	0.21	-10.74	0.46	-28.68	0.65	–	–	–	–	–
$vo10c_{2,4}$	0.03	68.80	2.25	151.74	3.69	–	–	–	–	–
$vo10c_{2,5}$	0.02	4.90	1.26	21.33	2.15	–	–	–	–	–
$vo10c_{2,6}$	0.09	-9.93	0.50	-19.18	0.57	–	–	–	–	–
$vo10c_{2,7}$	0.08	-0.64	0.65	2.11	0.94	–	–	–	–	–
$vo10c_{2,8}$	0.13	-4.93	0.46	-12.86	0.86	–	–	–	–	–
$vo10c_{2,9}$	0.11	-2.22	0.63	-6.71	1.15	–	–	–	–	–
$vnstd^*_{1,Y}$	-0.01	0.001	0.000	0.001	0.001	–	–	–	–	–
$vnstd^*_{2,Y}$	0.03	-0.002	0.000	-0.002	0.000	–	–	–	–	–
$vn_{1,Y}$	50.05	0.00	0.02	-0.07	0.03	–	–	–	–	–
$vn_{2,Y}$	50.08	0.00	0.00	0.00	0.00	–	–	–	–	–

5.2.1.7. Imputación múltiple

Para la aplicación de este método se empleó la librería NORM de SPLUS (Schafer, 1999). En las muestras de *Pobl* se hallaron sesgos $\hat{B}$ demasiados grandes en comparación con los métodos de casos totales, casos disponibles y *hot-deck* ya que, en $T_{nr} = 10\,\%$ el rango se ubicó de $-6.7\,\%$ a $77.1\,\%$ pero también el rango de variación de L_c fue muy grande, el valor mínimo fue de $-5.4\,\%$ a $11.5\,\%$. En el caso de $T_{nr} = 30\,\%$ los valores de $\hat{B}$ variaron entre $-20.9\,\%$ y $191.4\,\%$ y los de L_c entre $-8.6\,\%$ a $11.3\,\%$. Los valores positivos de los sesgos indica que los errores más grandes correspondieron a sobreestimaciones. También es notable que las estimaciones en las variables continuas $vnstd_1$ y $vnstd_2$, los valores de $\hat{B}$ y L_c fueron menores, en valor absoluto, a $10.0\,\%$ situación que no ocurrió en los demás métodos de imputación. En $PoblM$ el rango de variación de $\hat{B}$ tuvo como límites $-30.5\,\%$ y $43.6\,\%$ y el de L_c cuando $T_{nr} = 10\,\%$ fue $0.01\,\%$ a $11.5\,\%$. La tasa de no respuesta del $30\,\%$ ocasionó sesgos mayores de estimación $\hat{B}$, su rango se ubicó entre $-73.6\,\%$ y $107.4\,\%$. Los valores de L_c se encontraron entre $0.07\,\%$ y $15.0\,\%$. En este conjunto de muestras el método funcionó peor que en las muestras de *Pobl*. El método no se volvió impreciso a pesar de que se aumentaba el número de valores a estimar de una variable a otra con menos valores a estimar. Debido a la alta correlación negativa (-0.23) que tenía vn_1 con la variable auxiliar vn_2 sus estimaciones fueron muy precisas, sus sesgos fueron menores a $0.02\,\%$, en valor absoluto, y con los valores de L_c muy reducidos, también menores a $0.02\,\%$. En las variables $vnstd_1$ y $vnstd_2$ las varianzas de los sesgos de estimación fueron menores a 0.0001 en las dos tasas de no respuesta y en los conjuntos de muestras de *Pobl* y $PoblM$. Los sesgos de estimación en las dos variables, en ambas tasas de no respuesta y en los dos conjuntos de muestras fueron las más precisas de todos los métodos evaluados de no respuesta parcial ya que fueron menores a 0.001, en valor absoluto. (Ver Tabla 5.10).

Tabla 5.10: Método imputación múltiple. Valores de $\hat{B}$ y de L_c de proporciones en los niveles de medición de las variables ordinales y medias de las variables continuas

		Pobl					*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$			$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
$var_{k,nm}$	vae	$\hat{B}$	L_c	$\hat{B}$	L_c	vae	$\hat{B}$	L_c	$\hat{B}$	L_c
$vod_{1,0}$	0.77	-1.10	3.42	1.39	3.33	0.74	-1.34	4.14	-1.43	4.15
$vod_{1,1}$	0.23	3.67	11.46	-4.65	11.14	0.26	3.81	11.76	4.05	11.79
$vod_{2,0}$	0.62	0.58	0.06	1.16	0.10	0.78	-1.21	0.04	-3.11	0.07
$vod_{2,1}$	0.38	-0.93	0.09	-1.87	0.16	0.22	4.38	0.15	11.24	0.25
$vo4c_{1,0}$	0.72	-2.44	0.06	-6.27	0.09	0.46	-3.65	0.10	-9.37	0.17
$vo4c_{1,1}$	0.10	26.52	0.45	65.74	0.71	0.15	21.83	0.32	54.53	0.52
$vo4c_{1,2}$	0.06	4.65	0.44	10.14	0.71	0.05	-30.51	0.67	-73.60	1.02
$vo4c_{1,3}$	0.12	-8.93	0.21	-20.05	0.31	0.33	-0.21	0.02	-0.55	0.03
$vo4c_{2,0}$	0.34	-3.94	0.09	-10.38	0.14	0.37	-4.07	0.09	-10.08	0.14
$vo4c_{2,1}$	0.05	26.95	0.64	71.19	1.06	0.10	12.17	0.45	26.42	0.75
$vo4c_{2,2}$	0.08	25.00	0.47	59.86	0.74	0.35	6.16	0.12	16.29	0.21
$vo4c_{2,3}$	0.53	-3.64	0.07	-8.72	0.11	0.17	-11.43	0.21	-28.22	0.32
$vo10c_{1,0}$	0.09	-9.36	0.25	-20.90	0.39	0.06	-13.16	0.35	-33.67	0.57
$vo10c_{1,1}$	0.11	-7.04	0.21	-17.37	0.33	0.16	-1.44	0.08	-3.46	0.12
$vo10c_{1,2}$	0.01	38.24	2.05	78.06	3.03	0.19	-2.50	0.11	-6.17	0.16
$vo10c_{1,3}$	0.04	9.18	0.60	19.49	0.89	0.11	3.94	0.18	9.44	0.30
$vo10c_{1,4}$	0.04	15.94	0.60	36.59	0.93	0.14	4.29	0.22	10.28	0.36
$vo10c_{1,5}$	0.01	77.14	1.92	191.42	3.15	0.03	20.53	1.12	51.95	1.82
$vo10c_{1,6}$	0.06	11.32	0.48	29.16	0.74	0.03	43.61	1.14	107.38	1.79
$vo10c_{1,7}$	0.04	12.56	0.58	35.02	0.89	0.04	14.19	0.69	37.93	1.11
$vo10c_{1,8}$	0.23	-3.61	0.14	-8.60	0.22	0.08	-13.06	0.40	-33.36	0.61
$vo10c_{1,9}$	0.36	-3.23	0.08	-8.41	0.14	0.17	-6.38	0.16	-15.39	0.24
$vo10c_{2,0}$	0.16	-3.11	0.12	-8.84	0.21	0.13	-6.99	0.17	-17.20	0.29
$vo10c_{2,1}$	0.15	-3.54	0.15	-9.38	0.26	0.04	-17.82	0.55	-42.85	0.87
$vo10c_{2,2}$	0.01	31.22	1.38	87.60	2.26	0.10	-1.20	0.24	-3.86	0.41
$vo10c_{2,3}$	0.21	-2.48	0.15	-6.05	0.24	0.13	8.65	0.21	20.82	0.35
$vo10c_{2,4}$	0.03	24.47	0.77	62.65	1.22	0.13	7.09	0.27	16.37	0.43
$vo10c_{2,5}$	0.02	49.83	1.26	124.94	1.97	0.11	10.84	0.31	27.37	0.50
$vo10c_{2,6}$	0.09	5.15	0.35	12.00	0.55	0.10	2.15	0.33	5.93	0.52
$vo10c_{2,7}$	0.08	2.82	0.34	5.35	0.50	0.12	-4.94	0.23	-11.60	0.38
$vo10c_{2,8}$	0.13	-4.79	0.21	-11.83	0.32	0.04	-12.36	0.44	-28.66	0.73
$vo10c_{2,9}$	0.11	-6.73	0.21	-14.84	0.33	0.10	-5.72	0.19	-14.41	0.28
$vnstd^*_{1,Y}$	-0.01	0.000	0.000	0.000	0.000	0.01	-0.001	0.000	0.000	0.000
$vnstd^*_{2,Y}$	0.03	0.000	0.000	-0.001	0.000	0.01	0.000	0.000	0.000	0.000
$vn_{1,Y}$	50.05	0.00	0.01	-0.01	0.02	50.06	-0.01	0.01	0.00	0.01
$vn_{2,Y}$	50.08	0.00	0.00	0.00	0.00	49.92	0.00	0.00	0.00	0.00

5.2.2. Resumen de los rangos de los estimadores de sesgos de los métodos para el tratamiento de la no respuesta total y parcial, excluyendo las variables $vnstd_1$ y $vnstd_2$

En los métodos para la no respuesta total, se verificó un resultado esperado, tanto en *Pobl* como en *PoblM* los sesgos $\hat{B}$ y valores de L_c aumentaron cuando se incrementó la tasa de no respuesta. De acuerdo a las longitudes de los rangos de variación de $\hat{B}$ y L_c las estimaciones fueron mejores para las muestras de *PoblM*, en comparación con las de *Pobl*. El desempeño de los métodos de submuestreo de los no informantes y clases de ponderación fueron los peores de los cinco evaluados en *Pobl* y el que mejor funcionó fue el de ponderación simple. En las muestras de *PoblM* claramente el de peor desempeño fue el método de submuestreo de los no informantes y el que generó los menores valores de $\hat{B}$ y L_c fue el método de ponderación simple (Ver Tabla 5.11).

Tabla 5.11: Resumen de rangos de los métodos para la no respuesta total

		Pobl				*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
Mét.		$\hat{B}$	L_c	$\hat{B}$	L_c	$\hat{B}$	L_c	$\hat{B}$	L_c
Submuestr. no inf.	mín	-18.07	0.07	-40.39	0.31	-13.73	0.07	-32.80	0.10
	máx	-7.50	1.06	-21.01	4.90	-10.79	0.50	-27.53	0.76
Pond.Simp.	mín	-8.05	0.02	-17.72	0.03	-2.29	0.02	-5.31	0.03
	máx	5.78	1.19	14.49	2.03	1.53	0.57	3.49	0.97
Clases pond.	mín	-17.17	0.04	-25.72	0.05	-7.08	0.03	-10.47	0.05
	máx	23.74	2.07	39.70	2.87	5.80	1.28	10.52	1.76
Est.Razón	mín	-6.12	0.03	-12.73	0.05	1.68	0.04	3.91	0.06
	máx	8.22	1.19	20.55	2.08	3.91	0.57	9.27	1.07
Est.Regres.	mín	-8.59	0.02	-18.18	0.03	-2.30	0.01	-5.31	0.03
	máx	5.86	1.18	14.43	1.94	0.98	0.57	1.75	0.96

En los métodos para la no respuesta parcial, los mejores resultados en términos de los sesgos $\hat{B}$ se ubicaron en las muestras de *PoblM*. También se observó que a medida que se incrementó la tasa de no respuesta de 10 % a 30 % también aumentaron los sesgos de estimación. El método de imputación múltiple fue el de peor desempeño en ambos conjuntos de muestras, sin embargo, en *Pobl* el funcionamiento de método de regresión fue extremadamente deficiente. Los métodos de casos disponibles y *hot-deck* tuvieron los mejores desempeños de los cuatro métodos que son comparables para los conjuntos de muestras de *Pobl* y *PoblM* (Ver Tabla 5.12).

Tabla 5.12: Resumen de rangos de los métodos para la no respuesta parcial

		Pobl				*PoblM*			
		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$		$T_{nr} = 10\,\%$		$T_{nr} = 30\,\%$	
Mét.		$\hat{B}$	L_c	$\hat{B}$	L_c	$\hat{B}$	L_c	$\hat{B}$	L_c
C.completos	mín	-8.44	0.02	-17.66	0.03	-1.12	0.02	-1.56	0.03
	máx	5.87	1.18	14.46	1.94	1.51	0.52	3.48	0.94
C.disponibles	mín	-3.78	0.01	-7.27	0.02	-0.33	0.01	-0.59	0.02
	máx	2.63	0.77	5.80	1.31	0.70	0.38	1.39	0.62
Hot-deck	mín	-3.97	0.02	-6.96	0.03	-0.60	0.02	-0.85	0.03
	máx	2.66	1.08	5.44	1.65	0.69	0.50	1.39	0.82
Imp.regres.	mín	-16.76	0.02	-45.25	0.03	–	–	–	–
	máx	166.22	2.48	504.41	4.94	–	–	–	–
Imp.múlt.	mín	-9.36	0.01	-20.90	0.02	-30.51	0.01	-73.60	0.01
	máx	77.14	11.46	191.42	11.14	43.61	11.76	107.38	11.79

La razón por la cual se obtuvieron mejores estimaciones en las muestras de *PoblM* podría ser atribuible al hecho de que las variables que las compusieron tenían correlaciones casi nulas y por lo tanto la estimación de los valores de una variable en particular no dependía de las otras variables restantes. En las variables continuas de *PoblM*, generadas con una distribución normal multivariada de cuatro variables, los sesgos de estimación fueron muy similares a las continuas de *Pobl*, también generadas con una distribución normal multivariada subyacente de diez variables.

Conclusiones

La manera de presentar los resultados en este estudio sirvió para que se observara el impacto de las estimación con respecto al valor poblacional a estimar (proporciones en los niveles de respuesta de la variables ordinales y medias en las variables continuas). El mecanismo de omisión empleado fue de naturaleza no lineal y además sesgado hacia un conjunto de casos definido por lo que se esperaba que los métodos más simples generarían los sesgos mayores, situación que no se observó.

El primer hallazgo que fue aplicable a todos los métodos de no respuesta total y parcial, excepto el de submuestreo de los no informantes, fue que a medida que se incrementaron los valores a estimar dentro de una misma variable, también surgieron estimaciones con un sesgo mayor, en valor absoluto, en comparación con otra con menos valores a estimar. Otro resultado general para todos los métodos de no respuesta total y parcial es que la magnitud de L_c aumentó en los niveles de medición con las proporciones más bajas y en las variables continuas siempre se encontraron los menores valores, sin pérdida de generalidad, si $vo10c_{1,j} > vo10c_{1,k}$ entonces $L_{c,j} < L_{c,k}$. La presentación de los sesgos relativos $vnstd_1$ y $vnstd_2$ los hubiera mostrado desproporcionados y no habrían reflejado de ninguna manera el desempeño habitual de los métodos analizados ya que cuando se pretende estimar parámetros que son iguales o cercanos a cero por pequeña que sea la discrepancia de estimación la comparación con el valor verdadero indicará una enorme diferencia. La alternativa de una transformación lineal aplicada a dichas variables no hubiera resuelto el problema ya que este tipo de transformaciones modifica la magnitud de los sesgos pero la comparación

con el valor poblacional no se afecta. Variables de este no se hallan constantemente en casos prácticos pero, por ejemplo, en mediciones de niveles de plomo en sangre sus valores son muy pequeños. Por lo que la forma de evaluar sus estimaciones no debe presentarse en relación al parámetros poblacional.

El método de regresión para la no respuesta parcial generó resultados tan imprecisos para las variables ordinales debido al número de valores a estimar y además que el método ajusta un modelo de regresión (logístico o multinomial) donde algunas de las variables predictivas (que considera a cada uno de los niveles de medición también como variables predictivas) se eliminan del modelo, pero lo que se pudo observar es que en las variables continuas y correlacionadas significativamente funcionó correctamente.

El método de imputación múltiple también generó sesgos muy altos en la variables ordinales pero en la estimación de los parámetros de las variables continuas $vnstd_1$ y $vnstd_2$ tuvo un desempeño correcto, ya que en los otros métodos tanto de no respuesta total y parcial su principal inconveniente fue su incapacidad para generar sesgos de estimación menores al 10 %.

El buen funcionamiento de los métodos ponderación simple, caos totales, casos disponibles y *hot-deck* se debió a que el mecanismo de omisión seleccionó casos que aunque se excluyeron no afectaron las estimaciones de las proporciones medias ya que con mayor probabilidad se extrajeron los casos cuyos valores de vn_2 eran mayores, aunque sus estimaciones de los parámetros de $vnstd_1$ y $vnstd_2$ fueron muy imprecisas ya que incluso los sesgo calculados resultaron mayores al valor a estimar. El método de submuestreo de los no informantes generó sesgos considerables debido a que el mecanismo de omisión no coincidió con el esquema de muestreo de los no informantes. La literatura indica que el método de clases de ponderación mejora las estimaciones de la ponderación simple pero solamente si la clasificación genera conjuntos mucho más homogéneos; para este caso en particular la clasificación dependientes de la variable auxiliar vn_2, empeoró las estimaciones de los valores en comparación con la ponderación simple, por lo que sería adecuado evaluar las homogeneidad de las agrupaciones para considerar si es conveniente la clasificación que se pretende utilizar.

Los argumentos anteriores coinciden con la literatura revisada, que señala que para emplear cualquier método de imputación se debe tratar de modelar el mecanismo que genera la no respuesta —tal recomendación es aplicable

especialmente al submuestreo de los no informantes ya que implica un esfuerzo adicional de levantamiento de datos que puede empeorar las estimaciones en lugar de mejorarlas—. Las variables ordinales podrán ser imputadas adecuadamente si se reduce el número de valores a estimar y en el caso de las variables continuas se podrá aplicar los métodos de regresión e imputación múltiple si las variables a imputar se correlacionan de manera significativa.

El método de estimadores de regresión para el tratamiento de la no respuesta total, y los de regresión y de imputación múltiple para la no respuesta parcial son los más recomendables para ser empleados para la estimación de parámetros de variables continuas sin importar las magnitudes sus los valores poblacionales.

APÉNDICE A

Resultados auxiliares

A.1. Teorema de Bayes

El Teorema de Bayes es una herramienta estándar de la estadística matemática. Se emplea para relacionar distribuciones condicionales con distribuciones marginales. Si A y B son dos variables aleatorias con distribución condicional $Pr(A, B)$, entonces

$$Pr(A, B) = Pr(A|B)Pr(B) = Pr(A)Pr(B|A).$$

El Teorema de Bayes generalmente se enuncia como

$$P(A|B) = \frac{Pr(A)Pr(B|A)}{Pr(B)}.$$

Si existe una variable adicional, C, a la cual están condicionados

$$Pr(A, B|C) = Pr(A|B, C)Pr(B|C) = Pr(A|C)Pr(B|A, C)$$

y

$$P(A|B, C) = \frac{Pr(A|C)Pr(B|A, C)}{Pr(B|C)}.$$

Las distribuciones marginal y conjunta están relacionadas por

$$Pr(B) = \int Pr(A, B)dA \text{ y } Pr(B|C) = \int Pr(A, B|C)dA$$

donde $\int[\cdot]dA$ representa integración o suma sobre todos los valores posibles de A dependiendo si A es continuo o discreto. También se tiene que $\int Pr(A)dA = 1$ y $\int Pr(A|B)dA = 1$ para todos los posibles valores de B debido a que todas las distribuciones tienen que sumar uno.

A.2. Distribuciones subyacentes de variables ordinales

Las observaciones de una variable ordinal representan las respuestas a un conjunto de categorías ordenadas, como en una escala Likert de cuatro categorías. Si se asume que una persona que selecciona una categoría tiene más de una característica que si hubiera seleccionado una categoría menor pero se desconoce la magnitud de la diferencia. Las variables ordinales no son variables continuas y no deberían ser tratados como tales. Es una práctica común tratar las puntuaciones $1, 2, 3, \ldots$ asignados a categorías como si tuvieran propiedades métricas. Las variables ordinales no tienen origen o unidades de medición. La única información con la que se cuenta son las frecuencias de cada celda de una tabla de contingencia.

Para cada variable ordinal z se asume que existe una variable continua subyacente z^*. Ésta variable z^* representa la actitud subyacente a la respuesta a z y se supone que tiene un rango de $-\infty$ a $+\infty$. La variable subyacente z^* puede emplearse en análisis estadísticos concebidos para variables continuas en lugar de la observada z. La variable subyacente asigna una métrica a la variable ordinal.

Si la variable z tiene m categorías etiquetadas como $1, 2, \ldots, m$, la conexión entre z y z^* es

$$z = i \text{ si y sólo si } \tau_{i-1} < z^* < \tau_i, \quad i = 1, 2, \ldots, m,$$

donde

$$-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{m-1} < \tau_m = +\infty,$$

son conocidos como valores de los umbrales. Con m categorías implica que existan $m - 1$ parámetros de los umbrales $\tau_1, \tau_2, \ldots, \tau_{m-1}$.

Debido a que sólo se cuenta con información del orden de la variable z, la distribución de z^* esta determinada solamente por una transformación monótona. En un principio, se puede seleccionar cualquier distribución continua para z^*. Sin embargo, cualquier variable continua con una función de

densidad y de distribución se le puede aplicar una transformación monótona para convertirla en una distribución normal. Por lo que es conveniente escoger una distribución normal estándar con función densidad $\phi(u)$ y una función de distribución $\Phi(u)$ para z^*. Entonces la probabilidad de una respuesta en la categoría i es

$$\pi_i = Pr[z = i] = Pr[\tau_{i-1} < z^* < \tau_i] = \int_{\tau_{i-1}}^{\tau_i} \phi(u)du = \Phi(\tau_i) - \Phi(\tau_{i-1}), \quad \text{(A.1)}$$

por lo que

$$\tau_i = \Phi^{-1}(\pi_1 + \pi_2 + \cdots + \pi_i), \quad i = 1, \ldots, m-1, \quad \text{(A.2)}$$

donde Φ^{-1} es la función inversa de la distribución normal estándar. La cantidad $(\pi_1 + \pi_2 + \cdots + \pi_i)$ es la probabilidad de una respuesta en la categoría i ó en una menor.
Las probabilidades π_i con cantidades poblacionales desconocidas. En la práctica, π_i puede ser estimada consistentemente con el porcentaje p_i correspondiente a las respuestas en la categoría i. Las estimaciones de los umbrales se puede obtener como

$$\hat{\tau}_i = \Phi^{-1}(p_1 + p_2 + \cdots + p_i), \quad i = 1, \ldots, m-1. \quad \text{(A.3)}$$

La cantidad $(p_1 + p_2 + \cdots + p_i)$ es la proporción de casos en la muestra que respondieron la categoría i o una menor. La ecuación A.3 de hecho es el estimador de máxima verosimilitud de τ_i basado en una muestra de datos marginal y univariada. Pero el modelo es saturado; existen $m-1$ parámetros τ_i y $m-1$ proporciones muestrales independientes p_i. El ajuste es perfecto puesto que

$$\hat{\pi}_i = \Phi(\hat{\tau}_i) - \Phi(\hat{\tau}_{i-1}) = p_i.$$

Tabla A.1: Estimación de umbrales

Categoría i	Porcentaje	Porcentaje acumulado	Umbral τ_i
1	17	17	−0.954
2	42	59	0.228
3	33	92	1.405
4	8	100	

La estimación de umbrales se ilustra con la Tabla A.1. Las categorías 1, 2, 3, 4 tuvieron 17, 42, 33 y 8 %, respectivamente, de selección. Los valores acumulados se presentan en la tercera columna. El primer umbral se localiza en punto en donde el área bajo la curva normal a la izquierda del umbral corresponde al 17 %. El segundo umbral se localiza donde el área bajo la curva normal a la izquierda del umbral es 59 %. El tercer umbral se localiza en donde el área bajo la curva normal a la izquierda del umbral es 92 %. Por lo tanto $\hat{\tau}_1 = -0.954$, $\hat{\tau}_2 = 0.228$, $\hat{\tau}_3 = 1.405$.

APÉNDICE B

Rutinas de cómputo

En este capítulo se presentan el código en el lenguaje R para la generación de la base de datos para las dos poblaciones analizadas, la generación de los mecanismos de omisión total y parcial y el cálculo de las estimaciones de los parámetros de las variables para el tratamiento de la no respuesta total y parcial.

B.1. Generación de las poblaciones

```
## Se crea una matriz aleatoria de correlación, positiva definida
## para la generación de una distribución normal multivariada
Posdef <- function(n)
{Z <- matrix(ncol=n, rnorm(n^2))
Z <- t(Z)%*%Z
z <- 1/sqrt(diag(Z))
Z <- diag(z)%*%Z%*%diag(z)
return(Z)}

##Generación de la Población  N= 10000 con distribución normal
##multivariada con vectores  de medias (0,0,0,0,0,0,0,0,50,50)
##y desviaciones estándares (1,1,1,1,1,1,1,1,9.49,9.49)
Pobl <- rmvnorm(10000, mean=c(0,0,0,0,0,0,0,0,50,50),cov=Posdef(10),
sd=c(1,1,1,1,1,1,1,1,9.49,9.49))

##Se generaron vectores con números aleatorios de una distribución
```

```
##normal que se emplearon para categorizar las primeras
##seis variables.
aleN <- cbind(
{c(rnorm(1,mean=0, sd=1))},
{sort(c(rnorm(1,mean=0, sd=1)))},
{sort(c(rnorm(3,mean=0, sd=1)))},
{sort(c(rnorm(3,mean=0, sd=1)))},
{sort(c(rnorm(9,mean=0, sd=1)))},
{sort(c(rnorm(9,mean=0, sd=1)))} )

##Este procedimiento  categoriza las seis primeras variables
##en 2, 4 y 10 niveles de medición.
for (i in 1:6)
for (j in 1:length(Pobl[,1]))
if (i==1|i==2)
 {if (aleN[1,i]>=Pobl[j,i])  Pobl[j,i] <- 1
else if (aleN[1,i]<Pobl[j,i])  Pobl[j,i] <- 2}
else if (i==3|i==4)
{if (aleN[1,i]>=Pobl[j,i])    Pobl[j,i] <- 1
else if ((aleN[1,i]<Pobl[j,i])&(aleN[2,i]>=Pobl[j,i]))Pobl[j,i] <- 2
else if ((aleN[2,i]<Pobl[j,i])&(aleN[3,i]>=Pobl[j,i]))Pobl[j,i] <- 3
else if (aleN[3,i]<Pobl[j,i])  Pobl[j,i] <- 4}

else if (i==5|i==6)
{if (aleN[1,i]>=Pobl[j,i])   Pobl[j,i] <- 1
else if ((aleN[1,i]<Pobl[j,i])&(aleN[2,i]>=Pobl[j,i]))Pobl[j,i] <- 2
else if ((aleN[2,i]<Pobl[j,i])&(aleN[3,i]>=Pobl[j,i]))Pobl[j,i] <- 3
else if ((aleN[3,i]<Pobl[j,i])&(aleN[4,i]>=Pobl[j,i]))Pobl[j,i] <- 4
else if ((aleN[4,i]<Pobl[j,i])&(aleN[5,i]>=Pobl[j,i]))Pobl[j,i] <- 5
else if ((aleN[5,i]<Pobl[j,i])&(aleN[6,i]>=Pobl[j,i]))Pobl[j,i] <- 6
else if ((aleN[6,i]<Pobl[j,i])&(aleN[7,i]>=Pobl[j,i]))Pobl[j,i] <- 7
else if ((aleN[7,i]<Pobl[j,i])&(aleN[8,i]>=Pobl[j,i]))Pobl[j,i] <- 8
else if ((aleN[8,i]<Pobl[j,i])&(aleN[9,i]>=Pobl[j,i]))Pobl[j,i] <- 9
else if (aleN[9,i]<Pobl[j,i])  Pobl[j,i] <- 10}

##Generacion de Población con seis variables con distribución
##multinomial y cuatro variables con distribución normal
##multivariada con vectores de medias (0,0,50,50) y de
##desviaciones estándares (1,1,9.49,9.49)

##Este procedimiento genera aleatoriamente las probabilidades para
##las distribuciones bi(multi)nomiales marginales.
rm1 <- runif(2,0,1); rm2 <- runif(2,0,1)
rm3 <- runif(4,0,1); rm4 <- runif(4,0,1)
rm5 <- runif(10,0,1); rm6 <- runif(10,0,1)
```

```
rm1 <- rm1/sum(rm1); rm2 <- rm2/sum(rm2)
rm3 <- rm3/sum(rm3); rm4 <- rm4/sum(rm4)
rm5 <- rm5/sum(rm5); rm6 <- rm6/sum(rm6)

##Generación de la base de datos con distribución multinomial
##y normal multivariada
PoblM <-cbind(
sample(1:2,10000,prob=rm1,T),
sample(1:2,10000,prob=rm2,T),
sample(1:4,10000,prob=rm3,T),
sample(1:4,10000,prob=rm4,T),
sample(1:10,10000,prob=rm5,T),
sample(1:10,10000,prob=rm6,T),
rmvnorm(10000, mean=c(0,0,50,50),cov=Posdef(4), sd=c(1,1,9.49,9.49)))
```

B.2. Generación de los mecanismos de omisión al azar (MAR) total y parcial

```
##Generación de todos los posibles patrones de omisión de respuesta
##en nueve variables (2^9-1=511).
patM<-expand.grid(0:1, 0:1, 0:1, 0:1, 0:1, 0:1,0:1,0:1,0:1,0)
patM[patM==1]<-NA

##El renglón con cero omisiones se elimina.
patM <- remove.row(patM, 1, 1)

##Generación del patron de no respuesta total
patMT<-c(rep(NA,10))

##Simulación de mecanismo de omisión, tasa del 10%
##ln(0.90/(1-0.90))=2.2
##Simulación de mecanismo de omisión, tasa del 30%
##ln(0.70/(1-0.70))=0.85
##Mecanismo de omisión total.
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patMT

##Simulación de mecanismo de omisión, tasa del 10%
##ln(0.90/(1-0.90))=2.2
##Simulación de mecanismo de omisión, tasa del 30%
```

```
##ln(0.70/(1-0.70))=0.85
##Mecanismo de omisión parcial.
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patM[sample(1:511,1),]
```

B.3. Métodos para el tratamiento de no respuesta total

```
##Ponderación simple
##Distribución normal multivariada de las variables
## Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))

dos.time({
for (tau in 1:10){
##Proceso inicial de iteraciones
for( k in 1:100){
##Selección de una muestra de la población
##la variable Pobl indica la población a muestrear.
sampT<-Pobl[sample(1:10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10%
##ln(0.90/(1-0.90))=2.2
##Simulación de mecanismo de omisión, tasa del 30%
##ln(0.70/(1-0.70))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patMT

##Eliminación de los casos con omisiones
samp <- na.omit(samp)

## Calculo de la distribución de porcentaje
```

```
##en los niveles de medición en la muestra sin omisión
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]),mean(sampT[,8]),
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión
parmE <-c(c(tabulate(samp[,1],2),tabulate(samp[,2],2),
tabulate(samp[,3],4),tabulate(samp[,4],4),
tabulate(samp[,5],10),tabulate(samp[,6],10))/nrow(samp),
mean(samp[,7]),mean(samp[,8]),
mean(samp[,9]),mean(samp[,10]))

##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}} } }})

##Submuestreo de los no informantes
##Distribución normal multivariada de las variables
## Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))
dos.time({
for (tau in 1:10){
##Proceso inicial de iteraciones
for( k in 1:100){
##Selección de una muestra de la población
sampT<-PoblM[sample(1:10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10%
##ln(0.90/(1-0.90))=2.2
##Simulación de mecanismo de omisión, tasa del 30%
##ln(0.70/(1-0.70))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
```

```
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patMT

##Selección de no informantes con muestreo alet. simple del 10%
sampNI <- matrix(ncol=10,sampT[is.na(samp)])
sampNI <- sampNI[sample(1:nrow(sampNI),ceiling(nrow(sampNI)*.10)),]

##Eliminación de los casos con omisiones
samp <- na.omit(samp)

## Calculo de la distribución de porcentaje
##en los niveles de medición muestra sin omisiones
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]),mean(sampT[,8]),
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición muestra con omisiones
parmE <-c(c(tabulate(samp[,1],2),tabulate(samp[,2],2),
tabulate(samp[,3],4),tabulate(samp[,4],4),
tabulate(samp[,5],10),tabulate(samp[,6],10))/nrow(samp),
mean(samp[,7]),mean(samp[,8]),
mean(samp[,9]),mean(samp[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición de los no informantes
parmENi <-c(c(tabulate(sampNI[,1],2),tabulate(sampNI[,2],2),
tabulate(sampNI[,3],4),tabulate(sampNI[,4],4),
tabulate(sampNI[,5],10),tabulate(sampNI[,6],10))/nrow(sampNI),
mean(sampNI[,7]),mean(sampNI[,8]),
mean(sampNI[,9]),mean(sampNI[,10]))

##Cálculo de las estimaciones
parmE<-((parmE*nrow(samp)+parmENi*nrow(sampNI))/nrow(sampT))

##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
```

```
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}} } }})

##Clases de ponderación
##Distribución normal multivariada de las variables
## Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))
tr_0
dos.time({
for (tau in 1:10){
for (tau1 in 1:10){
##Proceso inicial de iteraciones
for( k in 1:10){
##Selección de una muestra de la población
sampT<-PoblM[sample(1:10000,660,replace=F),]
samp <- sampT
tr_tr+1
##Simulación de mecanismo de omisión, tasa del 10%
##ln(0.90/(1-0.90))=2.2
##Simulación de mecanismo de omisión, tasa del 30%
##ln(0.70/(1-0.70))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patMT

##Eliminación de los casos con omisiones
samp <- na.omit(samp)

##Segmentación en cuatro clases de ponderación
##para todas la variable sin omisiones vnstd1
norm<- (c(-.674,0.0,0.674)*stdev(sampT[,10])+mean(sampT[,10]))
sampc1<-matrix(ncol=10,samp[samp[,10]<norm[1]])
sampc2<-matrix(ncol=10,samp[samp[,10]>=norm[1]&samp[,10]<norm[2]])
sampc3<-matrix(ncol=10,samp[samp[,10]>=norm[2]&samp[,10]<norm[3]])
sampc4<-matrix(ncol=10,samp[samp[,10]>=norm[3]])

## Calculo de la distribución de porcentaje
##en los niveles de medición en la muestra sin omisión
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
```

```
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]),mean(sampT[,8]),
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión clase 1
parmEc1 <-c(c(tabulate(sampc1[,1],2),tabulate(sampc1[,2],2),
tabulate(sampc1[,3],4),tabulate(sampc1[,4],4),
tabulate(sampc1[,5],10),tabulate(sampc1[,6],10))/nrow(sampc1),
mean(sampc1[,7]),mean(sampc1[,8]),
mean(sampc1[,9]),mean(sampc1[,10]))*(nrow(sampc1)/nrow(samp))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión clase 2
parmEc2 <-c(c(tabulate(sampc1[,1],2),tabulate(sampc1[,2],2),
tabulate(sampc1[,3],4),tabulate(sampc1[,4],4),
tabulate(sampc1[,5],10),tabulate(sampc1[,6],10))/nrow(sampc2),
mean(sampc1[,7]),mean(sampc1[,8]),
mean(sampc1[,9]),mean(sampc1[,10]))*(nrow(sampc2)/nrow(samp))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión clase 3
parmEc3 <-c(c(tabulate(sampc3[,1],2),tabulate(sampc3[,2],2),
tabulate(sampc3[,3],4),tabulate(sampc3[,4],4),
tabulate(sampc3[,5],10),tabulate(sampc3[,6],10))/nrow(sampc3),
mean(sampc3[,7]),mean(sampc3[,8]),
mean(sampc3[,9]),mean(sampc3[,10]))*(nrow(sampc3)/nrow(samp))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión clase 4
parmEc4 <-c(c(tabulate(sampc4[,1],2),tabulate(sampc4[,2],2),
tabulate(sampc4[,3],4),tabulate(sampc4[,4],4),
tabulate(sampc4[,5],10),tabulate(sampc4[,6],10))/nrow(sampc4),
mean(sampc4[,7]),mean(sampc4[,8]),
mean(sampc4[,9]),mean(sampc4[,10]))*(nrow(sampc4)/nrow(samp))

##Estimación del promedio de sesgos de estimación y su varianza
x<-((parmEc1+parmEc2+parmEc3+parmEc4)-parmT)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}}}}}})
```

```
##Procedimiento con estimadores de razón
##Distribución normal multivariada de las variables
## Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))
dos.time( {
for (tau in 1:10){
##Proceso inicial de iteraciones
for( k in 1:100){
##Selección de una muestra de la población
sampT<-PoblM[sample(1:10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10%
##ln(0.90/(1-0.90))=2.2
##Simulación de mecanismo de omisión, tasa del 30%
##ln(0.70/(1-0.70))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patMT

##Eliminación de los casos con omisiones
samp <- na.omit(samp)

## Calculo de la distribución de porcentaje
##en los niveles de medición muestra sin omisiones
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]) ,mean(sampT[,8]),
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición muestra con omisiones
parmE <-c(c(tabulate(samp[,1],2),tabulate(samp[,2],2),
tabulate(samp[,3],4),tabulate(samp[,4],4),
tabulate(samp[,5],10),tabulate(samp[,6],10))/nrow(samp),
mean(samp[,7]),mean(samp[,8]),
mean(samp[,9]),mean(samp[,10]))*(mean(sampT[,10])/mean(samp[,10]))
```

```
##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}} }} })

##Procedimiento con estimadores de regresión
##Distribución normal multivariada de las variables
## Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))
dos.time({
for (tau in 1:10){
##Proceso inicial de iteraciones
for( k in 1:100){
##Selección de una muestra de la población
sampT<-PoblM[sample(1:10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10%
##ln(0.90/(1-0.90))=2.2
##Simulación de mecanismo de omisión, tasa del 30%
##ln(0.70/(1-0.70))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))
/(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patMT

##Eliminación de los casos con omisiones
samp <- na.omit(samp)

## Calculo de la distribución de porcentaje
##en los niveles de medición muestra sin omisiones
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]) ,mean(sampT[,8]),
```

```
mean(sampT[,9]),mean(sampT[,10]))

##Calculo de los valores de beta para cada variable
##como variable auxiliar vn2
b1<-(sum((samp[,1]-mean(samp[,1]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b2<-(sum((samp[,2]-mean(samp[,2]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b3<-(sum((samp[,3]-mean(samp[,3]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b4<-(sum((samp[,4]-mean(samp[,4]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b5<-(sum((samp[,5]-mean(samp[,5]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b6<-(sum((samp[,6]-mean(samp[,6]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b7<-(sum((samp[,7]-mean(samp[,7]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b8<-(sum((samp[,8]-mean(samp[,8]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
b9<-(sum((samp[,9]-mean(samp[,9]))*(samp[,10]-mean(samp[,10])))/
sum((samp[,10]-mean(samp[,10]))^2) )
dif<-(mean(sampT[,10])-mean(samp[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición, muestra con omisiones
parmE <-c(c(tabulate(samp[,1],2)+(dif*b1),tabulate(samp[,2],2)+(dif*b2),
tabulate(samp[,3],4)+(dif*b3),tabulate(samp[,4],4)+(dif*b4),
tabulate(samp[,5],10)+(dif*b5),tabulate(samp[,6],10)+(dif*b6))/nrow(samp),
mean(samp[,7])+(dif*b7),mean(samp[,8])+(dif*b8),
mean(samp[,9])+(dif*b9),mean(samp[,10]))

##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}} }} })
```

B.4. Métodos para el tratamiento de no respuesta parcial

```
##Casos completos
##Distribución normal multivariada de las variables
##Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))

dos.time({
for(tau in seq(0,990,10)){
##Proceso inicial de iteraciones
for( k in (tau+1):(tau+10)){
##Selección de una muestra de la población
sampT<-PoblM[sample(1:10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10% ln((1-0.10)/.10))=2.2
##Simulación de mecanismo de omisión, tasa del 30% ln((1-0.30)/.30))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))
/(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patMT[sample(1:511,1),]


##Eliminación de los casos con omisiones
samp <- na.omit(samp)

## Calculo de la distribución de porcentaje
##en los niveles de medición en la muestra sin omisión
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]),mean(sampT[,8]),
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión
parmE <-c(c(tabulate(samp[,1],2),tabulate(samp[,2],2),
tabulate(samp[,3],4),tabulate(samp[,4],4),
```

```
tabulate(samp[,5],10),tabulate(samp[,6],10))/nrow(samp),
mean(samp[,7]),mean(samp[,8]),
mean(samp[,9]),mean(samp[,10]))

##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
print(k)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}}}}})

##Casos completos
##Distribución normal multivariada de las variables
##Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))

dos.time({
##Proceso inicial de iteraciones
for( k in 901:1000){
##Selección de una muestra de la población
sampT<-Pobl[sample(10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10% ln((1-0.10)/.10))=2.2
##Simulación de mecanismo de omisión, tasa del 30% ln((1-0.30)/.30))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-2.2)))/
(1+exp(((samp[i1,10]-m0)/s0-2.2)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patM[sample(511,1),]


## Calculo de la distribución de porcentaje
##en los niveles de medición en la muestra sin omisión
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]),mean(sampT[,8]),
```

```
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión
parmE <-c(tabulate(samp[,1],2)/nrow(na.omit(samp[,1])),
tabulate(samp[,2],2)/nrow(na.omit(samp[,2])),
tabulate(samp[,3],4)/nrow(na.omit(samp[,3])),
tabulate(samp[,4],4)/nrow(na.omit(samp[,4])),
tabulate(samp[,5],10)/nrow(na.omit(samp[,5])),
tabulate(samp[,6],10)/nrow(na.omit(samp[,6])),
mean(samp[,7], na.rm=T),mean(samp[,8], na.rm=T),
mean(samp[,9], na.rm=T),mean(samp[,10], na.rm=T))

##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
print(k)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}} } })

##Hot-Deck Aleatorio

##Función para hacer la imputación
##Hot-Deck univariada
##Indica las posición de los datos
##faltantes y las omisiones.
Separa <- function(muestra)
{pos.m <- is.na(muestra)
pos.r <- !pos.m
datos.r <- muestra[pos.r]
datos.m <- muestra[pos.m]
m <- length(datos.m)
list(datos.r=datos.r, m=m, datos.m=datos.m, pos.m=pos.m)}

##Selecciona el conjunto de datos que reemplazaran los datos
##faltantes.
HotDeck <- function(datos.r,m)
{donantes.HD <- sample(datos.r, m,replace=T)
donantes.HD}

##Hace la imputación
Imputación <- function(muestra, donantes, pos.m)
```

```
{muestra[pos.m] <- donantes
muestra}



x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))

dos.time({
for( tau in seq(0,990,10)) {
##Proceso inicial de iteraciones
for( k in (tau+1):(tau+10)) {
##Selección de una muestra de la población
sampT<-PoblM[sample(1:10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10% ln((1-0.10)/.10))=2.2
##Simulación de mecanismo de omisión, tasa del 30% ln((1-0.30)/.30))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660)
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patM[sample(1:511,1),]


H1 <- Separa(samp[,1])
deck1 <- HotDeck(H1$datos.r,H1$m)
samp[,1] <- Imputación(samp[,1],deck1,H1$pos.m)

H2 <- Separa(samp[,2])
deck2 <- HotDeck(H2$datos.r,H2$m)
samp[,2] <- Imputación(samp[,2],deck2,H2$pos.m)

H3 <- Separa(samp[,3])
deck3 <- HotDeck(H3$datos.r,H3$m)
samp[,3] <- Imputación(samp[,3],deck3,H3$pos.m)

H4 <- Separa(samp[,4])
deck4 <- HotDeck(H4$datos.r,H4$m)
samp[,4] <- Imputación(samp[,4],deck4,H4$pos.m)

H5 <- Separa(samp[,5])
```

```
deck5 <- HotDeck(H5$datos.r,H5$m)
samp[,5] <- Imputación(samp[,5],deck5,H5$pos.m)

H6 <- Separa(samp[,6])
deck6 <- HotDeck(H6$datos.r,H6$m)
samp[,6] <- Imputación(samp[,6],deck6,H6$pos.m)

H7 <- Separa(samp[,7])
deck7 <- HotDeck(H7$datos.r,H7$m)
samp[,7] <- Imputación(samp[,7],deck7,H7$pos.m)

H8 <- Separa(samp[,8])
deck8 <- HotDeck(H8$datos.r,H8$m)
samp[,8] <- Imputación(samp[,8],deck8,H8$pos.m)

H9 <- Separa(samp[,9])
deck9 <- HotDeck(H9$datos.r,H9$m)
samp[,9] <- Imputación(samp[,9],deck9,H9$pos.m)

## Calculo de la distribución de porcentaje
##en los niveles de medición en la muestra sin omisión
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]),mean(sampT[,8]),
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión
parmE <-c(c(tabulate(samp[,1],2),tabulate(samp[,2],2),
tabulate(samp[,3],4),tabulate(samp[,4],4),
tabulate(samp[,5],10),tabulate(samp[,6],10))/nrow(samp),
mean(samp[,7]),mean(samp[,8]),
mean(samp[,9]),mean(samp[,10]))

##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
print(k)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}}}  }})
```

```
##Imputación múltiple


##Funciones para redondeo a los
##valores observados en la variables
##ordinales
v.obsd <- function(n)
{ if (n<=1) n <-1
else if(n>=2) n<-2
else n<- round(n,0)
return(n) }
v.obs4c <- function(n)
{ if (n<=1) n <-1
else if(n>=4) n<-4
else n<- round(n,0)
return(n) }
v.obs10c <- function(n)
{ if (n<=1) n <-1
else if(n>=10) n<-10
else n<- round(n,0)
return(n) }
##Distribución normal multivariada de las variables
##Preparación de los valores iniciales
x_c(rep(0,36))
xbar_c(rep(0,36))
xbar1_c(rep(0,36))
s2_c(rep(0,36))

dos.time({
for(tau in seq(0,90,10)){
##Proceso inicial de iteraciones
for( k in (tau+1):(tau+10)){
##Selección de una muestra de la población
sampT<-PoblM[sample(1:10000,660,replace=F),]
samp <- sampT

##Simulación de mecanismo de omisión, tasa del 10% ln((1-0.10)/.10))=2.2
##Simulación de mecanismo de omisión, tasa del 30% ln((1-0.30)/.30))=0.85
m0<-mean(samp[,10])
s0<-stdev(samp[,10])
for (i1 in 1:660){
if ((exp(( (samp[i1,10]-m0)/s0-0.85)))/
(1+exp(((samp[i1,10]-m0)/s0-0.85)))>runif(1,0,1))
samp[i1,]<-samp[i1,]+patM[sample(1:511,1),]}
```

```
##Cálculo de valores para iniciar imputaciones.
w <-prelim.norm(samp)

##Encuentra estimadores de máxima verosimilitud de medias
##desviaciones estándares y correlaciones.
thetahat <- em.norm(w,showits=F)
getparam.norm(w,thetahat,corr=T)

##Realiza 50 iteraciones de aumento de datos con los
##con los estimadores de MV y se generan cinco
##conjuntos de datos distintos.


rngseed(7654321)
theta <- da.norm(w,thetahat,steps=50)
imp1 <- imp.norm(w,theta,samp)

theta <- da.norm(w,theta,steps=50)
imp2 <- imp.norm(w,theta,samp)

theta <- da.norm(w,theta,steps=50)
imp3 <- imp.norm(w,theta,samp)

theta <- da.norm(w,theta,steps=50)
imp4 <- imp.norm(w,theta,samp)

theta <- da.norm(w,theta,steps=50)
imp5 <- imp.norm(w,theta,samp)

##Promedio de las cinco muestras
##imputadas
samp<-(imp1+imp2+imp3+imp4+imp5)/5

## Redondeo a los valores observados
##de las variables ordinales.
for (j in 1:nrow(samp))
{samp[,1]<-v.obsd(samp[j,1])
samp[j,2]<-v.obsd(samp[j,2])
samp[j,3]<-v.obs4c(samp[j,3])
samp[j,4]<-v.obs4c(samp[j,4])
samp[j,5]<-v.obs10c(samp[j,5])
samp[j,6]<-v.obs10c(samp[j,6])}
```

```
## Calculo de la distribución de porcentaje
##en los niveles de medición en la muestra sin omisión
parmT <-c(c(tabulate(sampT[,1],2),tabulate(sampT[,2],2),
tabulate(sampT[,3],4),tabulate(sampT[,4],4),
tabulate(sampT[,5],10),tabulate(sampT[,6],10))/nrow(sampT),
mean(sampT[,7]),mean(sampT[,8]),
mean(sampT[,9]),mean(sampT[,10]))

## Calculo de la distribución de porcentaje
##en los niveles de medición con omisión
parmE <-c(c(tabulate(samp[,1],2),tabulate(samp[,2],2),
tabulate(samp[,3],4),tabulate(samp[,4],4),
tabulate(samp[,5],10),tabulate(samp[,6],10))/nrow(samp),
mean(samp[,7]),mean(samp[,8]),
mean(samp[,9]),mean(samp[,10]))

##Estimación del promedio de sesgos de estimación y su varianza
x<-(parmE-parmT)
print(k)
for (m in 1:36)
{xbar1[m]<-xbar[m];
  xbar[m]<-(xbar[m]+(x[m]-xbar[m])/k);
{if (k==1)   s2[m]<-0
else
s2[m]<-(((1-1/(k-1)))*s2[m]+k*(xbar[m]-xbar1[m])^2)}}} } })
```

Referencias

[1] Allison, P. D. (2001). *Missing Data*. Sage University Paper Series on Quantitative Application in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

[2] Chapman, D. W. (1976). A survey of non-response imputation procedures. *American Statistical Association Proceedings Section Survey Research Methods*. 1976(1), 245-251.

[3] Chapman, D. W. (1983). The impact of substitution on survey estimates. En: Madow, W. G., Olkin, I., y Rubin, D. B. (Ed). *Incomplete data in sample surveys, volume 2, Theory and bibliographies.* Nueva York: Academic Press.

[4] Cochran, W. G. (1980). *Técnicas de muestreo*. México: Compañía Editorial Continental.

[5] Emrich, L. J. (1986). Randomized response. *American Statistical Association Proceedings Section Survey Research Methods*. 1986(1), 88-92.

[6] Ford, B. L. (1976). Missing data procedure: a comparative study. *Proceeding of the American Statistical Association, Social Statistical Section*, 324-329.

[7] Groves, R. M. (2002). *Survey errors and survey costs.* Nueva York: John Wiley and Sons.

[8] de Heer, W. F. y Israëls, A. Z. (1992). Response trends in Europe. *American Statistical Association Proceedings Section Survey Research Methods*. 92-101.

[9] Horvitz, D. G., Shah, B. V. y Simmons, W. R. (1967). The unrelated question randomized response model. *Proceeding of the American Statistical Association, Social Statistical Section*, 367-371.

[10] Kalton, G. y Kasprzyk, D. (1982) Imputing for missing survey responses. *American Statistical Association Proceedings Section Survey Research Methods*, 22-33.

[11] Lehtonen, R. y Pahkinen, E. J. (1996). *Practical methods for design and analysis of complex surveys*. Chichester: John Wiley and Sons.

[12] Lohr, S. L. (2000). *Muestreo: Diseño y análisis*. México: International Thomson Editores. Alli

[13] Little, R. J. A. y Rubin, D. B. (1986). *Statistical analysis with missing values*. Nueva York: Wiley.

[14] McKnight, P. E., McKnight, K. M., Sinadi, S. y Figueredo A. J. (2007). *Missing data: a gentle introduction*. Nueva York: The Guilford Press.

[15] Oh, H. L. y Scheuren, F. J. (1983). Weighting adjustment for unit nonresponse. En *Incomplete data sample in surveys*. Vol 2. Editado por Madow, W. G., Olkin, I. y Rubin, D. B., 143-184. Nueva York: Academic Press.

[16] Rubin, D. B. (1978). Multiple imputation in sample surveys –a phenomenological Bayesian approach to nonresponse. *American Statistical Association Proceedings Section Survey Research Methods*. 20-34.

[17] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Nueva York: John Wiley and Sons.

[18] Sande, I. G. (1982). Imputation in surveys: coping with reality. *The American Statistician*. 36(3); 145-152.

[19] Särndal, C. E., Swensson, B., y Wretman, J. (1992). *Model Assisted Survey Sampling*. Nueva York: Springer-Verlag.

[20] Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Londres: Chapman and Hall.

[21] Schafer, J.L. (1999). *Multivariate Normal Multiple Imputation Algorithms Version (5/95) modified for S-PLUS 4.0 for Windows (1/98)*, disponible en http://www.stat.psu.edu/ jls/misoftwa.html.

[22] Schafer, J. L. y Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*. 33(4), 545-571.

[23] Stevens, S. S. (1946). On the theory of scales of measurement. *Science*. Vol. 103(2684); 677-680.

[24] van Buuren, S. y Oudshoorn C.G.M. (2000). *Multivariate imputation by chained equations: MICE V1.0*, disponible en http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm.

[25] Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*. 60, 66-69.