Encuesta de Microemprendimiento 2013

DISEÑO MUESTRAL

INSTITUTO NACIONAL DE ESTADÍSTICAS

Diciembre / 2013

DEPARTAMENTO DE INVESTIGACIÓN Y DESARROLLO
DEPARTAMENTO DE ESTUDIOS LABORALES

Encuesta de Microemprendimiento 2013 - Diseño Muestral

Instituto Nacional de Estadísticas.

Diciembre / 2013.

Jefe Departamento de Investigación y Desarrollo: Charles Durán A.

Jefe Departamento de Estudios Laborales: Alexandra Rueda

Jefe de proyecto III EME: David Niculcar

Coordinador Sección Estadísticas Sociales: Miguel Guerrero H.

Analista(s) Investigador(es): Denisse López A.

Juan Carlos Herrera O.

Miguel Guerrero H.

ÍNDICE

ÍNDICE DE CUADROS

ÍNDICE DE GRÁFICOS

INTRODUCCIÓN

El presente documento describe las características del diseño muestral así como la metodología de cálculo de los factores de expansión de la Tercera Encuesta de Microemprendimiento (III EME). En los primeros dos capítulos se describen los aspectos relacionados con el diseño muestral, exponiéndose los detalles e insumos necesarios para la determinación del tamaño muestral, las unidades muestrales, así como también las características del marco y unidades seleccionadas. El tercer capítulo está focalizado en el desarrollo y construcción del factor de expansión. En él se detallan las probabilidades de selección, el ponderador base (inverso de las probabilidades de selección), el ajuste por falta de respuesta y la calibración a stock de total de trabajadores independientes según macrozona y sexo[1], además se detalla el procedimiento de suavizamiento de los ponderadores. Finalmente, en el cuarto capítulo, se especifica la forma de utilizar las variables que definen el diseño muestral en la estimación y respectivos errores.

[1] Ver más detalles en capítulo 3.3

1. ANTECEDENTES DEL DISEÑO MUESTRAL

A continuación se exponen los objetivos del estudio, población objetivo, unidad de información y nivel de estimación, utilizados para definir la estrategia de muestreo.

1.1. Objetivo General

- Lograr, a través de la implementación de una encuesta a hogares, una caracterización de la heterogénea realidad de los microemprendimientos del país, sus dueños y trabajadores, y su evolución en el tiempo.

- Complementar la información en el tiempo que permita evaluar el desempeño empresarial y el emprendimiento en el país.

1.2. Objetivos Específicos

- Identificar y caracterizar la situación de formalidad bajo distintas dimensiones (Registros contables, inscripción en servicios de impuestos internos, declaración de impuestos, organización jurídica, generación de empleo formal e informal, etc.) y sus determinantes.
- Indagar acerca de la relación que tiene el negocio con el sistema financiero, a través del acceso y trabas al financiamiento, sus características y usos del mismo.
- Estudiar la motivación y las razones del surgimiento de los emprendimientos. Si éstos son motivados por necesidad, por oportunidad o bien, causados por situaciones del entorno.
- Identificar los obstáculos que dificultan el desarrollo de las unidades productivas, tales como las restricciones en materia de acceso a tecnología, capacitación, financiamiento, entre otros. Conocer la situación laboral actual del trabajador independiente, así como sus experiencias o fracasos anteriores como emprendedor.
- Conocer el nivel educacional con que cuentan los emprendedores, además de las áreas más importantes donde ha recibido capacitación en los últimos tres años.
- Realizar una recopilación de datos que permita comparar los resultados con estadísticas internacionales sobre industrias y emprendimiento.

1.3. Población Objetivo

El estudio está enfocado a las unidades productivas de menor tamaño, es decir, al emprendedor tradicional, que es por lo general informal y más precario, que puede ser captado mediante una encuesta a hogares, en contraposición de un emprendedor de alto impacto que puede ser captado por otras fuentes.

Debido a que no existe un consenso entre los especialistas en emprendimiento sobre una definición de quiénes son emprendedores, la Subsecretaría de Economía ha optado por definir como población objetivo a todos quienes sean "Trabajadores por Cuenta Propia" o "Empleadores", quienes forman el conjunto de trabajadores “Independientes” del país. Esto evita cometer un sesgo de selección al truncar la muestra sólo a una definición particular, sino que se da espacio a capturar a toda la gama de emprendedores.

En este contexto, la población objetivo son todos los trabajadores por cuenta propia y empleadores, denominados trabajadores independientes, que residen en viviendas particulares ocupadas del territorio nacional.

1.4. Unidad de información

La unidad de información es el trabajador por cuenta propia o el empleador que reside en la vivienda particular y que haya sido entrevistado en la Encuesta Nacional de Empleo, y clasificado en dicha categoría laboral.

1.5. Nivel de estimación

Se entiende por nivel de estimación aquellas desagregaciones geográficas o características sociodemográficas, para las cuales se desean obtener estimaciones con márgenes de error adecuados y buena cobertura geográfica.

La muestra de la III EME fue seleccionada aleatoriamente a fin de representar tantos las áreas urbanas y rurales de las 15 regiones del país, sin embargo el diseño muestral fue concebido con la finalidad de obtener estimaciones a nivel nacional, y por lo tanto para mayores desagregaciones no garantiza buenos márgenes de error.

2. DISEÑO MUESTRAL

La tercera versión de la Encuesta de Microemprendimiento, posee un diseño muestral bifásico, en que la primera fase corresponde a un muestreo probabilístico, estratificado y bietápico, donde las unidades primarias corresponden a manzanas en el área urbana y secciones en el área rural; mientras que las unidades de segunda etapa son las viviendas particulares. Las unidades primarias (manzanas en el área urbana y secciones en el área rural) fueron seleccionadas en forma proporcional al tamaño, mientras que las unidades de segunda etapa se seleccionaron de forma sistemática y con igual probabilidad. Así, las unidades seleccionadas y encuestadas en la Encuesta Nacional de Empleo (ENE) del período MAM[2] 2013 fueron utilizadas como marco de muestreo para la III EME, pues permitió identificar las viviendas donde residen trabajadores por cuenta propia y empleadores (según la clasificación en la ENE).

En la segunda fase, se clasificaron las viviendas en dos grupos, de acuerdo a si éstas contenían o no, en el período de referencia, al menos un trabajador por cuenta propia o empleador. Las viviendas que no poseían trabajadores independientes fueron descartadas, formando el marco de muestreo sólo aquellas viviendas con unidades elegibles (con trabajadores independientes). Posteriormente, se seleccionaron con igual probabilidad y de forma sistemática las viviendas a formar parte de la muestra. Luego, se listaron todos los trabajadores independientes al interior de la vivienda y se seleccionaron de forma aleatoria tantos trabajadores como tipos de actividad que éstos desempeñen. No obstante, si al interior de una vivienda existía más de un trabajador independiente desempeñando la misma actividad económica, entonces sólo se seleccionó uno por cada actividad para efectos de no duplicar información.

En las siguientes secciones se describen las características de los marcos de muestreo de ambas fases, y la estimación y distribución del tamaño muestral.

[2] Trimestre móvil marzo, abril y mayo de 2013.

2.1. Características del Marco Muestral

A continuación se describen las características del marco muestral a partir del cual se seleccionó la muestra de la III EME. Como las unidades seleccionadas en la EME proceden desde la ENE, se deben revisar las características del marco de muestreo asociados a la fase 1 (ENE) y la fase 2.

2.1.1. Cobertura geográfica

La cobertura es una propiedad estadística asociada al marco muestral que se utiliza para la selección de la muestra. Así, el ámbito geográfico de la cobertura muestral, comprende el área urbana y rural del país. Sin embargo, se deben hacer algunas especificaciones de ciertas áreas que no cubre la encuesta.

La III EME, posee un diseño muestral bifásico, por lo tanto comparte las propiedades de cobertura de dos marcos muestrales, primero el utilizado para la selección de las viviendas de la ENE (período MAM 2013); y segundo el marco utilizado para la selección de los "independientes" para la III EME.

El marco muestral del INE, utilizado como base para la ENE y todas las encuestas de hogares, cubre sólo a la población que reside en viviendas particulares ocupadas y, por lo tanto, excluye a la población que habita en viviendas colectivas como: hogares de ancianos, hospitales, cárceles, conventos, etc.; y también a la población que reside en la calle. Sin embargo, se incluye a los hogares de personas que habitan y trabajan dentro de dichos centros, como porteros, conserjes y otros.

Además, el marco muestral de la ENE, excluye las viviendas ubicadas en las 22 áreas geográficas catalogadas por el INE como áreas de difícil acceso (ADA) o alto costo (que corresponden al 0,3% del total viviendas)[3]. Por otro lado, para optimizar el trabajo de campo y dadas las características de las unidades muestrales del área urbana (manzanas) se descartan del marco muestral, previo a la selección, las manzanas con 7 o menos viviendas. En total, el marco de la ENE excluye alrededor del 1,03% de las viviendas del país, según el Censo de Población y Vivienda del año 2002.

[3] Ver Anexo N°1

Finalmente, en la elaboración del marco muestral de III EME, se excluyen intencionadamente todas las viviendas que no poseen un "trabajador independiente", es decir, que no poseen unidades elegibles[4].

2.1.2. Estratificación del Marco Muestral

El Marco de Muestreo de la ENE fue estratificado según su condición geográfica (División Político Administrativa) y según el número de viviendas y población que contenían al CENSO 2002, además de una segregación dependiendo de la actividad económica preponderante en el área.

La estratificación del Marco de la ENE da origen a los siguientes estratos:

- Ciudades o grandes Centros Urbanos (CD): Conformadas por ciudades o conjuntos de ciudades adyacentes con 40.000 ó más habitantes.
- Resto de Área Urbana (RAU): Conformadas por conjuntos de Centros Urbanos con menos de 40.000 habitantes.
- Área Rural (R): Conformado por el conjunto de entidades clasificadas como rurales de acuerdo a un tamaño poblacional menor a 1.000 habitantes o entre 1.001 y 2.000 habitantes con predominio de Población Económicamente Activa (según información del Censo de Población y Vivienda del año 2002) dedicada a actividades primarias[5].

En la segunda fase, la III EME tiene cobertura del área urbana y rural del país, estratificada de forma natural de acuerdo a las 15 regiones que posee el país.

Cabe señalar que para fines de análisis y ajustes de los factores de expansión, las regiones fueron agrupadas en cuatro macrozonas: Norte, Centro, Sur, y Región Metropolitana. En el cuadro 1 se detalla la composición de cada macrozona.

[4] Se entiende como unidad elegible a los trabajadores clasificados como independiente en la ENE en el período MAM 2013

[5] Se entiende por Actividad Primaria a toda aquella actividad relacionada con la extracción de recursos naturales (agricultura, caza, pesca, minería, etc.).

Cuadro 1. Composición de macrozonas

Macrozona	Región
Norte	Arica y Parinacota
	Tarapacá
	Antofagasta
	Atacama
	Coquimbo
Centro	Valparaíso
	Libertador General Bernardo O'Higgins
	Maule
	Biobío
Sur	La Araucanía
	Los Ríos
	Los Lagos
	Aisén del General Carlos Ibáñez del Campo
	Magallanes y La Antártica Chilena
Metropolitana	Metropolitana de Santiago

Fuente: Elaboración propia

2.1.3. Depuración del listado de trabajadores independientes.

En correspondencia con el diseño muestral de la III EME, se elaboró un listado de unidades que permitiera la identificación de los trabajadores independientes. Para esto, a partir de la información recogida en la Encuesta Nacional de Empleo en el trimestre MAM 2013, se creó un listado de personas, clasificadas como trabajador independiente, el cual fue utilizado como marco muestral para la selección de la muestra de trabajadores independientes a entrevistar en la III EME.

Al momento de construir el listado o marco definitivo de la EME se realizó una revisión de las personas clasificadas como trabajadores independientes, a través de la revisión de variables como rama de actividad económica (específicamente para descartar aquellos temporeros agrícolas que se autoclasifican como trabajadores independientes), grupo ocupacional (específicamente para descartar aquellas ocupaciones asociadas a personas que trabajan como junior de almacén u oficina, empaquetadores y que se autoclasifican como independientes), número de trabajadores que posee el negocio o actividad, tipo de ingreso, entre otras variables. Esto, porque la ENE es contestada por un informante idóneo (proxy), quien responde por él y por todos los integrantes de su hogar, lo que constituye una fuente

de error no muestral de clasificación, propio de las encuestas a hogares, según los cuales una persona pudiera ser clasificada como trabajador independiente en la ENE, pero que en la realidad no lo sea, y viceversa.

En el cuadro 2, se presentan las variables "Total Independientes ENE", correspondiente al total de personas clasificadas en la ENE como trabajadores independientes, en el período MAM 2013; junto con la variable "Total Independientes EME", la cual hace referencia al universo de personas independientes luego de la depuración de la base de la ENE, utilizado para la selección de la muestra en la III EME. En total, la depuración del marco corresponde a 7%[6] de casos descartados por ser potenciales unidades no elegibles[7], observándose los mayores cambios en la región de O'Higgins (9,3%) y los menores en la región del Biobío con un 4,6%.

Cuadro 2. Distribución del total de independientes muestrales según ENE (MAM 2013) y según Marco EME

Macrozona	Región	Total Independientes ENE	Total Independientes EME
Total		**11.555**	**10.712**
Norte	Arica y Parinacota	493	467
	Tarapacá	469	434
	Antofagasta	242	218
	Atacama	305	289
	Coquimbo	736	685
Total Norte		**2.245**	**2.093**
Centro	Valparaíso	1.424	1.315
	Libertador General Bernardo O'Higgins	580	526
	Maule	804	749
	Biobío	1.328	1.267
Total Centro		**4.136**	**3.857**
Sur	La Araucanía	896	845
	Los Ríos	452	427
	Los Lagos	1.019	925
	Aisén del General Carlos Ibáñez del Campo	325	298
	Magallanes y La Antártica Chilena	126	119
Total Sur		**2.818**	**2.614**
Metropolitana	Metropolitana de Santiago	2.356	2.148
Total Metropolitana		**2.356**	**2.148**

Fuente: Elaboración propia

[6] $\frac{11.555-10.712}{11.555} = 0{,}07$

[7] En la EME, se entiende por unidades no elegibles, aquellos individuos que en la ENE fueron clasificados como trabajadores independientes, según información proporcionada por informante proxy, sin embargo, al momento de realizar el trabajo de campo se observa que la persona seleccionada, en el período de referencia de la ENE no era un trabajador independiente.

2.2. Estimación y Distribución del tamaño muestral

La III EME al poseer un diseño bifásico, considera que sus unidades serán seleccionadas a partir de otra encuesta o listado, en particular de la ENE. En este contexto, los parámetros a utilizar para la determinación del tamaño muestral fueron extraídos de la ENE, para las subpoblaciones específicas de Trabajadores por Cuenta Propia y Empleadores, los que conforman los llamados "Trabajadores Independientes".

2.2.1. Tamaño de la muestra

La estimación del tamaño muestral, se obtuvo a partir de un muestreo aleatorio simple en cada nivel de estimación, al cual se le aplican principalmente tres correcciones: la primera da cuenta del diseño muestral a partir de un estadígrafo denominado efecto del diseño (deff); la segunda da cuenta que la población en estudio es finita; y la tercera, corrige el tamaño para compensar la falta de respuesta, pérdida usual en este tipo de estudios.

El parámetro de estudio o variable de interés (pivote) para el cual se necesita obtener estimaciones precisas en la población U o nivel de estimación (nacional), es una razón entre dos variables:

$$R_{Y/X} = \frac{N° \, Trabajadores \, independientes}{N° \, personas \, Ocupadas} = \frac{Y}{X} = \frac{\sum_{k \in U} y_k}{\sum_{k \in U} x_k} \qquad (1)$$

La variable pivote considerada fue la proporción de trabajadores independientes, debido a que el objetivo principal de la encuesta se centra en caracterizar dicha población.

El método a utilizar para estimar un tamaño muestral adecuado en términos de precisión de acuerdo a los requerimientos, se basa en la relación entre el error estándar[8] y el tamaño de muestra empleado para obtenerlo.

[8] El error estándar de la estimación es simplemente la raíz cuadrada de la varianza de la estimación, esto es: $SE_{\hat{P}} = \sqrt{V(\hat{P})}$, o alternativamente, la varianza es igual al cuadrado del error estándar, $V(\hat{P}) = SE_{\hat{P}}^2$

El Error Estándar SE del estimador $\hat{P}$ en relación al porcentaje de individuos con cierta característica, en el contexto de un muestreo polietápico, está dado aproximadamente por la expresión:

$$V(\hat{P}) = SE_{\hat{P}}^2 \approx \left(1 - \frac{m}{M}\right) \frac{S_{\hat{P}}^2 \cdot Deff_{\hat{P}}}{m} \quad (2)$$

En esta expresión, $Deff_{\hat{P}}$ es el efecto del diseño[9], $f = \frac{m}{M}$ es la fracción de muestreo y $1 - f$ es la corrección por finitud o factor de corrección de la varianza en muestreo de poblaciones finitas, siendo m el número de viviendas a encuestar y M el número de viviendas en la población del nivel de estimación requerido.

El error absoluto de la estimación del parámetro P, denotado como $E_A(\hat{P})$, está relacionado con la varianza de esta misma estimación por la expresión:

$$E_A(\hat{P}) = Z_{1-\alpha/2} \cdot SE_{\hat{P}} = Z_{1-\alpha/2} \cdot \sqrt{V(\hat{P})} \quad (3)$$

Siendo $Z_{1-\alpha/2}$ el percentil $1 - \alpha/2$ de la distribución Normal Estándar, asociada a una estimación por intervalos de $1 - \alpha$ de nivel de confianza. Por lo general, se usa un nivel de confianza del 95%, por lo cual el percentil equivale al 97,5% y el valor usado es entonces. $Z_{1-\alpha/2} = 1{,}96$

Luego, para determinar el tamaño muestral se deben fijar ciertos parámetros, como: la tasa de no respuesta (Tnr), el error absoluto $E_A(\hat{P}) = e_0$, y el nivel de confianza $1 - \alpha$.

Finalmente el tamaño muestral se determina mediante la siguiente fórmula,

$$m = \frac{Z_{1-\alpha/2}^2 \cdot S_{\hat{P}}^2 \cdot Deff_{\hat{P}}}{e_0^2 + \frac{Z_{1-\alpha/2}^2 \cdot S_{\hat{P}}^2 \cdot Deff_{\hat{P}}}{M}} \cdot \frac{1}{(1 - Tnr)} \quad (4)$$

En el apartado siguiente se detalla el cálculo del tamaño muestral.

[9] Se puede interpretar como el aumento o disminución en la varianza, debido a considerar un muestreo complejo (es decir. estratificado, bietápico, por conglomerados) en vez de un muestreo aleatorio simple de viviendas. Aproximadamente, es el cociente entre la varianza de un muestreo multietápico y la de un muestreo aleatorio simple de viviendas.

2.2.2. Estimación del Tamaño Muestral

De acuerdo a lo señalado anteriormente, primero se determinó el tamaño muestral bajo las dos primeras correcciones: el efecto del diseño y por finitud. De acuerdo a esto, el tamaño muestral es de 5.897 viviendas, tamaño determinado con un nivel de confianza del 95%, y un error absoluto de 1,46%.

Cuadro 3. Total de viviendas a encuestar sin considerar corrección por no respuesta.

	Parámetros Obtenidos ENE		Tamaño sin Tnr		
Nivel de Estimación	**Estimación P[10]**	**Deff**	**N° viviendas Esperado**	**Error Absoluto E A**	**Error Relativo E R**
Nacional	23,7%	2,772	5.897	1,46%	6,2%

Fuente: Elaboración propia

Todas las encuestas de hogares sufren la pérdida de unidades debido al agotamiento del informante, o unidades no elegibles debido a desactualización del marco de muestreo, rechazos, etc. En encuestas donde el diseño muestral es bifásico, dicho problema puede acrecentarse debido a que la condición que hace a la unidad elegible puede cambiar en el tiempo. En el caso de la III EME, la condición de "trabajador independiente" puede cambiar de un período a otro, por lo tanto es más probable obtener un menor número de unidades con información al finalizar el proceso de levantamiento.

Al considerar una tasa de no respuesta esperada de alrededor del 15%, el total de viviendas a seleccionar y enviar a terreno es de 6.880, de las cuales se espera obtener información de al menos 5.897 unidades.

Cuadro 4. Tamaño muestral (Total de viviendas) determinado según la proporción de independientes.

Nivel de Estimación	**Estimación P[11]**	**Deff**	**N° viviendas seleccionar**
Nacional	23,7%	2,772	6.880

Fuente: Elaboración propia

[10] P corresponde a la razón entre el total de trabajadores independientes y total de personas ocupadas en el período de referencia.
[11] P corresponde a la razón entre el total de trabajadores independientes y total de personas ocupadas en el período de referencia.

El tamaño de la muestra teórica[12] es de 6.880 viviendas aproximadamente, sujeto a un nivel de estimación nacional y error absoluto fijo de 1,46%. Dichas unidades fueron distribuidas de forma proporcional en las 15 regiones del país, de acuerdo a la estructura observada en la ENE para el trimestre de referencia. Sin embargo, la encuesta sólo tendrá representatividad a nivel nacional.

2.2.3. Distribución de la muestra entre regiones según submuestra

Una vez obtenido este tamaño muestral requerido de acuerdo a los objetivos de precisión a nivel nacional - 6.880 viviendas - se distribuyeron éstas en los distintos subniveles de desagregación en forma proporcional al tamaño, según la cantidad de trabajadores independientes reportados en la ENE. Debido que al momento de diseñar la muestra aún no se contaba con el total de independientes del período MAM 2013, se decidió utilizar como información auxiliar el total de independientes reportado en el trimestre móvil MAM 2012, según región y mes de levantamiento.

Como la muestra de la ENE está subdividida en tres meses o períodos de levantamiento, con el objetivo de disminuir el tiempo transcurrido entre el levantamiento de información de la ENE y la EME y con ello tener una menor atrición, se distribuyó la muestra de la III EME en tres meses de levantamiento independientes entre sí, Mayo, Junio y Julio, de acuerdo al mes de levantamiento de la ENE, Marzo, Abril y Mayo, respectivamente.

La distribución regional se realizó de forma proporcional en cuanto al total de viviendas con al menos un trabajador independiente reportado en MAM 2012. Es decir, en aquellas regiones donde se observó un mayor número de trabajadores independientes se le asignó un mayor número de viviendas a encuestar. Posteriormente, al interior de cada región la muestra fue subdividida en tres partes iguales, cuando ello fuera posible, según el mes de levantamiento. Así, las viviendas a encuestar en el mes de mayo en la III EME deberán ser aquellas viviendas que fueron entrevistadas en Marzo 2013 en la ENE.

A continuación se ilustra la distribución de la muestra según mes de levantamiento y región.

[12] Cabe mencionar que los errores efectivos se calculan con la muestra efectivamente lograda en terreno, ante lo cual los errores pueden ser mayores a los teóricos. Este tamaño corresponde al obtenido de las simulaciones adicionales, considerando una tasa de no-respuesta del 15%, aproximadamente.

Cuadro 5. Total de viviendas seleccionadas según región y mes de levantamiento.

Macrozona	Región	Mes Levantamiento III EME			Total
		Mayo	Junio	Julio	
Total EME		**2.294**	**2.293**	**2.293**	**6.880**
Norte	Arica y Parinacota	92	93	92	277
	Tarapacá	88	88	88	264
	Antofagasta	42	42	42	126
	Atacama	62	61	61	184
	Coquimbo	149	149	150	448
Total Norte		**433**	**433**	**433**	**1.299**
Centro	Valparaíso	286	286	287	859
	Libertador General Bernardo O'Higgins	123	123	123	369
	Maule	167	167	167	501
	Biobío	261	260	260	781
Total Centro		**837**	**836**	**837**	**2.510**
Sur	La Araucanía	178	178	178	534
	Los Ríos	93	93	93	279
	Los Lagos	210	209	209	628
	Aisén del General Carlos Ibáñez del Campo	65	65	65	195
	Magallanes y La Antártica Chilena	24	25	24	73
Total Sur		**570**	**570**	**569**	**1.709**
Metropolitana	Metropolitana de Santiago	454	454	454	1.362
Total Metropolitana		**454**	**454**	**454**	**1.362**

Fuente: Elaboración propia

Cabe señalar que la distribución del total de viviendas a encuestar según región, está dada por la distribución del total de independientes observados en la ENE en MAM 2012. Sin embargo, el marco de muestreo desde el cual se seleccionó la III EME, puede tener una distribución similar pero no idéntica.

2.3. Selección de Unidades

La Encuesta Nacional de Empleo registra para cada miembro del hogar de 15 o más años, la información necesaria para caracterizarlos de acuerdo a si éstos pertenecen o no a la Fuerza de Trabajo. Además de ello, registra información que permite la categorización de las personas "ocupadas" según la Clasificación Internacional de la Situación de Empleo (CISE), lo que permite identificar la población objetivo, es decir, "los trabajadores por cuenta propia y empleadores". Esta variable es la que permite la construcción del Marco de Muestreo de la EME, a partir del cual se seleccionaron las viviendas y personas. En el cuadro 6 se presenta la distribución del total de viviendas y personas según región.

Cuadro 6. Total de viviendas y personas Marco EME

Macrozona	Región	ENE MAM 2013		Selección EME	
		Total Independientes	Total viviendas	Total Independientes	Total viviendas
Total EME		**10.712**	**9.304**	**7.632**	**6.880**
Norte	Arica y Parinacota	467	411	306	277
	Tarapacá	434	356	308	264
	Antofagasta	218	195	143	126
	Atacama	289	250	200	184
	Coquimbo	685	605	495	448
Total Norte		**2.093**	**1.817**	**1.452**	**1.299**
Centro	Valparaíso	1.315	1.158	935	859
	Libertador General Bernardo O'Higgins	526	466	405	369
	Maule	749	648	558	501
	Biobío	1.267	1.114	858	781
Total Centro		**3.857**	**3.386**	**2.756**	**2.510**
Sur	La Araucanía	845	710	618	534
	Los Ríos	427	365	314	279
	Los Lagos	925	783	705	628
	Aisén del General Carlos Ibáñez del Campo	298	251	221	195
	Magallanes y La Antártica Chilena	119	108	79	73
Total Sur		**2.614**	**2.217**	**1.937**	**1.709**
Metropolitana	Metropolitana de Santiago	2.148	1.884	1.487	1.362
Total Metropolitana		**2.148**	**1.884**	**1.487**	**1.362**

Fuente: Elaboración propia

La selección se realizó en dos etapas, primero sobre las viviendas y luego en su interior a los independientes. Las viviendas fueron seleccionadas con igual probabilidad, de forma sistemática al interior de cada región. Las unidades seleccionadas fueron ordenadas previamente de acuerdo a las variables que identifican la división política administrativa (región, provincia, comuna, distrito censal, zona censal, manzana) y área (urbano-rural). De esta manera se garantiza que estén representadas todas las áreas geográficas en la misma medida como éstas se encuentran en el marco de muestreo.

Posteriormente, en cada vivienda seleccionada, se identificaron todos los independientes y las actividades económicas en que éstos se desenvuelven. Luego, se seleccionaron de forma aleatoria y con igual probabilidad tantos independientes como actividades identificadas. Sin embargo, en caso de encontrar más de un independiente en el hogar ejecutando la misma actividad económica, se tomó el resguardo de seleccionar sólo a un representante por actividad.

3. FACTORES DE EXPANSIÓN

La muestra de la III EME fue diseñada para lograr representatividad a nivel nacional. En atención a los errores que se desea alcanzar y al presupuesto disponible, se determinó como tamaño óptimo la recolección de 6.800 viviendas aproximadamente. Para compensar las pérdidas asociadas a la no respuesta, la muestra efectiva fue sobre-dimensionada en un 15%, aproximadamente.

Los factores de expansión se obtienen como el inverso de las probabilidades de selección, además de la aplicación de diversos ajustes. En este caso, las probabilidades de selección asociados a los trabajadores independientes tienen varias componentes:

1. Probabilidad de que la vivienda hubiera sido seleccionada y contestado la ENE período MAM 2013.
    - Probabilidad de seleccionar el conglomerado de pertenencia.
    - Probabilidad de seleccionar la vivienda dado que el conglomerado al que pertenece fue seleccionado.
    - Probabilidad de responder ENE.
2. Probabilidad de seleccionar una vivienda para EME, dado que la vivienda posee trabajadores independientes.
3. Probabilidad de seleccionar un trabajador independiente, dado que su vivienda fue seleccionada.

Mientras que respecto a los ajustes que se deben realizar, éstos son:

1. Ajuste por falta de respuesta (probabilidad de que el trabajador independiente participe en EME III).
2. Ajuste a un stock poblacional dado un período de referencia.

Respecto a los elementos utilizados en los cálculos del factor de expansión, se puede especificar que:

1. Lo referido a las probabilidades de selección de la primera fase, se extraen directamente de la Encuesta Nacional de Empleo, ya que son éstas las utilizadas en el factor de expansión de la ENE (previo a la post-estratificación por sexo y tramo de edad).
2. Tanto las probabilidades de selección de las viviendas y de las personas se extraerán directamente desde el Marco de la III EME.

3. Respecto al ajuste por falta de respuesta, se deben utilizar los grupos o “celdas de ajuste”, creadas a partir de la información existente, tanto de los que responden como los que no, en la III EME.
4. Calibración al stock poblacional, el cual fue creado a partir de los datos recogidos en la ENE en el período de referencia donde se seleccionó la muestra (MAM 2013), pero ajustados al crecimiento poblacional estimado a partir de las proyecciones poblacionales de junio del 2013, según macrozona y sexo.

En los apartados siguientes se detalla el proceso de cálculo de las probabilidades de selección, así como también de los factores. Se hablará indistintamente de factores de expansión y de ponderadores.

3.1. Ponderador Base

El ponderador base se define como el factor de expansión obtenido sólo con las probabilidades de selección, sin ajustes ni correcciones.

En la EME III, las personas seleccionadas, corresponden a un subconjunto de personas que participaron durante el proceso de encuestaje del trimestre MAM 2013 de la ENE. Por lo tanto, uno de los insumos fundamentales del ponderador base, son los factores de expansión de vivienda de la ENE, que dan cuenta de la probabilidad de que una vivienda haya sido seleccionada y entrevistada en la ENE. La sección 3.1.1 expone las probabilidades de selección y respuesta de la ENE; la sección 3.1.2 expone la formula explicita de la probabilidad condicional de selección de un trabajador independiente; finalmente, en la sección 3.1.3 se expone la fórmula matemática del ponderador base.

3.1.1. Probabilidad de selección y entrevista de las viviendas en la muestra de la ENE –MAM 2013.

El diseño muestral de la Encuesta Nacional de Empleo, corresponde a un muestreo probabilístico, estratificado y bietápico, donde las unidades primarias corresponden a manzanas en el área urbana y secciones en el área rural; mientras que las unidades de segunda etapa son las viviendas particulares. Las unidades primarias fueron seleccionadas en forma proporcional al tamaño, mientras que al interior de cada manzana o sección las unidades de segunda etapa se seleccionaron de forma sistemática y con igual probabilidad. El factor de expansión de la ENE posee un ajuste por no respuesta implícito, es decir, el peso de las unidades que no

responden es distribuido en el resto de las viviendas del conglomerado al cual pertenecen.

La expresión que se detalla a continuación fue extraída desde el documento "Manual conceptual y Metodológico del diseño muestral de la ENE[13]", que corresponde al ponderador inicial o teórico corregido por no respuesta.

$$F_{hij}^{1} = \underbrace{\left(\frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}^{T}}\right)}_{\substack{Factor\ de\ expansión \\ teórico}} \cdot \overbrace{\frac{m_{hi}^{T}}{\left(m_{hi}^{T} - m_{hi}^{NR}\right)}}^{\substack{Ajuste \\ no\ respuesta}} = \frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}}$$

Donde:

h: Subíndice que representa el estrato de muestreo ENE.

i: Subíndice que representa el conglomerado i.

j: Subíndice que representa la vivienda j.

M_h: Total de viviendas en el estrato h, según el Marco de muestreo de la ENE.

n_h: Total de conglomerados seleccionados en el estrato h en la ENE.

M_{hi}: Total de viviendas particulares que contiene el conglomerado i del estrato h, según información del Marco muestral.

M'_{hi}: Total de viviendas particulares que contiene el conglomerado i del estrato h , según información recogida en enumeración.

m_{hi}^{T}: Total de viviendas seleccionadas en el conglomerado i del estrato h

m_{hi}^{NR}: Total de viviendas seleccionadas en el conglomerado i del estrato h que no responden.

m_{hi}: Total de viviendas que responde en la ENE en el período MAM 2013.

En consecuencia, la probabilidad de haber sido seleccionada y entrevistada la vivienda j, del conglomerado i, en el estrato h en el trimestre móvil MAM 2013 en la ENE, está dado por:

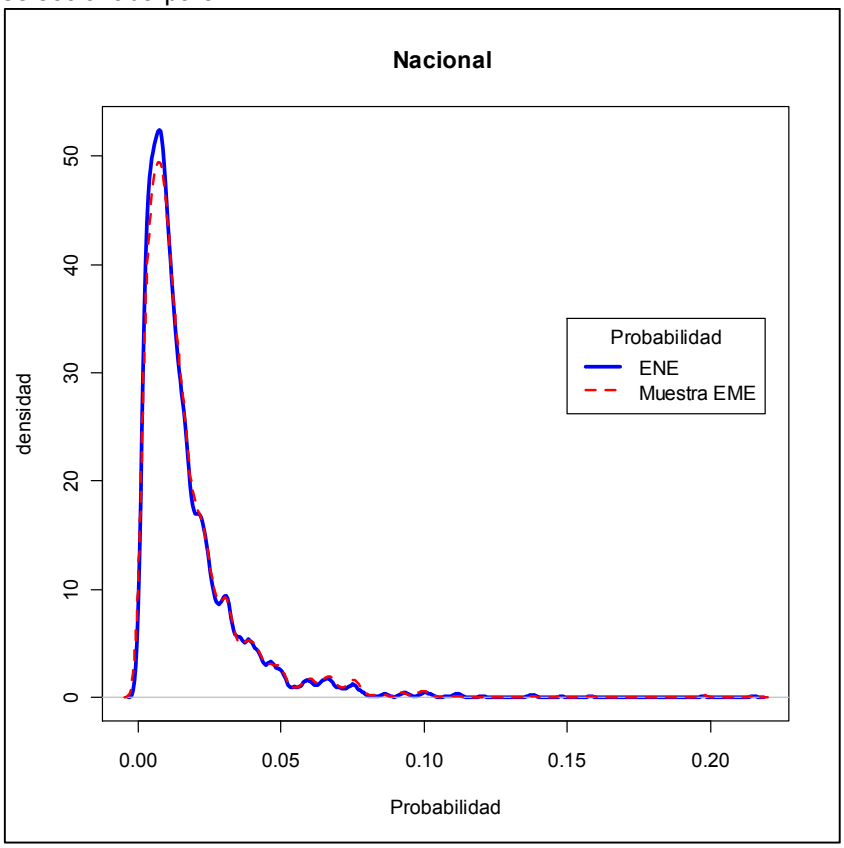
$$P_{hij}^{v} = \frac{1}{F_{hij}^{1}}$$

En el gráfico 1 se ilustra la función de densidad de la probabilidad de selección y entrevista de las viviendas en la muestra de la ENE y EME. Ambas funciones son similares pues poseen distribución asimétrica cargada a la derecha, lo cual se

[13] http://www.ine.cl/canales/chile_estadistico/mercado_del_trabajo/empleo/metodologia/pdf/031110/manual_metodologico031110.pdf

explica por la existencia de un pequeño número de viviendas con ponderadores chicos. A partir de este gráfico, se puede observar que cerca del 95% de las unidades - totalidad de las unidades de la ENE y el subconjunto seleccionado para la EME- poseen una probabilidad de selección y respuesta en la ENE igual o inferior al 5% (eje x igual a 0,05).

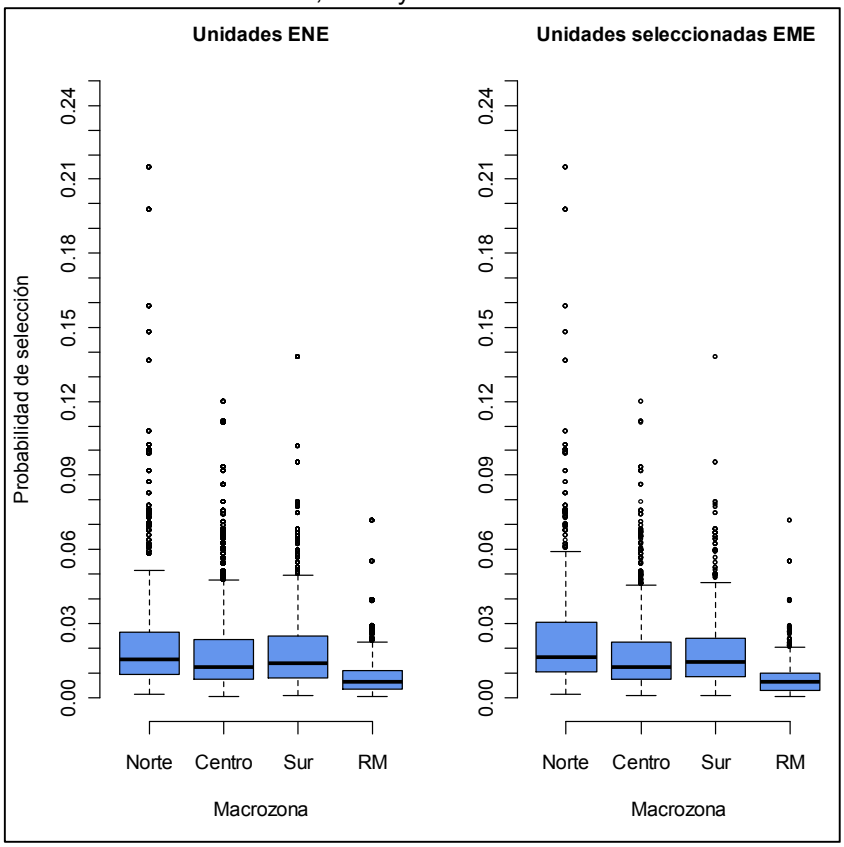
Gráfico 1. Función de densidad de las probabilidades de selección y entrevista de las viviendas en la ENE – completa versus muestra seleccionada para EME.



Fuente: Elaboración propia

Al realizar el análisis según macrozona (ver gráfico 2), se observa que la Región Metropolitana presenta las probabilidades de selección y respuesta más pequeñas y con menor variabilidad, pues el 99% de los casos (bigote superior del grafico de caja), presenta probabilidades inferiores a 2,5%. Por otro lado, la macrozona Norte es la que presenta mayor variabilidad, ya que el 50% de sus observaciones poseen probabilidades inferiores al 2,5%, mientras que el 50% restante posee probabilidades entre 2,5% y 25% aproximadamente.

Gráfico 2. Distribución de las probabilidades de selección y respuesta de las unidades de la ENE, Total y sólo seleccionadas EME.



Fuente: Elaboración propia

A continuación, se especifica la forma de cálculo de las probabildiades de selección de los trabajadores independientes, tanto las probabilidades condicionales a la selección en la ENE como aquellas incondicionales. Además, se explicita la forma en que se contruyó el ponderador base.

3.1.2. Probabilidad de selección de los independientes

La selección de los independientes se realizó en dos etapas. Primero, se seleccionaron con igual probabilidad viviendas que contenían al menos un independiente, según mes de levantamiento al interior de cada región.

Así, la probabilidad de selección de una vivienda que posee al menos un independiente está dada por:

$$p_{Rj}^{v} = \frac{m_R^{indep}}{M_R^{indep}}$$

Donde:

R: Subíndice que representa la región de pertenencia. $R = 1, \ldots, 15$.

j: Subíndice que representa la vivienda j.

p_{Rj}^{v}: Corresponde a la probabilidad de seleccionar la vivienda j perteneciente a la región R, según el listado de viviendas de la ENE que poseen al menos un independiente.

M_R^{indep}: Corresponde al total de viviendas con al menos un independiente en la región R, de acuerdo a la clasificación de la ENE.

m_R^{indep}: Corresponde al total de viviendas seleccionadas con al menos un independiente en la región R.

Luego, una vez seleccionada la vivienda se seleccionan los independientes. La probabilidad de seleccionar al independiente k al interior de la vivienda j, perteneciente a la región R, dado que la vivienda fue seleccionada, puede ser aproximada por:

$$p_{Rjk}^{indep|v} = \frac{S_{Rj}^{indep}}{T_{Rj}^{indep}}$$

Donde:

T_{Rj}^{indep}: Corresponde al total de independientes identificados en la ENE, en la vivienda j, perteneciente a la región R.

S_j^{indep}: Corresponde al total de independientes seleccionados, en la vivienda j, perteneciente a la región R.

Luego la probabilidad condicional de seleccionar el independiente k, en la vivienda j, de la región R, puede ser aproximada por la siguiente expresión:

$$p_{Rjk}^{indep} = p_{Rj}^{v} \cdot p_{Rjk}^{indep|v}$$

En el cuadro 7 se observa que, en general, la probabilidad de seleccionar trabajadores independientes (2), dado que respondió la ENE, oscila entre 22% y 8%. Sin embargo, en la Región Metropolitana estas probabilidades se encuentran concentradas en torno a 0,72%, observándose que menos del 5% de las unidades poseen probabilidades inferiores al 30%.

Las altas probabilidades que se presentan en las distintas macrozonas se explican principalmente por la selección, en algunas regiones, de un gran número de trabajadores independientes en comparación al total de unidades disponibles para seleccionar.

Cuadro 7. Estadísticas descriptivas de la probabilidad de selección de las viviendas y personas, según macrozona

Estadísticas descriptivas	Macrozona							
	Norte		Centro		Sur		Región Metropolitana	
	Probabilidad de selección vivienda(1)	Probabilidad de selección personas (2)	Probabilidad de selección vivienda(1)	Probabilidad de selección personas (2)	Probabilidad de selección vivienda(1)	Probabilidad de selección personas (2)	Probabilidad de selección vivienda(1)	Probabilidad de selección personas (2)
Recuento	1.299	1.299	2.510	2.510	1.709	1.709	1.362	1.362
Moda	0,74	0,74	0,74	0,74	0,80	0,80	0,72	0,72
Mínimo	0,65	0,22	0,70	0,23	0,68	0,26	0,72	0,24
Percentil 05	0,65	0,65	0,70	0,70	0,75	0,68	0,72	0,72
Percentil 25	0,67	0,67	0,70	0,70	0,75	0,75	0,72	0,72
Mediana	0,74	0,74	0,74	0,74	0,76	0,76	0,72	0,72
Percentil 75	0,74	0,74	0,77	0,77	0,80	0,80	0,72	0,72
Percentil 95	0,74	0,74	0,79	0,79	0,80	0,80	0,72	0,72
Percentil 99	0,74	0,74	0,79	0,79	0,80	0,80	0,72	0,72
Máximo	0,74	0,74	0,79	0,79	0,80	0,80	0,72	0,72
Media	0,72	0,70	0,74	0,73	0,77	0,76	0,72	0,71

Fuente: Elaboración propia

3.1.3. Inverso de las probabilidades de selección o Ponderador Base

El ponderador base, es aquel que da cuenta de las probabilidades de selección de las viviendas en la fase 1, y las probabilidades de selección de los independientes en la fase 2, condicional a que la vivienda de residencia fue seleccionada en la ENE y que éstas participaran en el período MAM 2013.

Así, calculadas las probabilidades de selección y participación de una vivienda en la ENE en el trimestre MAM 2013 y la probabilidad de seleccionar un independiente desde la ENE, el ponderador base se calcula como:

$$F_{Rjk}^{base} = \left(\frac{1}{P_{hij}^{v}}\right) \cdot \left(\frac{1}{p_{Rjk}^{indep}}\right)$$

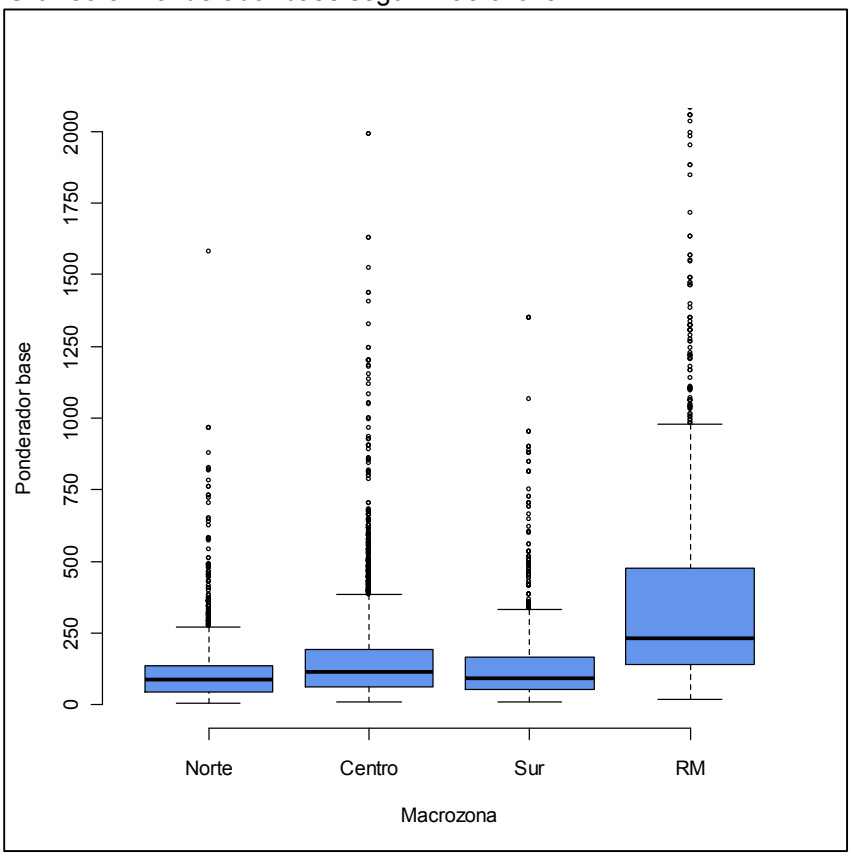
En el cuadro 8 se observa que la Región Metropolitana presenta mayores ponderadores base, mientras que las macrozonas Norte y Sur poseen menores valores, pero entre ellos similares distribuciones y variabilidad de sus ponderadores. Según el gráfico 3, el mayor valor del ponderador se observa en la Región Metropolitana, siendo hasta cuatro veces más grande que los valores extremos de las restantes macrozonas.

Cuadro 8. Estadísticas descriptivas del ponderador base

Estadísticas descriptivas	Macrozona				Total País
	Norte	Centro	Sur	Región Metropolitana	
Recuento	1.452	2.756	1.937	1.487	7.632
Moda	120,4	140,1	32,4	111,7	120,4
Mínimo	6,9	11,9	9,6	19,4	6,9
Percentil 05	17,9	27,4	27,6	65,2	26,2
Percentil 25	45,1	61,6	54,6	142,5	63,2
Mediana	86,9	114,1	93,1	232,5	118,3
Percentil 75	136,0	191,6	167,1	478,1	215,1
Percentil 95	359,6	505,9	354,4	1.276,3	612,2
Percentil 99	703,5	994,4	812,6	2.776,0	1.334,4
Máximo	1.579,8	2.133,6	1.350,5	6.804,5	6.804,5
Media	119,7	170,1	137,2	410,1	198,9
Error típico de la media	3,3	3,61	3,2	15,0	3,58
Suma	**173.818,1**	**468.679,6**	**265.753,5**	**609.780,9**	**1.518.032,1**

Fuente: Elaboración propia

Gráfico 3. Ponderador base según macrozona



Fuente: Elaboración propia

También existen valores extremos en cada macrozona. Sin embargo, los casos más preocupantes son los considerados “casos influyentes”, pues las características de un individuo pueden representar hasta 6.804 personas, es decir al 1,1% de la población estimada (suma del ponderador) de la Región Metropolitana. Para minimizar el efecto en las estimaciones de ponderadores de esta magnitud, se implementó un método de verificación de valores extremos y suavizamiento de los mismos.

En el siguiente apartado se revisa la pertinencia de suavizar el ponderador base y el método de suavizamiento.

3.1.4. Suavizamiento del Ponderador Base

En la etapa de construcción del ponderador base, se observaron 5 casos con valores mayores a las 5.000 unidades, los que en conjunto representan a un 5% de los independientes de la Región Metropolitana. Las restantes observaciones poseen ponderadores inferiores o iguales a 3.820. Para identificar la presencia de casos influyentes y reducir su impacto, se implementó un procedimiento de suavizamiento de los factores de expansión que puede ser resumido en 5 pasos,

i. Inspeccionar la existencia de valores extremos en la distribución del ponderador,
ii. Determinar puntos de corte a partir de los cuales realizar el suavizamiento,
iii. Suavizar los valores extremos identificados,
iv. Estimar el error cuadrático medio (ECM) para los distintos puntos de corte,
v. Elegir la opción de corte que minimice el ECM,

El gráfico 4, que muestra los factores de expansión base de forma ordenada, permite identificar al menos tres puntos de discontinuidad del ponderador base: dos casos extremos (los que se encuentran al interior de la elipse continua) que superan las 6.500 unidades, cinco ponderadores, que se encuentran al interior de la elipse semi-continua, que poseen valores superiores a 5.000 y aquellos casos que superan 2.500 unidades. Sin embargo, al realizar este mismo análisis según macrozona, se pueden identificar discontinuidades en valores más pequeños. Por ejemplo, en la macrozona Norte y Sur, se observa una discontinuidad importante a partir del valor 1.500.

Gráfico 4. Dispersión del factor de expansión base o inicial

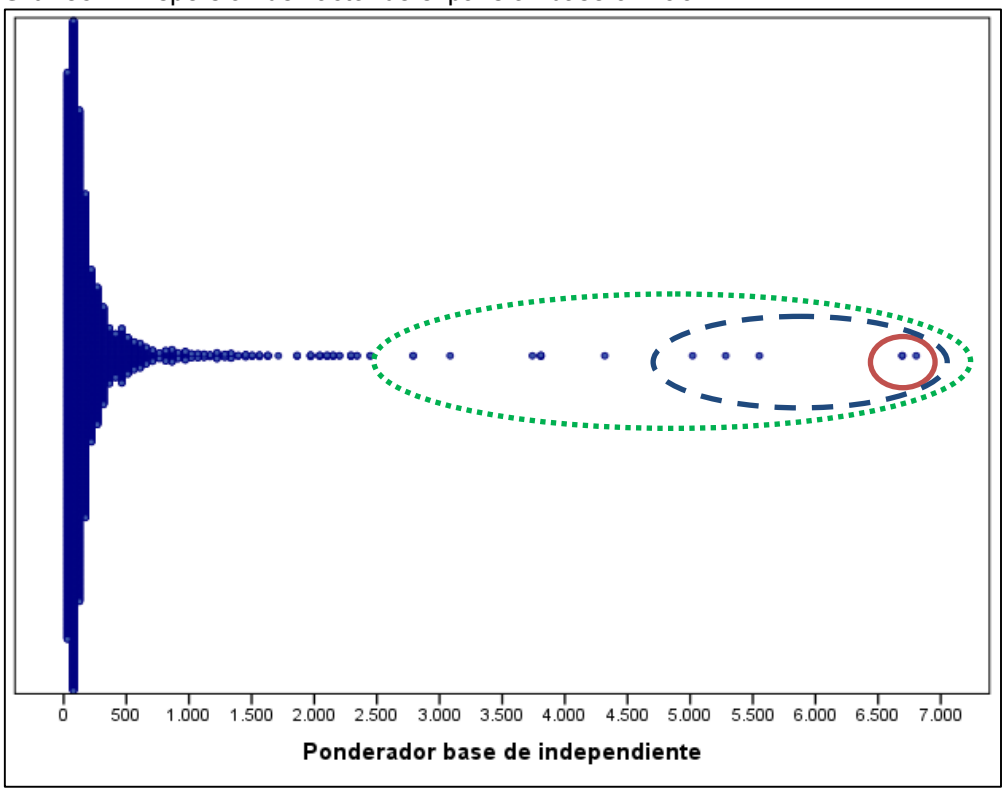
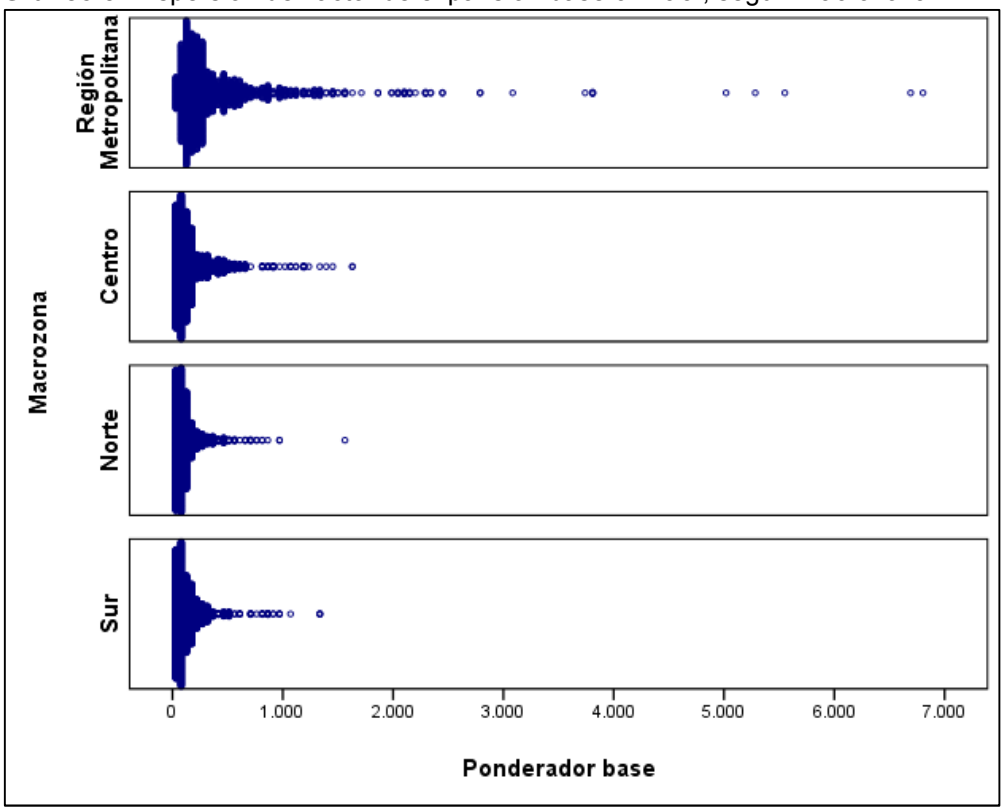


Gráfico 5. Dispersión del factor de expansión base o inicial, según macrozona



Fuente: Elaboración propia

Para inspeccionar la existencia de valores extremos, se utilizaron dos estrategias: (1) identificar discontinuidades, de forma visual, en la distribución del ponderador base; (2) identificar valores extremos a partir de una distancia determinada entre el ponderador promedio y cada valor del ponderador al interior de cada macrozona[14]. Considerando lo anterior, se analizaron 6 puntos de corte distintos:

$$c_p = \begin{cases} c_1 = \frac{1}{5} = 0{,}20 \\ c_2 = \frac{1}{6} = 0{,}16\bar{6} \\ c_3 = \frac{1}{8} = 0{,}125 \\ c_4 = 0{,}002 \\ c_5 = \frac{1}{9} = 0{,}11\bar{1} \\ c_6 = \frac{1}{10} = 0{,}100 \end{cases}$$

Por otro lado, para realizar el suavizamiento se procede a truncar aquellos ponderadores identificados como valores extremos de la siguiente forma,

$Si\ P \neq 4$

$$T_{Rjk,g} = \begin{cases} F_{Rjk}^{base} & si\ \ \delta_{Rjk,g} \geq c_p \\ \dfrac{\bar{F}_{Rjk,g}^{base}}{c_p} & si\ \delta_{Rjk,g} < c_p \end{cases}$$

$Si\ P = 4$

$$T_{Rjk,g} = \begin{cases} F_{Rjk}^{base} & si\ \dfrac{1}{F_{Rjk}^{base}} \geq c_p \\ 5.000 & si\ \dfrac{1}{F_{Rjk}^{base}} < c_p \end{cases}$$

Donde,

g: Subíndice de la macrozona de procedencia de las unidades.

$\bar{F}_{Rjk,g}^{base}$: Corresponde al ponderador base promedio en macrozona g.

$\delta_{Rjk,g} = \frac{\bar{F}_{Rjk,g}^{base}}{F_{Rjk}^{base}}$: Distancia entre el ponderador base de la Región R, vivienda j persona k, perteneciente a la macrozona g, con el ponderador base promedio de la macrozona g.

Si se suman todos los valores $T_{Rjk,g}$, se obtiene un total de unidades estimadas inferior que al sumar los ponderadores base, por lo tanto se deben distribuir los

[14] Al chequear grafico 5 se observó que el comportamiento de los ponderadores es distinto al interior de cada macrozona, por lo tanto se determinó realizar el suavizamiento de forma independiente al interior de cada uno de estos.

pesos faltantes en el resto de los ponderadores que no fueron truncados. Los pesos fueron distribuidos al interior de cada macrozona de la siguiente forma:

$Si\ P \neq 4$

$$F_{Rjk}^{Tr} = \begin{cases} T_{Rjk,g} \cdot \dfrac{(\sum_{k \in g} F_{Rjk}^{base} - \sum_{k \in g \cap \delta_{Rjk,g} \geq c_p} T_{Rjk,g})}{\sum_{k \in g \cap \delta_{Rjk,g} < c_p} T_{Rjk,g}} & si \quad \delta_{Rjk,g} \geq c_p \\ \dfrac{\bar{F}_{Rjk,g}^{base}}{c_p} & si \quad \delta_{Rjk,g} < c_p \end{cases}$$

$Si\ P = 4$

$$F_{Rjk}^{Tr} = \begin{cases} T_{Rjk,g} \cdot \dfrac{(\sum_{k \in g} F_{Rjk}^{base} - \sum_{k \in g \cap \delta_{Rjk,g} \geq c_p} T_{Rjk,g})}{\sum_{k \in g \cap \delta_{Rjk,g} < c_p} T_{Rjk,g}} & si \quad \dfrac{1}{F_{Rjk}^{base}} \geq \quad c_p \\ 5.000 & si \quad \dfrac{1}{F_{Rjk}^{base}} < \quad c_p \end{cases}$$

Donde,

$k \in g \cap \delta_{Rjk,g} < c_p$: persona k, en la macrozona g, tal que la distancia de su ponderador, con el ponderador promedio es inferior al punto c_p establecido.

$k \in g \cap \delta_{Rjk,g} \geq c_p$: corresponde al universo de personas k, en la macrozona g, tal que la distancia de su ponderador, con el ponderador promedio es superior al punto c_p establecido.

Esto es, aquellos ponderadores identificados como valores extremos son truncados al valor máximo establecido ($\bar{F}_{Rjk,g}^{base}/c_p$ o 5.000), mientras que sobre el resto de los ponderadores es distribuido el peso “sobrante” de los ponderadores truncados.

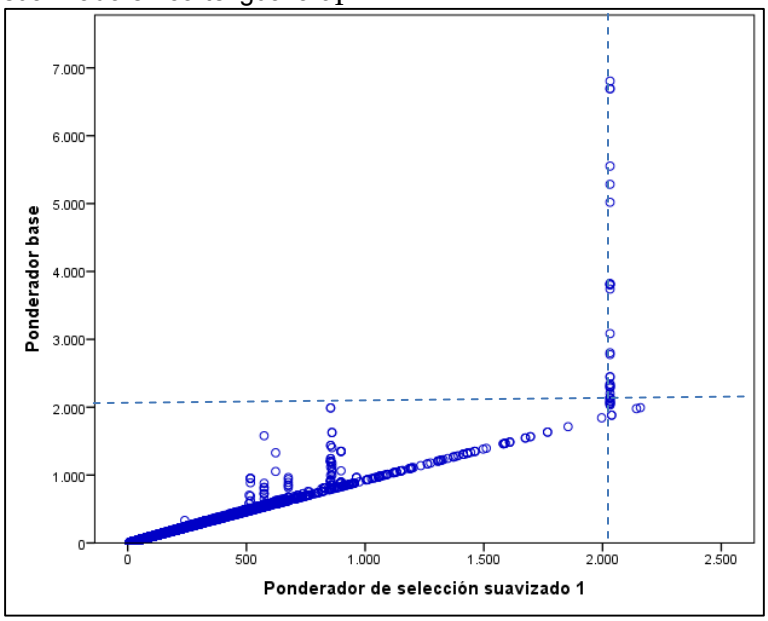
En el cuadro 9 se exponen las estadísticas descriptivas de cada uno de los ponderadores base suavizados según cada uno de los cortes establecidos. Se observa que el ponderador, en promedio, no sufre cambios importantes en la estimación. Sin embargo, el error de estimación asociado decrece. Por otro lado, respecto a los estadísticos relacionados a la forma de la distribución, el coeficiente de asimetría, pareciera mejorar (la distribución es más simétrica) con cada uno de los suavizamientos. Asimismo, se aprecia una mejora importante en el coeficiente de curtosis, pues este estadígrafo se reduce a la mitad, o más, con cada uno de los distintos puntos de suavizamiento. En términos generales se observa que, mientras más exigente es el punto de corte determinado, el número de unidades suavizadas es mayor y, por tanto, los estadígrafos tienen un mejor comportamiento (se reducen los valores extremos, se reduce la variabilidad, mejora el coeficiente de asimetría y curtosis, etc.).

Cuadro 9. Estadísticas descriptivas del ponderador base y ponderador base suavizados en distintos puntos de corte

Estadístico		Ponderador base	Ponderador suavizado c_1	Ponderador suavizado c_2	Ponderador suavizado c_3	Ponderador suavizado c_5	Ponderador suavizado c_6	Ponderador suavizado c_4
Rango		6.798	2.152	2.468	3.078	3.078	3.885	4.311
Mínimo		7	7	7	7	7	7	7
Máximo		6.804	2.159	2.474	3.084	3.084	3.892	4.318
Suma		1.518.032	1.518.032	1.518.032	1.518.032	1.518.032	1.518.032	1.518.032
Media	Estimación	196	197	197	197	197	196	196
	Error típico	3,644	3,003	3,140	3,267	3,274	3,387	3,440
Desv. típ.		300	247	258	269	269	279	283
Varianza		89.842	61.005	66.698	72.217	72.520	77.584	80.064
Asimetría	Estimación	8	4	4	5	5	6	6
	Error típico	0,030	0,030	0,030	0,030	0,030	0,030	0,030
Curtosis	Estimación	126	20	27	36	36	53	63
	Error típico	0,060	0,060	0,060	0,060	0,060	0,060	0,060

Fuente: Elaboración propia

Gráfico 6. Dispersión del ponderador base versus ponderador suavizado en corte igual a c_1.



Fuente: Elaboración propia

Pese a lo indicado anteriormente, se debe revisar el comportamiento de aquellos valores del ponderador que siendo “grandes”, no caen en la categoría de valores extremos, pues al momento de redistribuir los pesos “sobrantes”, es probable que superen el umbral establecido. Esto puede ser visualizado en el gráfico 6, al observar que ciertos ponderadores base, con valores iguales o próximos a 2.000,

luego de ser suavizados, superan el umbral establecido (línea vertical segmentada), lo que significa que el umbral establecido no es óptimo. Ante esto se descartaron inmediatamente dos de los 6 puntos establecidos (c_1 y c_2).

Luego, para determinar el punto de corte donde se realizará finalmente el suavizamiento, se calculó un estadígrafo que diera cuenta del sesgo y de la variabilidad. Para esto se obtuvo el Error Cuadrático Medio (ECM) asociado a la variable de interés. Como en esta encuesta se pretende caracterizar los trabajadores independientes, se estableció analizar la estructura de la variable "Rama de actividad" (reducida a aquellas categorías más importantes[15]) y sobre estas categorías se calculó el sesgo utilizando la siguiente fórmula:

$$sesgo\left(\hat{P}_{c_p}\right) = P_{base} - \hat{P}_{c_p}$$

Después de calcular el sesgo de cada categoría, se calculó el efecto sobre la variable completa, a través de la suma del valor absoluto del sesgo de cada categoría. En el cuadro 10 se aprecia que el punto de corte $\boldsymbol{c_4}$, es el que posee menor sesgo en la categoría G, y por tanto a nivel agregado parece ser el menos sesgado. Le sigue el punto $\boldsymbol{c_6}$ y por último el punto $\boldsymbol{c_3}$ el cual presenta un mayor sesgo.

Cuadro 10. Estimación del sesgo de la estructura de la rama de actividad económica.

	Sesgo			
Rama Actividad Reducida	$\boldsymbol{c_3}$	$\boldsymbol{c_4}$	$\boldsymbol{c_5}$	$\boldsymbol{c_6}$
A. Agricultura, ganadería, caza y silvicultura	0,000	0,000	0,000	0,000
D. Industrias manufactureras	-0,001	-0,001	-0,001	-0,001
F. Construcción	-0,001	-0,001	-0,001	-0,001
G. Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	0,005	0,003	0,005	0,004
I. Transporte, almacenamiento y comunicaciones	-0,001	0,000	-0,001	-0,001
K. Actividades inmobiliarias, empresariales y de alquiler	-0,001	-0,001	-0,001	-0,001
O. Otras actividades de servicios comunitarios, sociales y personales	0,001	0,001	0,001	0,001
Sector Primario	0,000	0,000	0,000	0,000
Servicios	-0,001	-0,001	-0,001	-0,001
Total	0,012	0,007	0,011	0,009

Fuente: Elaboración Propia.

Finalmente, se calcula el ECM para cada categoría como:

$$ECM\left(\hat{P}_{c_p}\right) = Sesgo^2\left(\hat{P}_{c_p}\right) + Var(\hat{P}_{c_p})$$

[15] Ver más detalles en capítulo 4.2

Al revisar el cuadro 11, se observa que en términos del ECM no existen diferencias entre los puntos de suavizamiento.

Por lo tanto, en términos de sesgo, el mejor punto es el c_4. Sin embargo, al tener el mismo ECM que el punto c_6, significa que posee mayor variabilidad. En consecuencia, al comparar todos los ajustes, el punto de corte c_6[16] es aquel que mejor combina un sesgo pequeño junto con una variabilidad mínima, razón por la cual es elegido como el umbral o punto de corte.

Cuadro 11. Estimación del ECM de la estructura de la rama de actividad económica.

	ECM			
Rama Actividad Reducida	c_3	c_4	c_5	c_6
A. Agricultura, ganadería, caza y silvicultura	0,0000	0,0000	0,0000	,0000
D. Industrias manufactureras	0,0001	0,0001	0,0001	,0001
F. Construcción	0,0000	0,0000	0,0000	,0000
G. Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	0,0002	0,0002	0,0002	,0002
I. Transporte, almacenamiento y comunicaciones	0,0000	0,0000	0,0000	,0000
K. Actividades inmobiliarias, empresariales y de alquiler	0,0001	0,0001	0,0001	,0001
O. Otras actividades de servicios comunitarios, sociales y personales	0,0000	0,0001	0,0000	0,0001
Sector Primario	0,0000	0,0000	0,0000	0,0000
Servicios	0,0000	0,0000	0,0000	0,0000
Suma	0,0005	0,0005	0,0005	0,0005
Promedio	0,0001	0,0001	0,0001	0,0001

Fuente: Elaboración Propia.

[16] Sólo 11 observaciones son truncadas bajo este umbral.

En el cuadro 12 se observa que, en términos de distribución, al comparar el ponderador base versus el ponderador suavizado no se registran cambios de los factores de la macrozona Sur. Sin embargo, en las restantes macrozonas los valores extremos fueron suavizados. El mayor cambio se encuentra en la Región Metropolitana, donde el valor máximo 6.804 disminuye a 3.892 unidades.

Cuadro 12. Estadísticas descriptivas del ponderador base y ponderador suavizado.

Estadísticas descriptivas	Macrozona									
	Norte		Centro		Sur		Región Metropolitana		Total	
	Ponderador base	Ponderador c_6	Ponderador base	Ponderador c_6	Ponderador base	Ponderador c_6	Ponderador base	Ponderador c_6	Ponderador base	Ponderador c_6
Recuento	1417	1417	2656	2656	1901	1901	1446	1446	7.420	7.420
Moda	120	120	140	140	32	32	211	217	120	120
Mínimo	7	7	12	12	10	10	19	20	7	7
Percentil 05	18	18	27	27	27	27	64	66	26	26
Percentil 25	45	46	62	62	53	53	143	146	63	64
Mediana	87	88	113	114	93	93	234	240	118	119
Percentil 75	136	136	189	191	165	165	478	490	214	217
Percentil 95	360	363	504	504	343	343	1267	1298	603	616
Percentil 99	648	650	925	935	813	813	2448	2507	1.321	1.350
Máximo	1580	964	1989	1627	1350	1350	6804	3892	6.804	3.892
Media	119	119	169	169	136	136	408	408	197	197
Error típico de la media	3	3	4	4	3	3	15	13	4	3
Suma	168581	168581	448032	448032	257945	257945	589606	589606	1.464.165	1.464.165

Fuente: Elaboración Propia

Posteriormente, utilizando como insumo el ponderador base suavizado, se realiza el ajuste por falta de respuesta, el cual se detalla en el siguiente apartado.

3.2. Ponderador ajustado por falta de respuesta

En las encuestas de hogares se puede observar falta de respuesta de sus unidades por diversas causas, como por ejemplo: no se identifica la dirección, no contacto con el informante, informante cambia de domicilio, informante con dificultad física o mental, rechazo de la entrevista, etc.

En la III EME la información recabada, corresponde a los trabajadores independientes, por lo tanto la ausencia de sus respuestas debe ser corregida con la finalidad de reducir sesgos provocados por este tipo de errores no muestrales. Sin embargo, se debe señalar que la ausencia de información se corrige sólo para algunos casos, es decir cuando, el informante rechazó la entrevista; la vivienda de residencia del informante se encuentra sin moradores presentes en todas las visitas efectuadas; a la fecha de la visita el informante ha fallecido; al momento de la visita el informante se ha cambiado de domicilio o se encuentra fuera del país; al momento de la visita el informante posee dificultades físicas o mentales para contestar la encuesta; el informante no domina el idioma bajo el cual se realiza la encuesta; se impidió el acceso a la vivienda de residencia del informante (administrador, conserjes, etc. niegan el acceso a la vivienda).

Existen otras causas de no respuesta que quedan fuera del ámbito de corrección del factor de expansión, ya que corresponden a viviendas o personas sin encuestar debido a que no debieron pertenecer al marco de muestreo y por lo tanto, no debieron ser seleccionados para responder la III EME (técnicamente no elegibles). Estos casos incluyen situaciones, donde la vivienda de residencia del informante ha cambiado de estado - colectiva, de uso temporal, desocupada temporalmente, incendiada, demolida etc.- (viviendas no elegibles) o por otro lado, se identifica que los individuos fueron clasificados erróneamente como trabajadores independientes en la ENE (individuos no elegibles).

En la III EME, de un total de 6.880 viviendas seleccionadas, se seleccionaron 7.632 trabajadores independientes. De éstos, 7.420 fueron clasificados como elegibles (97,2%), de los cuales 6.758 respondieron la encuesta[17]. Por lo tanto, la tasa de respuesta de la EME, ajustada por elegibilidad, es de 91,1%.

Es posible que la falta de respuesta afecte sólo la precisión de la estimación. Sin embargo, si existe alguna relación entre las unidades faltantes y la variable de

[17] Mayor detalle ver Anexo N° 2

interés, es posible obtener estimaciones sesgadas. Por lo tanto, es recomendable realizar algún método de ajuste para compensar estas pérdidas, y mitigar dichos problemas.

El método a implementar para compensar la falta de respuesta fue el método de estratificación mediante "propensity score". De acuerdo, a lo indicado por Valliant[18], este método consiste en modelar la probabilidad de respuesta en la III EME como la realización de un proceso de variables latentes ($R_i^* = x_i^T\beta + u_i$), es decir, un conjunto de variables que inciden en la "motivación" (R^*) de participar de una unidad.

Así, mediante un conjunto de variables conocidas para quienes responden y quienes no responden se busca estimar la probabilidad de responder en la encuesta ($P(R_i^* > \theta)$). Dentro de los modelos paramétricos, se utilizan generalmente tres, los que responde a distintas características:

i. **Modelo Probit**. La probabilidad es modelada como si los valores fueran iguales a los de la función de distribución acumulada de la Normal. Por lo tanto, está bajo un supuesto de Normalidad.
ii. **Modelo Logístico**. Si bien modela la probabilidad de responder al igual que un modelo probit, la diferencia fundamental se encuentra en la función de enlace[19] (expresión matemática), que si bien es simétrica, ésta no requiere un supuesto de normalidad.
iii. **Modelo c-log-log.** La probabilidad de responder es modelada bajo la función de enlace de la distribución log-Weibull. El uso de este modelo es equivalente a suponer que el error asociado al proceso de variables latentes (u_i), tiene una distribución de valores extremos.

Cabe mencionar que para implementar el modelo probit se debiera contar con un set de variables latentes que en conjunto tengan distribución normal. Sin embargo, en la III EME, el potencial conjunto de variables (sexo, tramo etario, categoría ocupacional, nivel educacional, etc.) corresponden a variables de tipo categóricas lo que dificulta el cumplimiento de dicho supuesto. Por otro lado, para utilizar el modelo c-log-log se debiera suponer que en la III EME, el error asociado a la estimación de la probabilidad de responder – a través de un set de variables latentes - estaría explicado por un comportamiento anómalo o difícil de explicar. Como las viviendas y

[18] Mayor detalle ver Valliant, R. Drever, J. Kreuter, F. (2013, section 13.5) "Practical Tools for Designing and Weighting Survey Samples", New York. Springer.
[19] Mayor detalle ver Valliant, R. Drever, J. Kreuter, F. (2013, section 13.5, página 323) "Practical Tools for Designing and Weighting Survey Samples", New York. Springer.

personas seleccionadas ya participaron en la ENE, tanto la respuesta como el rechazo de éstos en la III EME, responden a un comportamiento más bien predecible.

De acuerdo a lo anterior y según el comportamiento de los datos de la III EME, el modelo adecuado a utilizar es el modelo logístico. Así, el método de estratificación para el ajuste de no respuesta puede ser resumido en los siguientes pasos:

1. Determinar las variables que se incluirán en el modelo de regresión logística con el cual se realizará la predicción de la probabilidad de respuesta de una persona elegible.
2. A través del modelo elegido, calcular la probabilidad de responder de cada una de las unidades elegibles que fueron utilizadas en el modelo.
3. Ordenar las unidades de menor a mayor, según la probabilidad estimada.
4. Crear los estratos o "celdas de ajustes" donde se realizarán las correcciones de no respuesta[20].

Una vez creadas las celdas de ajustes, se procede a estimar el factor de ajuste por falta de respuesta, el cual está dado por la siguiente expresión:

$$\hat{R}_c^{NR} = \frac{\sum_{k \in S_c} F_{Rjk}^{base_{tr}}}{\sum_{k \in S_{c,R}} F_{Rjk}^{base}}$$

Donde:
c: Es el subíndice de la celda de ajuste por falta de respuesta. $c = 1, \ldots, 5$
S_c: Total de independientes seleccionados y elegibles en la celda c
$S_{c,R}$: Total de independientes seleccionados en la celda c y que responde la encuesta.
F_{Rjk}^{base}: Corresponde al factor de expansión base para la persona *k,* de la vivienda *j,* en la región *R.*

Así, la expresión del ponderador de no respuesta es,

$$F_{Rjk}^{NR} = F_{Rjk}^{base_{tr}} \cdot \hat{R}_c^{NR}$$

Así, de acuerdo a la metodología antes expuesta, son 5 las celdas en las cuales se realizarán los ajustes por falta de respuesta. En el cuadro 13 se presentan las tasas de respuesta para cada una de estas celdas, así como también el factor de ajuste

[20] Mayor detalle ver Anexo N°2

por no respuesta ($\hat{R}_c^{NR}$). Se observa que el grupo 5 presenta menor tasa de respuesta, por lo que cada factor base fue "abultado" en 42% aproximadamente.

Cuadro 13. Total unidades elegibles, que responde y tasa de respuesta.

Celda Ajuste	Total Responde	Total Elegibles	Tasa de Respuesta	$\hat{R}_c^{NR}$
Total	**6.765**	**7.420**	**91,2%**	
1	1.448	1.484	97,6%	1,03
2	1.421	1.484	95,8%	1,04
3	1.390	1.484	93,7%	1,09
4	1.388	1.484	93,5%	1,13
5	1.118	1.484	75,3%	1,42

Fuente: Elaboración Propia

En el cuadro 14 presenta las principales estadísticas descriptivas del ponderador base suavizado y del ponderador ajustado por falta de respuesta. En promedio, existe un aumento de los ponderadores al ser ajustados de aproximadamente un 10%, observándose en la Macrozona Centro el mayor crecimiento promedio de los ponderadores (12%). Por otro lado, el mayor ponderador se encuentra en la Región Metropolitana el cual no supera las 4.500 unidades. En el siguiente apartado se revisará la pertinencia de un nuevo suavizamiento de los ponderadores, utilizando las mismas estrategias aplicadas para el ponderador base.

Cuadro 14. Estadísticas descriptivas del ponderador ajustado por falta de respuesta.

Estadísticas descriptivas	Macrozona									
	Norte		Centro		Sur		Región Metropolitana		Total	
	F_{Rjk}^{Tr}	F_{Rjk}^{NR}	F_{Rjk}^{Tr}	F_{Rjk}^{NR}	F_{Rjk}^{Tr}	F_{Rjk}^{NR}	F_{Rjk}^{Tr}	F_{Rjk}^{NR}	F_{Rjk}^{Tr}	F_{Rjk}^{NR}
Recuento	1.417	1.302	2.656	2.369	1.901	1.757	1.446	1.337	7.420	6.765
Moda	120	110	140	192	32	35	217	125	120	35
Mínimo	7	7	12	13	10	10	20	20	7	7
Percentil 05	18	19	27	30	27	31	66	69	26	29
Percentil 25	46	49	62	69	53	58	146	155	64	69
Mediana	88	94	114	130	93	102	240	266	119	131
Percentil 75	136	149	191	218	165	179	490	516	217	236
Percentil 95	363	386	504	559	343	396	1.298	1.351	616	667
Percentil 99	650	680	935	982	813	874	2.507	2.746	1.350	1.405
Máximo	964	1.092	1.627	1.748	1.350	1.416	3.892	4.348	3.892	4.348
Media	119	129	169	189	136	147	408	441	197	217
Error típico de la media	3,2	3,6	3,5	4,1	3,2	3,6	13,2	14,5	3,3	3,7
Suma	168.581	168.566	448.032	448.678	257.945	257.945	589.606	589.718	1.464.165	1.464.907

Fuente: Elaboración propia

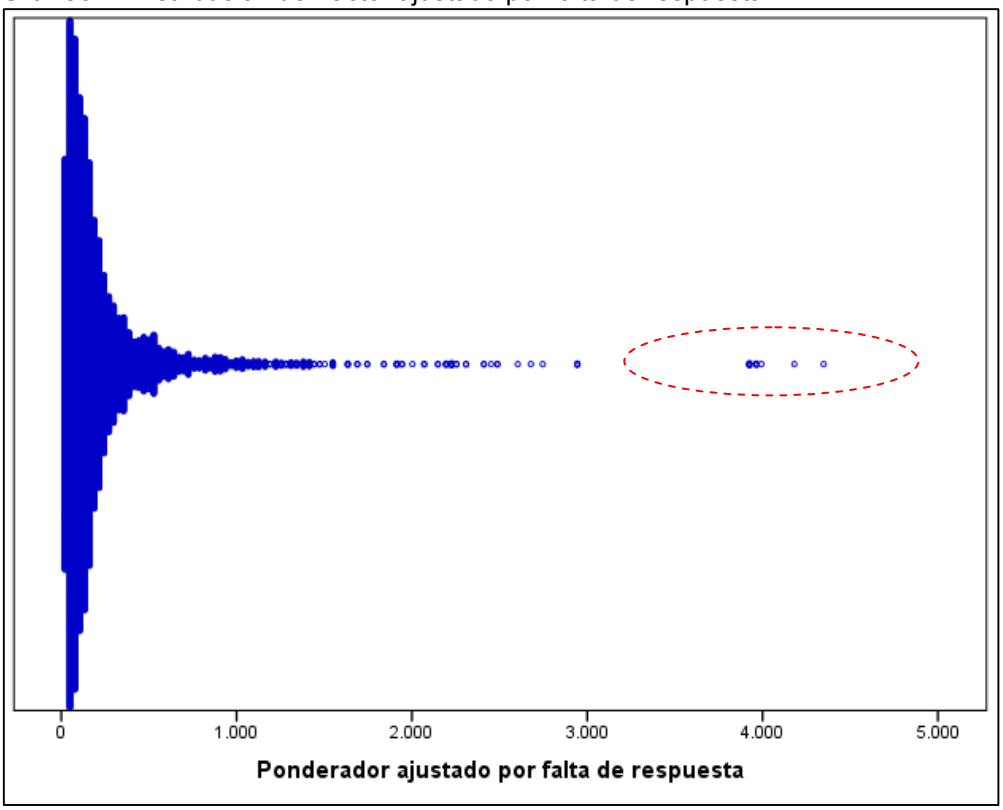
3.2.1. Suavizamiento del Ponderador ajustado por falta de respuesta

Para los ponderadores ajustados por no respuesta, se evaluó la pertinencia de realizar suavizamiento de acuerdo al último punto de corte o criterio establecido para el ponderador base, $c_6 = 0{,}10$.

En el gráfico 7 se observa que existen valores grandes para el ponderador ajustado por falta de respuesta, sin embargo se debe evaluar si, de acuerdo a los criterios establecidos, son o no valores extremos.

Si a partir del cuadro 14 se realiza el cuociente entre el valor promedio y el valor máximo observado del ponderador, por macrozona se obtiene que, cada uno de dichos valores supera el umbral 0,1[21]. En consecuencia, al interior de cada macrozona no fue necesario realizar suavizamiento.

Gráfico 7. Distribución de Factor ajustado por falta de respuesta.



Fuente: Elaboración propia

[21] Según Macrozona,: Norte= 0,119 ; Centro=0,108; Sur=0,104 ;RM=0,101

3.3. Ponderador calibrado

En general, en todas las encuestas de hogares el ponderador final o factor de expansión se encuentra calibrado, con el objetivo de alcanzar algún stock poblacional obtenido de una fuente externa a la encuesta. Por ejemplo, los factores de expansión de la Encuesta Nacional de Empleo son calibrados, cada trimestre móvil, al total de población estimado[22] por sexo y tramo de edad (menores de 15 años y 15 o más años) para cada estrato ENE, con fecha 15 de cada mes, central del período de levantamiento; mientras que la Encuesta Casen 2011 fue calibrada al stock poblacional residente en viviendas particulares ocupadas según región, con fecha 30 de noviembre. Los ponderadores de la EANNA 2012, por otro lado, fueron calibrados según sexo y tramo de edad con fecha 30 de marzo del 2012.

En los tres ejemplos expuestos anteriormente la población objetivo corresponde a personas que poseen ciertos atributos demográficos, cuantificados en los Censos de Población y Vivienda, lo que permite obtener una estimación de la población desagregada a esos niveles. Para la EME en cambio, existe un inconveniente, no existe una estimación "oficial" o de referencia, respecto a los "trabajadores independientes" (formales e informales) a nivel del país.

Por otro lado, la muestra seleccionada en la III EME está anclada a la población de referencia del trimestre MAM 2013 de la ENE, lo cual implica que la EME hace un seguimiento a los trabajadores independientes que se encontraban en ese período clasificados como trabajadores independientes, sin tomar en cuenta los flujos de entrada a esa condición laboral.

Dado lo anterior, se decidió utilizar la estimación del total de independientes del trimestre MAM 2013 de la ENE actualizada al período del trabajo de campo de la III EME. Para esto, se utilizó el crecimiento proyectado (crecimiento natural de la población según las estimaciones del CENSO 2002) para el mes central del período de levantamiento de la encuesta, es decir junio 2013. En definitiva, la estimación utilizada en la calibración del ponderador de la EME se obtuvo a través de los siguientes pasos:

1. Primero, se considera toda la información levantada para la ENE en el período MAM 2013.

[22] Estimaciones realizadas por el departamento de demografía del INE, a partir de información auxiliar.

2. Segundo, se calcula un nuevo factor de expansión, considerando las proyecciones de población a Junio del 2013.

En el período MAM 2013, la ENE utilizó el siguiente cálculo:

$$F_{hij}^{2} = \frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}} \cdot \frac{P_{hs}^{4}}{\hat{P}_{hs}}$$

Donde:

$$\hat{P}_{hs} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} p_{hijs}$$

p_{hijs}: Corresponde al total de personas de sexo y tramo de edad s, en la vivienda j, del conglomerado i, del estrato ENE h.

P_{hs}^{4}: Total de población del sexo s, del estrato ENE h, proyectado al 15 de abril de 2013.

Para obtener la estimación del total de independientes para la EME se calculó con la misma fórmula, sin embargo el stock poblacional utilizado corresponde al proyectado con fecha junio 2013. Es decir,

$$F_{hij}^{2} = \frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}} \cdot \frac{P_{hs}^{6}}{\hat{P}_{hs}}$$

En el cuadro 15, se presenta el total de independientes estimado a partir de la publicación de la ENE, período MAM 2013; y según total de personas estimado con la información levantada en MAM 2013, pero con proyecciones actualizadas a la fecha de levantamiento de la EME (en adelante I_{gs}.)

Como se observa en el cuadro 15, el total de “trabajadores independientes” estimados y publicados oficialmente son 1.850.643 personas. Sin embargo, al actualizar las proyecciones de población este total asciende a 1.855.389, lo que equivale a un incremento del 0,26% a nivel nacional, un 0,34% en la macrozona Norte, un 0,29% en el Centro, un 0,31% en el Sur y un 0,17% en la Región Metropolitana.

Cuadro 15. Total de independientes estimado a partir de la ENE- Período MAM 2013

Macrozona	Sexo	Total Independientes	
		Factor Expansión Oficial ENE – MAM	Factor Expansión Información ENE - ajustado Junio
	Hombre	1.151.400	1.154.359
Total	Mujer	699.243	701.030
	Total	**1.850.643**	**1.855.389**
	Hombre	137.299	137.784
Norte	Mujer	82.128	82.398
	Total	**219.426**	**220.182**
	Hombre	350.950	351.990
Centro	Mujer	213.139	213.728
	Total	**564.089**	**565.719**
	Hombre	236.767	237.508
Sur	Mujer	129.960	130.372
	Total	**366.727**	**367.879**
	Hombre	426.384	427.077
Metropolitana	Mujer	274.017	274.532
	Total	**700.400**	**701.609**

Fuente: Elaboración propia

Finalmente, el ponderador calibrado, se le asigna a cada una de las personas entrevistadas en la EME. El procedimiento de cálculo de este ponderador se resume en tres pasos:

1. Estimar el total de trabajadores independientes según sexo, para cada macrozona a partir de la EME 2013. Es decir, se estimó el total de independientes a través de la utilización del ponderador de no respuesta, tal como se muestra a continuación:

$$\hat{P}_{gs} = \sum_{j=1}^{m_g} \sum_{k=1}^{p_g} F_{Rjk}^{NR} \cdot p_{jks} \quad \begin{matrix} g = 1,2,3,4 \\ s = 1,2 \end{matrix}$$

Donde:

g: Subíndice de la macrozona de procedencia de las unidades.
p_g: Número de personas independientes entrevistadas en la vivienda g.
m_g: Número de viviendas entrevistadas en la macrozona g.

$$p_{jks} = \begin{cases} 1, & si\ persona\ k\ es\ sexo\ s \\ 0, & en\ otro\ caso \end{cases}$$

2. Construir el ajuste a la población total, mediante la razón entre la estimación del total de independientes de acuerdo a fuentes externas (ENE), y la estimación de la encuesta obtenida en el paso (1):

$$\hat{R}_{gs} = \frac{I_{gs}}{\hat{P}_{gs}}$$

3. Construir el Factor de Expansión final, o Ponderador Calibrado, como el producto entre el ponderador ajustado por falta de respuesta con el ajuste a la población total, calculado en el paso 2.

$$F_{gjs}^{cal} = F_{Rjk}^{NR} \cdot \hat{R}_{gs}$$

Al usar el ponderador calibrado, se debe tener en consideración que éste expande al total de "trabajadores independientes", de sexo *s* y residentes en la macrozona *g*, estimados a partir de la Encuesta Nacional de Empleo, en el trimestre móvil MAM 2013, actualizado al crecimiento poblacional de junio 2013 - mes central de levantamiento de EME.

En el cuadro 17 se observa un incremento en los ponderadores, y por tanto un aumento en los casos más extremos. Por ejemplo, en la Región Metropolitana un trabajador independiente, mujer, representaba a 3.994 personas, sin embargo al ajustar según sexo y macrozona, esta persona representa 5.106 individuos. Cabe señalar, que en el caso de un hombre de la misma macrozona su ponderador cambió de 4.348 a 4.953, lo cual puede ser revisado en el cuadro 16.

Cuadro 16. Estadísticas descriptivas del ponderador ajustado por falta de respuesta y calibrado a stock de independientes, según sexo.

Estadísticas descriptivas	Sexo					
	Hombre		Mujer		Total	
	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}
Recuento	4.143	4.143	2.622	2.622	6.765	6.765
Moda	60	90	92	112	35	90
Mínimo	7	10	7	9	7	9
Percentil 05	30	40	27	34	29	38
Percentil 25	72	97	67	85	69	91
Mediana	134	177	126	160	131	169
Percentil 75	241	314	231	293	236	307
Percentil 95	667	834	660	843	667	837
Percentil 99	1.416	1.784	1.332	1.702	1.405	1.754
Máximo	4.348	4.953	3.994	5.106	4.348	5.106
Media	220	279	211	267	217	274
Error típico de la media	5	6	6	7	4	4
Suma	912.353	1.154.359	552.555	701.030	1.464.907	1.855.389

Fuente: Elaboración propia

Cuadro 17. Estadísticas descriptivas del ponderador ajustado por falta de respuesta y calibrado a stock de independientes, según macrozona.

Estadísticas descriptivas	Macrozona									
	Norte		Centro		Sur		Región Metropolitana		Total	
	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}
Recuento	1.302	1.302	2.369	2.369	1.757	1.757	1.337	1.337	6.765	6.765
Moda	110	32	192	40	35	90	125	128	35	90
Mínimo	7	9	13	16	10	15	20	23	7	9
Percentil 05	19	25	30	38	31	43	69	84	29	38
Percentil 25	49	65	69	86	58	81	155	185	69	91
Mediana	94	123	130	165	102	146	266	311	131	169
Percentil 75	149	195	218	276	179	257	516	612	236	307
Percentil 95	386	514	559	705	396	569	1.351	1.581	667	837
Percentil 99	680	887	982	1.245	874	1.153	2.746	3.329	1.405	1.754
Máximo	1.092	1.494	1.748	2.186	1.416	2.116	4.348	5.106	4.348	5.106
Media	129	169	189	239	147	209	441	525	217	274
Error típico de la media	3,6	4,7	4,1	5,1	3,6	5,1	14,5	17,3	3,7	4,5
Suma	168.566	220.182	448.678	565.718	257.945	367.880	589.718	701.609	1.464.907	1.855.389

Fuente: Elaboración Propia

Además se observa que la Región Metropolitana posee la mayor variabilidad en sus ponderadores, así como también los trabajadores independientes hombres. Sin embargo, el valor mayor registrado es para un trabajador independiente mujer. En el siguiente apartado se revisará la pertinencia de realizar suavizamiento a los ponderadores calibrados.

3.3.1. Suavizamiento de Ponderador Calibrado

Al analizar la existencia de valores extremos de acuerdo a los criterios establecidos para el ponderador base, se concluyó que bajo el criterio más estricto (factor de expansión es 5 veces o más el factor promedio de la macrozona), 78 observaciones debería ser truncadas, mientras que bajo el criterio más conservador y el utilizado en el primer suavizamiento, no debería aplicarse ningún suavizamiento, como se ilustra en el cuadro 18.

Cuadro 18. Número de observaciones a truncar según criterio o punto de corte

	Valor extremo		
Punto corte	NO	Si	Total
C_1	6.687	78	6.765
C_2	6.720	45	6.765
C_3	6.745	20	6.765
C_5	6.758	7	6.765
C_6	6.765	0	6.765

Fuente: Elaboración Propia

Ante estos resultados se implementaron cuatro suavizamientos y se comparó el resultado en aquellos ponderadores truncados, mediante gráficos y estimaciones del ECM.

Al revisar los gráficos 8 y 9 se puede observar que ninguno de los criterios es apropiado, ya que al momento de redistribuir los pesos “sobrantes”, otros ponderadores superan el umbral. Esto sucedió con cada uno de los umbrales analizados.

En consecuencia, de acuerdo a estos resultados y en línea con el criterio establecido para el ponderador base y para el ajustado por falta de respuesta, en esta etapa no se realizó suavizamiento de los factores de expansión.

Gráfico 8. Dispersión entre factor calibrado y ajustado según criterio 1.

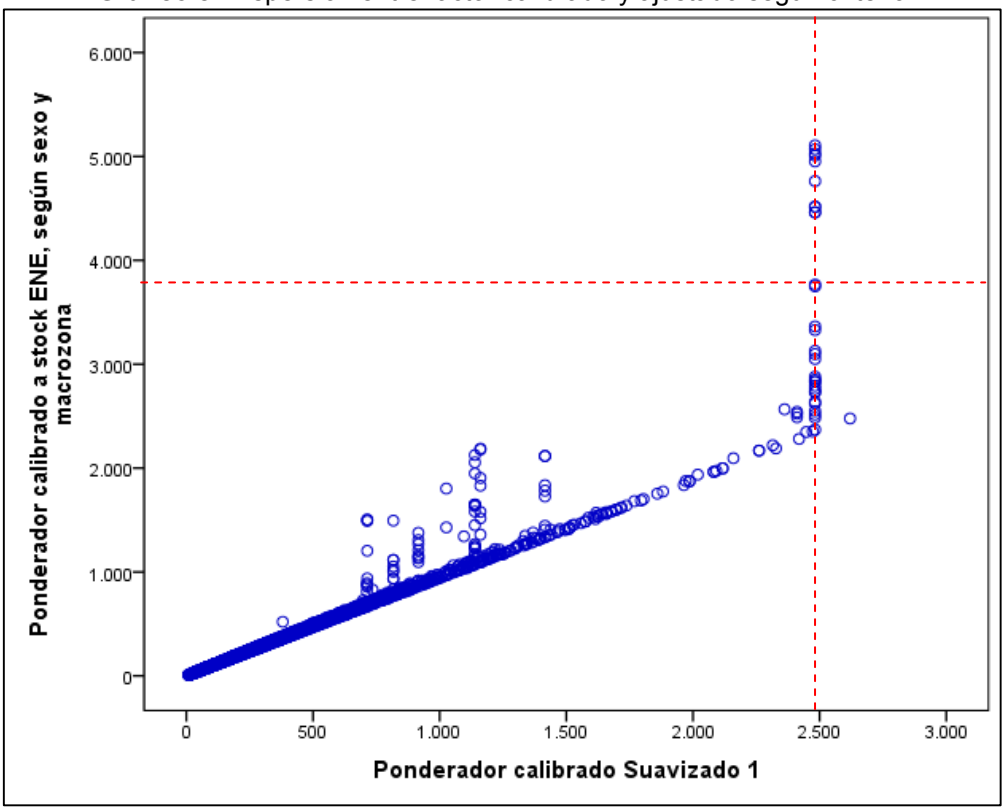
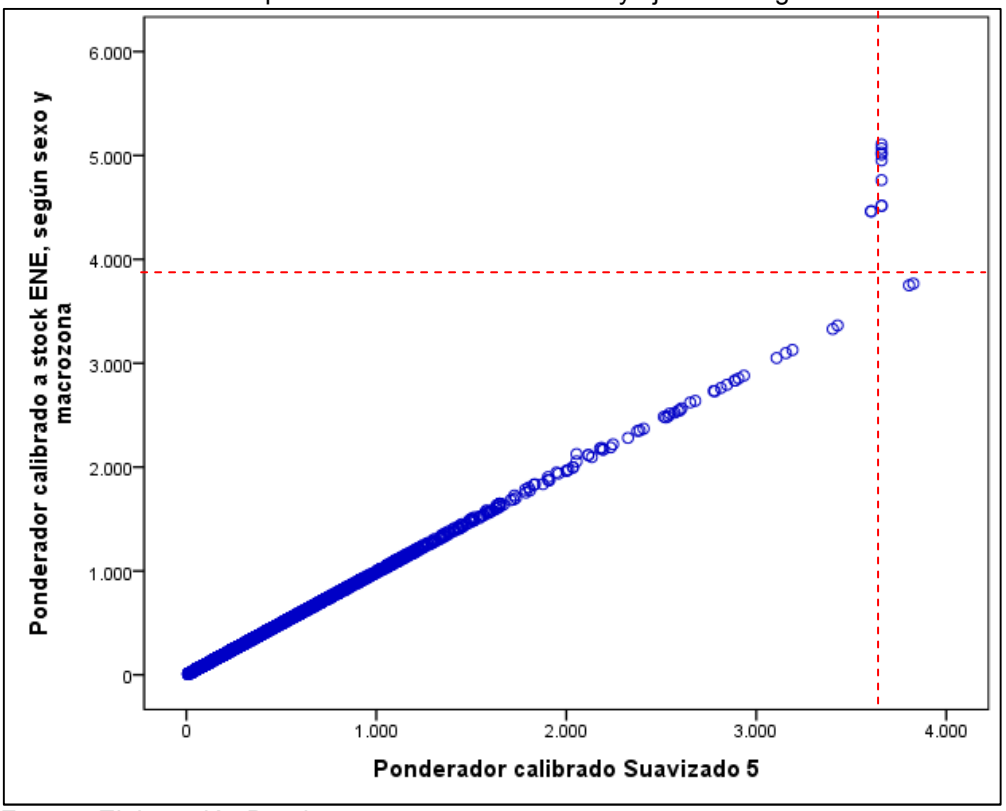


Gráfico 9. Dispersión entre factor calibrado y ajustado según criterio 5.



Fuente: Elaboración Propia

4. ESTIMACIÓN DE VARIANZA

De acuerdo a lo descrito en los apartados anteriores, el diseño muestral de la III EME es bifásico y complejo, por lo tanto las probabilidades de selección de los individuos son desiguales. Así, cualquier análisis que se desee realizar a partir de la III EME, se debe hacer utilizando el factor de expansión. Si no se usa el factor de expansión se obtendrán estimaciones sesgadas y que sólo darán cuenta del comportamiento de las unidades seleccionadas, pero no de la población total.

Por otro lado, al momento de realizar un estudio, se sugiere a los analistas, estimar la variabilidad muestral asociada a la estimación puntual. Para ello, existen diversos paquetes estadísticos en STATA (svyset), SPSS (csplan, en muestras complejas), R (Survey, svydesign), SAS (PROC surveyfreq, Proc surveymeans), etc. que utilizan fórmulas convencionales (aun cuando muchas veces éstas no tienen una fórmula explícita) para la estimación de las varianzas, bajo un supuesto de muestreo aleatorio con reemplazo con ponderadores, lo que facilita los cálculos.

Para la utilización de los paquetes estadísticos de forma apropiada, se requiere identificar las variables que definen el diseño muestral de la encuesta. En este contexto, en los siguientes apartados se exponen las variables que identifican el diseño muestral, así como también su implementación en Spss y Stata. Para ello, se definió como variable de análisis principal la estructura de la rama de actividad de los trabajadores independientes. Como existen algunas categorías como pesca; electricidad, gas y agua; entre otras, en las cuales se observa una baja prevalencia, se crea una variable más agregada denominada “rama reducida”. Es sobre esta variable que en la sección 4.2 se realizan las estimaciones de los errores.

4.1. Variables que identifican el diseño

El diseño muestral de la III EME posee las características de un diseño muestral bifásico y complejo. La primera fase se caracteriza por poseer un diseño muestral complejo, pues es estratificado geográficamente y la selección de las viviendas que participan en la ENE se realizó en dos etapas, seleccionando en primera instancia los conglomerados de forma sistemática y con probabilidad proporcional al tamaño, mientras que las viviendas en su interior fueron seleccionadas de forma sistemática pero con igual probabilidad. La segunda fase se caracteriza porque las viviendas se seleccionaron a partir de un listado de viviendas de la ENE tal que en su interior residen al menos un trabajador independiente, luego en su interior se seleccionaron

sistemáticamente, tantos trabajadores independientes como actividades únicas se identifican en su interior.

En este contexto las variables que identifican el diseño muestral de ambas fases, corresponden a una variable llamada “Estrato” que identifica los estratos geográficos de la ENE; y una variable “IdDirectorio” que corresponde a una variable ficticia que identifica de forma única los conglomerados en la ENE. Las variables que identifican el diseño de la III EME, corresponden a las mismas de la ENE, ya que la selección de las unidades se realizó al interior de cada región de forma independiente, por lo tanto, como por construcción los estratos de la ENE no combina regiones, entonces el cruce Región versus Estrato da como resultado los mismos estratos de la ENE.

En general, cuanto más complejo es el diseño muestral bajo el cual se implementa una encuesta, más complejo se vuelve la forma de determinar los errores muestrales. Tanto así, que no existen fórmulas exactas y/o explícitas para esto. Sin embargo, paquetes estadísticos en software especializados, facilitan los cálculos a través de aproximaciones realizadas mediante distintos modelos o métodos de estimación, para lo cual se debe identificar las variables que definen el diseño muestral (estratos, conglomerados) y el factor de expansión apropiado (considerando todos los ajustes pertinentes).

En ocasiones pueden existir algunas dificultades en la implementación de la estimación de los errores mediante un paquete estadístico, originadas por las características del diseño muestral, por ejemplo: más de una fase de muestreo; muestreo multietápico de las unidades muestrales, selección de unidades sin reemplazo, estratos de muestreo con sólo una unidad primaria con unidades elegibles; variabilidad de los tamaños de los conglomerados.

En el caso de la III EME, se observan principalmente tres dificultades: (1) Diseño muestral bifásico y complejo; (2) existen estratos de muestreo (los de la ENE) que poseen solo un conglomerado (manzana o sección); (3) el número de unidades seleccionadas y que responde en cada conglomerado es desigual y muy variable. A fin de minimizar los problemas señalados anteriormente, y siguiendo las recomendaciones internacionales[23], los errores fueron estimados a partir de modelos que buscan dar cuenta, lo más fielmente posible del diseño muestral. Para ello se agruparon estratos y conglomerados a fin de que estos nuevos pseudo-estratos y pseudo-conglomerados, garanticen la estimación de varianzas en cada

[23] Ver Capítulo 15.5 en Valliant *et al.* (2013).

nuevo estrato, y de ésta forma no subestimar los errores. A continuación se detallan los procedimientos y criterios utilizados en la creación de dichas variables.

4.1.1. Creación de pseudo-estratos

Los estratos ficticios o pseudo-estratos son construidos con el objetivo de corregir los problemas generados por la existencia de estratos con solo un conglomerado (estratos unitarios), esto es subestimar la varianza de cualquier variable de interés.

Los pseudo-estratos son construidos a través de la agrupación de dos o más estratos originales, los que pueden ser unitarios o no, de acuerdo a un patrón u ordenamiento jerárquico de variables geográficas o de tamaño, de modo que estos contengan al menos dos conglomerados, los que a su vez deberán contener al menos 15 unidades que responden en su interior.

A continuación se detalla el procedimiento de construcción de los pseudo-estratos;

i. Primero se contabiliza, al interior de cada estrato original, el total de individuos que participa en la encuesta. Si el estrato contiene menos de 30 (2•15) unidades entonces deberá ser colapsado con otro.
ii. Se ordenan todos los estratos, geográficamente, de acuerdo a la división político administrativa en urbanos y rurales, y luego al interior de cada región según ordenamiento del estrato.
iii. Finalmente, al interior de la misma área geográfica y región se colapsan aquellos estratos con menos de 30 unidades, lo más cercano geográficamente, pero sin que en conjunto estos superen las 60 unidades.

De un total de 160 estratos que posee la ENE, en la III EME se seleccionaron unidades desde 159 estratos, de los cuales 3 de ellos contienen unidades seleccionadas en solo un conglomerado. Sin embargo, existen 80 estratos con 30 o menos unidades (personas). Así el total de pseudo-estratos creados desciende a 102 unidades.

Cuadro 19. Total de estratos y de pseudo-estratos, según macrozona.

Macrozona	Estratos	Pseudo-estrato
Total	**159**	**102**
Norte	25	18
Centro	62	41
Sur	25	19
Metropolitana	47	24

Fuente: Elaboración propia

4.1.2. Creación de pseudo-conglomerados

Los conglomerados ficticios o pseudo-conglomerados fueron construidos con el objetivo de reducir los problemas generados a causa de la diversidad de tamaños de los conglomerados (número de unidades que participa en ellos), pues a mayor variabilidad en el tamaño de los conglomerados, la varianza de los estimadores tiende a incrementarse y volverse más inestable.

Los pseudo-conglomerados fueron creados a partir de un ordenamiento jerárquico, según comuna y total de unidades que responde, al interior de cada pseudo-estrato. Luego, se unieron los conglomerados a fin que estos en conjunto reunieran 15 unidades aproximadamente.

A continuación se detalla el procedimiento de construcción de los pseudo-conglomerados;

i. Primero se contabiliza, al interior de cada conglomerado original, el total de individuos que participa en la encuesta. Si el conglomerado contiene menos de 15 unidades entonces deberá ser colapsado con otro.
ii. Se ordenan todos los conglomerados geográficamente según área (urbana o rural); región, provincia y comuna (RPC); y total de unidades que responde, al interior de cada pseudo-estrato.
iii. Finalmente, al interior de cada pseudo-estrato se colapsan aquellos estratos con menos de 15 unidades, los más cercanos geográficamente, pero sin que en conjunto estos superen las 30 unidades.

La ENE posee un total de 4.126 conglomerados, en 2.837 de estos conglomerados se seleccionaron trabajadores independientes, los que se transformaron en 341 pseudo-conglomerados.

Con el objetivo que la unión de conglomerados y estratos no se crucen según sea el área, urbana o rural, es que se dejaron 2 pseudo-conglomerados con 13 unidades y 13 con 14 unidades. El máximo de unidades que se reporta en un pseudo-conglomerado es de 25 unidades, en un único caso.

En el cuadro 20 se expone el total de conglomerados y pseudo-conglomerados según macrozona.

Cuadro 20. Total de conglomerados y de pseudo-conglomerados, según macrozona

Macrozona	Conglomerados	Pseudo-Conglomerados
Total	**2.837**	**341**
Norte	549	68
Centro	1.102	120
Sur	597	86
Metropolitana	589	67

Fuente: Elaboración propia

4.2. Estimación de variables y varianzas en Spss y Stata

Diversos paquetes estadísticos poseen algoritmos que permiten la estimación de los errores muestrales bajo diseños muestrales complejos a través de métodos como, el método de linearización de Taylor; métodos de replicación repetido (Jackknife, Bootstrap), entre otros. Sin embargo, para que éstos sean más simples de implementar se deben realizar algunos supuestos: se asume que la selección de las unidades, en las distintas etapas, se realizó de forma independiente y con reemplazo (esto simplifica los cálculos y las expresiones matemáticas); por otro lado, aun cuando el diseño muestral de la encuesta posea muchas etapas sólo se da cuenta de la primera etapa, pues es esta la que aporta la mayor variabilidad al error total.

Tanto en Stata como en Spss, previo a la estimación de la variable en estudio y los errores asociados a ella, se debe definir el diseño muestral bajo el cual se realizarán las estimaciones. Las variables, que se encuentran en la base de datos y que definen el diseño muestral de la III EME son:

i. Fact_EME: corresponde al factor de expansión que da cuenta de las probabilidades de selección, de la fase 1 y 2, ajuste por falta de respuesta y calibración.
ii. Pseudo-estrato: variable que identifica el estrato de muestreo, tal que éste contiene al menos dos conglomerados, para garantizar la estimación de la varianza.
iii. Pseudo-conglomerado: variable que identifica el conglomerado, tal que éste contiene al menos 15 unidades aproximadamente.

Así, para revisar la estructura de la actividad en la cual se desenvuelven los trabajadores independientes, previamente, el investigador debiera hacer lo siguiente:

i. Determinar y construir la variable de interés, si ésta no está definida.
ii. Especificar las variables que definen el diseño complejo
iii. Realizar la estimación correspondiente

Considerando la estructura de la rama de actividad económica (CIIU rev. 3) para los trabajadores independientes como la variable de interés -se observa la existencia de categorías en las que la proporción de trabajadores independientes observados es pequeña, lo que conlleva a obtener estimaciones con gran variabilidad o error

muestral-, por lo cual se agruparon las categorías de baja prevalencia en dos grandes grupos, dando origen a una nueva variable denominada “rama de actividad reducida”, según como se indica en la cuadro 21.

Cuadro 21. Rama de actividad económica según CIIU Rev 3. vs Rama de actividad reducida

Rama de actividad Económica	Rama de actividad Económica Reducida
A. Agricultura, ganadería, caza y silvicultura	A. Agricultura, ganadería, caza y silvicultura
B. Pesca	Sector Primario
C. Explotación de minas y canteras	Sector Primario
D. Industrias manufactureras	D. Industrias manufactureras
E. Suministro de electricidad, gas y agua	Sector Primario
F. Construcción	F. Construcción
G. Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	G. Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres
H. Hoteles y restaurantes	Servicios
I. Transporte, almacenamiento y comunicaciones	I. Transporte, almacenamiento y comunicaciones
J. Intermediación financiera	Servicios
K. Actividades inmobiliarias, empresariales y de alquiler	K. Actividades inmobiliarias, empresariales y de alquiler
L. Administración pública y defensa; planes de seguridad social de afiliación obligatoria	Servicios
M. Enseñanza	Servicios
N. Servicios sociales y de salud	Servicios
O. Otras actividades de servicios comunitarios, sociales y personales	O. Otras actividades de servicios comunitarios, sociales y personales
P. Hogares privados con servicio doméstico	Servicios

Fuente: Elaboración Propia

A continuación se presenta un resumen con la estimación de la rama de actividad reducida, en la cual fueron clasificados los trabajadores independientes, según las estimaciones realizadas en Stata y Spss[24].

[24] Ver Anexo N°4

Cuadro 22. Estructura de la Actividad económica en la cual se desenvuelven los trabajadores independientes- estimación realizada en SPSS

Rama de actividad económica	Estimación	Error típico	Intervalo de confianza al 95%		Coeficiente de variación	Efecto del diseño
			Inferior	Superior		
A. Agricultura, ganadería, caza y silvicultura	10,5%	0,6%	9,4%	11,9%	6,0%	2,698
D. Industrias manufactureras	13,0%	0,9%	11,4%	14,8%	6,6%	4,114
F. Construcción	9,3%	0,5%	8,3%	10,3%	5,4%	1,886
G. Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	33,8%	1,1%	31,7%	36,0%	3,2%	3,449
I. Transporte, almacenamiento y comunicaciones	8,3%	0,5%	7,4%	9,3%	5,7%	1,900
K. Actividades inmobiliarias, empresariales y de alquiler	8,3%	0,7%	7,0%	9,7%	8,3%	3,919
O. Otras actividades de servicios comunitarios, sociales y personales	6,8%	0,7%	5,5%	8,3%	10,3%	4,999
Sector Primario	1,9%	0,3%	1,5%	2,5%	13,1%	2,170
Servicios	8,0%	0,7%	6,8%	9,4%	8,2%	3,769
No responde	0,0%	0,0%	0,0%	0,1%	71,0%	0,470

Fuente: Elaboración Propia

Cuadro 23. Estructura de la Actividad económica en la cual se desenvuelven los trabajadores independientes- estimación realizada en Stata

Rama reducida	**Proporción**	**Std. Err.**	Intervalo de confianza al 95%	
			Inferior	Superior
A. Agricultura, ganadería, caza y silvicultura	10,5%	0,6%	9,3%	11,8%
D. Industrias manufactureras	13,0%	0,9%	11,3%	14,7%
F. Construcción	9,3%	0,5%	8,3%	10,2%
G. Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	33,8%	1,1%	31,7%	36,0%
I. Transporte, almacenamiento y comunicaciones	8,3%	0,5%	7,4%	9,3%
K. Actividades inmobiliarias, empresariales y de alquiler	8,3%	0,7%	6,9%	9,6%
O. Otras actividades de servicios comunitarios, sociales y personales	6,8%	0,7%	5,4%	8,2%
Sector Primario	1,9%	0,3%	1,4%	2,4%
Servicios	8,0%	0,7%	6,7%	9,3%
No responde	0,0%	0,0%	0,0%	0,0%

Fuente: Elaboración Propia

Como se puede apreciar en los cuadros 22 y 23, no existen diferencias en términos de estructura (estimación de la proporción), ni tampoco en términos del error estándar o típico, al realizar estimaciones a través de Spss o Stata. Sin embargo, se observa que existe una pequeña diferencia en los intervalos de confianza. En ambos casos fue obtenido con un 95% de confianza, pero la distribución que utilizan es diferente. Spss por su parte, hace el supuesto que el promedio de los estimadores siguen una distribución normal estándar, por lo tanto utilizan el percentil de la distribución normal, es decir 1,96 aproximadamente. Mientras que Stata utiliza el supuesto de una distribución con colas más pesadas, la T-Student, donde los grados de libertad (gl) dan cuenta del diseño muestral, y estos son calculados como la diferencia entre el número de conglomerados y el número de estratos[25]. A nivel nacional la III EME posee $gl = 341 - 102 = 239$. Luego, en la construcción del intervalo de confianza en Stata se utilizó el percentil de la distribución T-Student con 239 grados de libertad, es decir 1,969939.

Respecto a la estructura de la rama de actividad de los trabajadores independientes, se observa que éstos se concentran principalmente, en Comercio, seguido de Industria Manufacturera y Agricultura, actividades que en conjunto reúnen a más del 55% de los trabajadores independientes.

[25] Un análisis más detallado lo puede encontrar en *Applied Survey Data Analysis*.

BIBLIOGRAFÍA

1. Valliant, R. Drever, J. Kreuter, F. (2013). Practical Tools for Designing and Weighting Survey Samples", Springer, New York.

2. Heeringa, S., West, B., and Berglund, P. (2010). Applied Survey Data Analysis. Chapman and Hall, CRC Press, Boca Raton, Florida

3. Dobson, A. (2002) An Introduction to Generalized Linear Models. CRC Press.

4. Burgueño, M; García-Bastos, J; González-Buitrago, J.(1993). Las curvas ROC en la evaluación de pruebas diagnósticas. Medicina Clínica Vol. 104. Núm. 17.1.995. España.

5. Montgomery. D; Peck, E; Vining, G. (2006). Introducción al Análisis de Regresión Lineal. 1era Edición español. Cía. Editorial continental. México.

6. The American Association for Public Opinion Research (2011). Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys.

ANEXOS

1. Anexo N°1. Áreas de Difícil acceso o Alto Costo

Cuadro 24. Áreas geográficas excluidas del Marco de Muestreo del INE, clasificadas como ADA's.

Región	Nombre Provincia	Nombre Comuna	Total Viviendas Censo 2002
Arica y Parinacota	Parinacota	General Lagos	447
Tarapacá	Tamarugal	Colchane	1.395
Antofagasta	El Loa	Ollagüe	287
Valparaíso	Valparaíso	Juan Fernández	257
	Isla de Pascua	Isla de Pascua	1.416
Los Lagos	Llanquihue	Cochamó	1.676
	Palena	Chaitén	2.305
		Futaleufú	853
		Hualaihué	2.553
		Palena	760
Aisén del General Carlos Ibáñez del Campo	Coihaique	Lago Verde	590
	Aisén	Guaitecas	463
	Capitán Prat	O'Higgins	249
		Tortel	187
Magallanes y de La Antártica Chilena	Magallanes	Laguna Blanca	267
		Río Verde	197
		San Gregorio	603
	Antártica Chilena	Cabo de Hornos (Ex - Navarino)	626
		Antártica	24
	Tierra el Fuego	Primavera	459
		Timaukel	172
	Última Esperanza	Torres del Paine	260
Total Viviendas ADA's			**16.046**

Fuente: Elaboración propia

2. Anexo N°2. Códigos de disposición última visita

En el cuadro 25, aparece el código de disposición de las viviendas en su última visita. Así, las categorías que en la variable “elegible” dice sí, corresponden a unidades elegibles y sobre las cuales se realizan los ajustes por falta de respuesta; las restantes unidades fueron clasificadas como no elegibles. Cabe señalar que aquellas unidades no legibles no son contabilizadas en el ajuste por falta de respuesta.

Cuadro 25. Códigos de disposición final de la última visita a la vivienda

Código de disposición de la última visita al hogar	Frecuencia	Porcentaje	Elegible
11. Entrevista lograda	6758	89,50%	Sí
12. Entrevista lograda de forma parcial	7	0,090%	Sí
21. Se rechazó la entrevista	118	1,54%	Sí
22. Vivienda ocupada sin moradores presentes	174	2,27%	Sí
23. Informante inubicable por cambio de domicilio o fuera del país	339	4,44%	Sí
24. Muerte del informante	5	0,06%	Sí
25. Informante con dificultad física, mental o cognitiva para contestar	19	0,24%	Sí
26. Problemas del Idioma	0	0,00%	Sí
27. Fuera de muestra	189	2,47%	No
31. Se impidió acceso a la vivienda(administrador, conserje o junta de vigilancia niega acceso)	0	0,00%	Sí
32. No fue posible localizar la dirección	2	0,02%	No
33. Área de difícil acceso o peligrosa	3	0,03%	No
41. Inmueble para uso no habitacional (empresa, oficina, vivienda colectiva, institución pública, etc.)	1	0,01%	No
42. Vivienda en demolición, incendiada, destruida o erradicada	0	0,00%	No
43. Vivienda particular desocupada (en arriendo, en venta, otro.)	16	0,20%	No
44. Vivienda de uso temporal (vacaciones, descanso, etc.)	1	0,01%	No
Total	**7632**	**100,00%**	**7420**

Fuente: Elaboración propia

3. Anexo N°3. Regresión logística implementada en la construcción de celdas para ajustes de no respuestas

Para la selección del mejor modelo que permita estimar la probabilidad de responder de un trabajador independiente seleccionado para participar en la III EME, se consideraron tres análisis de elegibilidad; 1) Descriptivo, 2) Modelación y 3) Sensibilidad del modelo. El objetivo del análisis descriptivo fue tener una primera aproximación de las variables que influyeron en la respuesta de las personas y de esta manera entender de forma intuitiva nuestro fenómeno de estudio. Luego, para la modelación de la variable de respuesta, se seleccionaron un conjunto de variables que permitieran ajustar mejor la respuesta de interés, para así llegar a la selección del modelo ideal. Finalmente, en la etapa de Sensibilidad del modelo se determinará qué "tan bueno" es nuestro ajuste, específicamente a través de la Curva ROC.

3.1. Especificación del Modelo

Dado que nuestra variable de interés tiene dos categorías provenientes de una respuesta binaria (Responde vs No responde), se utiliza un modelo que considera esta característica a medir. Los modelos ampliamente usados para estudiar este fenómeno, están dentro de una clase mayor de modelos llamados modelos lineales generalizados. Primeramente, se define la variable aleatoria binaria como:

$$Y_i = \begin{cases} 1, & si\ la\ i-ésima\ persona\ responde\ dado\ pertenece\ a\ una\ unidad\ elegible \\ 0, & si\ la\ i-ésima\ persona\ no\ responde\ dado\ pertenece\ a\ una\ unidad\ elegible \end{cases} \quad (1)$$

Con $P(Y=1)=\pi$ y con $P(Y=0)=1-\pi$. Si hay n variables aleatorias $Y_1, Y_2, \dots, Y_n$, independientes entre sí, con $P(Y_i=1)=\pi_i,\ \forall\ i=1,\dots,n$, entonces su función de probabilidad conjunta es:

$$\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = exp\left[\sum_{i=1}^{n} y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} log(1-\pi_i)\right] \quad (2)$$

La cual es miembro de la familia exponencial.

Al considerar la siguiente función de enlace[26]:

[26] Nuestro interés es modelar $E(Y_i)=\pi_i$ con, $\pi_i \in [0,1]$, a través, de $\boldsymbol{x}_i^t\beta$. Sin embargo, no existe una relación lineal entre π_i y $\boldsymbol{x}_i^t\beta$, tal que $E(Y_i)=\pi_i=\boldsymbol{x}_i^t\beta$, por lo general esta

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^t\beta \quad (3)$$

Con $\boldsymbol{x}_i^t = \left(1, x_1, x_2, \dots, x_p\right)^t$ y $\beta = \begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{pmatrix}$, tal que, $\boldsymbol{x}_i^t\beta = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$.

Se tiene que la probabilidad del suceso es:

$$\pi_i = \frac{exp\left(\boldsymbol{x}_i^t\beta\right)}{1 + exp\left(\boldsymbol{x}_i^t\beta\right)} = P\left(Y_i = 1/\boldsymbol{x}_i^t\beta\right) \quad (4)$$

La estimación de los parámetros se realiza mediante un proceso iterativo que aproxima la log-verosimilitud mediante el algoritmo de Newton Raphson o por aproximación Scoring de Fisher. A continuación, se detallan los pasos para la obtención de estos parámetros.

3.2. Estimación de Parámetros

3.2.1. Estimación Máxima verosimilitud

Sean $Y_1, \dots, Y_n$ n variables aleatorias independientes, es decir, cada una con función de densidad de probabilidad $f_i(y_i;\theta)$ donde el vector de parámetro $\theta = \left(\theta_1, \dots, \theta_p\right)^t$ es un elemento del espacio paramétrico Ω que comprende todos los valores a priori admisibles.

---

relación es de tipo no lineal. Para resolver esto, se necesita una función g que relacione la respuesta media con los regresores a estimar, es decir, $g(\pi_i) = \boldsymbol{x}_i^t\beta$, de tal forma que, $E(Y_i) = \pi_i = g^{-1}(\pi_i)$, entonces, se dice que g es una función de enlace. Ahora bien, si Y_i se puede expresar de forma general como $f(y;\pi) = \exp\left[a(y)b(\pi) + c(\pi) + d(y)\right]$, se dice que Y_i pertenece a la familia exponencial. Además, si $a(y) = y$ se dice que la distribución es de la forma canónica (o, estándar) y $b(\pi)$ se llama el parámetro natural de la distribución. Nuestra variable de interés sigue una distribución binomial, es decir, $Y_i \sim Binomial(1, \pi_i)$, se sabe que esta variable aleatoria pertenece a la familia exponencial con parámetro natural $b(\pi_i) = \log\left(\pi_i/1-\pi_i\right)$ y eso nos permite tomar este parámetro natural como función de enlace para $\boldsymbol{x}_i^t\beta$, de tal forma que, $\log(\pi_i/1-\pi_i) = \boldsymbol{x}_i^t\beta$. Finalmente, nuestro modelo a estimar es $Y_i \sim Binomial\left(1, \frac{exp(\boldsymbol{x}_i^t\beta)}{1+exp(\boldsymbol{x}_i^t\beta)}\right)$. [Para mayor detalle consultar Dobson (2002)]

La distribución de densidad conjunta de n observaciones independientes $\boldsymbol{y} = (y_1, \dots, y_n)^t$ es:

$$f(y;\theta) = \prod_{i=1}^{n} f_i(y_i;\theta) = L(\theta, \boldsymbol{y}). \quad (5)$$

La expresión $L(\theta, \boldsymbol{y})$ es vista como una función del vector de parámetro desconocido θ dada la muestra $\boldsymbol{y}$ (o datos), denominada función de verosimilitud. A menudo, se trabaja con el logaritmo natural de la función de verosimilitud, llamada función de log - verosimilitud:

$$logL(\theta;\boldsymbol{y}) = \sum_{i=1}^{n} log\, f_i(y_i;\theta). \quad (6)$$

Para encontrar el conjunto de soluciones para el vector de parámetro θ, dada la muestra $\boldsymbol{y}$, que maximice la función de verosimilitud o log verosimilitud, consideramos el principio de máxima verosimilitud que postula la elección de $\hat{\theta}$ perteneciente al espacio paramétrico Ω que maximice la función de log-verosimilitud. Para esto, se define el estimador máximo verosímil como $\hat{\theta}$ tal que:

$$logL\left(\hat{\theta};\boldsymbol{y}\right) \geq L(\theta, \boldsymbol{y}), \forall\, \theta. \quad (7)$$

3.2.1.1. Vector Score

Una forma clásica de encontrar los estimadores máximo verosímil es derivar la función de log – verosimilitud respecto a θ. El procedimiento que calcular la primera derivada de la función log-verosimilitud es llamada la función score de Fisher y es denotada por:

$$u(\theta) = \frac{\partial}{\partial\theta} logL(\theta, \boldsymbol{y}). \quad (8)$$

Se debe notar que el vector de score es un vector de las primera derivada parcial, para cada uno de los elementos de θ[27].

[27] Dado que la transformación logarítmica es una función monótona, esta es apropiada para maximizar $L(\theta, \boldsymbol{y})$ en lugar de $L(\theta, \boldsymbol{y})$. [Para mayor detalle consultar Dobson (2002)]

Para encontrar el estimador máximo verosímil el vector score se iguala a cero, y se resuelve el sistema de ecuaciones[28]:

$$u(\hat{\theta}) = \mathbf{0}. \quad (9)$$

Siendo $\mathbf{0}$ el vector de ceros.

3.2.1.2. Matriz de información

Una propiedad estadística del vector aleatorio score es que el valor verdadero del parámetro θ tiene media cero.

$$E[u(\theta)] = \mathbf{0}, \qquad (10)$$

La matriz de covarianza del vector $u(\theta)$ nos da la matriz de información:

$$Var[u(\theta)] = E[u(\theta)u^t(\theta)] = \boldsymbol{I}(\theta). \quad (11)$$

Bajo ciertas condiciones de regularidad, la matriz de información puede ser obtenida como el valor negativo del valor esperado de la segunda derivada de la log-verosimilitud:

$$\boldsymbol{I}(\theta) = -E\left[\frac{\partial^2 logL(\theta)}{\partial\theta\partial\theta^t}\right]. \quad (12)$$

La matriz negativa de las segundas derivadas es llamada la matriz de información observada.

3.2.1.3. Newton-Raphson y Fisher Scoring

El cálculo del estimador máximo verosímil requiere de un proceso iterativo que considere expandir la función score, evaluando en el estimador máximo verosímil $\hat{\theta}$ en torno a un valor θ_0 usando una serie de Taylor de primer orden, tal que:

$$u(\hat{\theta}) \approx u(\theta_0) + \frac{\partial u(\theta)}{\partial\theta}(\hat{\theta} - \theta_0). \quad (13)$$

[28] La primera derivada de la función log-verosimilitud es necesariamente un punto crítico (máximo, mínimo o inflexión). Y si la segunda derivada es menor a cero (cóncava) o si θ es un vector del Hessiano de tal forma que éste definido no negativo, se trata de un máximo. [Para mayor detalle consultar Dobson (2002)]

Dado el Hessiano denotado por $\boldsymbol{H}$ o matriz de segundas derivadas de la función log-verosimilitud, representado por:

$$\boldsymbol{H}(\theta) = \frac{\partial^2 L}{\partial\theta\partial\theta^t} = \frac{\partial u(\theta)}{\partial\theta}. \quad (14)$$

Se considera la expresión (13) y se multiplica $\boldsymbol{H^{-1}}$ por la izquierda, obteniendo lo siguiente:

$$\mathbf{0} = \boldsymbol{H^{-1}}(\theta_0)u(\theta_0) + \left(\hat{\theta} - \theta_0\right), \quad (15)$$

Despejando se tiene:

$$\hat{\theta} = \theta_0 - \boldsymbol{H^{-1}}(\theta_0)u(\theta_0). \quad (16)$$

Este resultado proporciona la base para un enfoque iterativo para el cálculo de la estimación máxima verosimilitud conocida como la técnica de Newton-Raphson. Teniendo en cuenta un valor de prueba θ_0, usando la ecuación (16) para obtener una estimación mejorada y repetir el proceso hasta que las diferencias entre las estimaciones sucesivas son lo suficientemente próximas a cero (o hasta que los elementos del vector de primeras derivadas son lo bastante cercanos a cero).

Un procedimiento alternativo sugerido por Fisher es reemplazar $-\boldsymbol{H^{-1}}(\theta_0)$ por su valor esperado, la matriz de información $-\boldsymbol{I^{-1}}(\theta_0)$. El procedimiento resultante es una estimación mejorada, denotada por,

$$\hat{\theta} = \theta_0 + \boldsymbol{I^{-1}}(\theta_0)u(\theta_0). \quad (17)$$

Este resultado es conocido como Scoring de Fisher.

3.3. Test de Hipótesis

A continuación, se presentan algunos elementos que se necesitan para realizar pruebas de hipótesis.

3.3.1. Test de Wald

Bajo ciertas condiciones de regularidad, el estimador máximo verosimilitud $\hat{\theta}$ tiene una distribución aproximadamente p −normal con vector de media θ y matriz de covarianza dada por la matriz de información inversa $\boldsymbol{I^{-1}}(\theta)$, de modo que:

$$\hat{\theta} \sim N_p\left(\theta, \boldsymbol{I^{-1}}(\theta)\right) \quad (18)$$

Dentro de las condiciones de regularidad, se debe considerar que el parámetro a estimar pertenezca al espacio paramétrico, la función de log-verosimilitud debe ser tres veces diferenciable y delimitada.

Este resultado proporciona una base para la construcción de pruebas de hipótesis e intervalos de confianza. Por ejemplo, consideremos la siguiente hipótesis:

$$H_0: \theta = \theta_0$$

Para un vector con valor fijo θ_0, la forma cuadrática es:

$$W = \left(\hat{\theta} - \theta_0\right)^t \boldsymbol{I^{-1}}(\theta)\left(\hat{\theta} - \theta_0\right), \quad (19)$$

Bajo H_0, es aproximadamente chi-cuadrado con p grados de libertad. Por otro lado, cuando se requiera evaluar o docimar un parámetro en particular, es decir $H_0: \theta_j = 0$, el estadístico de prueba se construye entre el cociente del valor estimado $\hat{\theta}_j$ y el elemento j −ésimo de la diagonal de la matriz de información inversa en raíz cuadrada. Para este caso el estadístico de Wald es:

$$z = \frac{\hat{\theta}_j}{\sqrt{Var(\hat{\theta}_j)}} \sim N(0,1). \quad (20)$$

Denominado estadístico z.

3.3.2. AIC

Para la selección del modelo más parsimonioso existen varios métodos, destacando entre ellos los criterios de información. Para el caso de la III EME se utilizará el criterio de Akaike (AIC)[29], el cual toma un valor igual a 2 veces la función de log-verosimilitud penalizado por el número de parámetros a estimar, dado por:

$$AIC = -2\left[\log L\left(\hat{\theta}, \boldsymbol{y}\right) + p\right]. \qquad (21)$$

Luego, se elige el modelo que tenga el menor AIC.

3.4. Indicadores estadísticos para evaluar el desempeño de un procedimiento diagnóstico

3.4.1. Sensibilidad y especificidad

La **sensibilidad** y la **especificidad** son las medidas tradicionales y básicas del valor diagnóstico de un modelo. Miden la discriminación diagnóstica de un modelo en relación a un criterio de referencia, que se considera la verdad.

La **sensibilidad** (S) indica la capacidad del modelo para detectar a un sujeto que responde, es decir, expresa cuan "sensible" es la prueba a la presencia de personas que responden. Para cuantificar su expresión se utilizan términos probabilísticos: si la persona responde, ¿cuál es la probabilidad de que el resultado sea positivo?

La **especificidad** (E) indica la capacidad que tiene el modelo para identificar a las personas que no responden cuando efectivamente no responden.

[29] Los criterios de información fueron construidos como estimadores aproximadamente insesgados de la log-verosimilitud esperada $E_{G(z)}\left(\ln f\left(Y, \hat{\theta}\right)\right)$, o, equivalentemente, de la discrepancia de la Información de Kullback – Leibler entre la verdadera distribución $g(z)$ y un modelo estadístico $f\left(Y, \hat{\theta}\right)$, desde un punto de vista predictivo. En la actualidad estos criterios de información son ampliamente utilizados para la selección de modelo estadístico, en la literatura se pueden encontrar otros criterios de información como por ejemplo: el Criterio con enfoque Bayesiano de Swarchz (BIC), denotado por, $BIC = -2\log L\left(\hat{\theta}, \boldsymbol{y}\right) + \ln\ (n)p$, donde penaliza el número de parámetros p con $\ln\ (n)$. También se puede considerar el Criterio de Hannan-Quinn $HQIC = -2\log L\left(\hat{\theta}, \boldsymbol{y}\right) + 2\ln\left(\ln(n)\right)p$ como una variante del BIC con una pequeña penalización de la magnitud del tamaño muestral. La utilización del modelo AIC se utilizó para fines prácticos bajo el principio de parsimonia que establece que *todo modelo debe ser más simple que los datos en los que se basa.* [Para mayor detalle consultar Rao (2008). McCullagh(1989) y Caballero (2011) entre otros]

Considerando un espacio de unidades elegibles y las personas que responden la encuesta versus las que no, se definen los siguientes cuantificadores para la variable de respuesta:

VP: Verdaderos positivos, número de personas que respondieron la encuesta y fueron diagnosticados como positivos por el modelo.

FP: Falsos positivos, número de personas que no respondieron y fueron diagnosticados como positivos por el modelo.
FN: Falsos negativos, números de personas que respondieron y fueron diagnosticado como negativos por el modelo.
VN: Verdaderos negativos, número de personas que no respondieron y fueron diagnosticado como negativos por el modelo.

Con estos términos, la Matriz de confusión puede expresarse así:

		Criterio de Verdad		**Total**
		Responden	**No responden**	
Prueba Diagnóstica	**Positivos**	VP	FP	VP+FP
	Negativos	FN	VN	FN+VN
	Total	VP+FN	FP+VN	N=(VP+FP+FN+VN)

Fuente: Elaboración propia

$$Sensibilidad(S) = \frac{Verdaderos\ positivos}{Total\ de\ Responden} = \frac{VP}{VP + FN}$$

$$Especificidad(E) = \frac{Verdaderos\ negativos}{Total\ de\ No\ responden} = \frac{VN}{VN + FP}$$

3.4.2. Valores predictivos

A pesar de que la S y la E se consideran las características operacionales fundamentales de una prueba diagnóstica, en la práctica su capacidad de cuantificación de la incertidumbre es limitada. Se necesita más bien evaluar la medida en que sus resultados modifican realmente el grado de conocimiento que se tenía sobre el estado de la persona. Concretamente, le interesa conocer la probabilidad de que un individuo para el que se haya obtenido un resultado positivo, sea efectivamente una persona que responde; y lo contrario, conocer la probabilidad de que un individuo con un resultado negativo este efectivamente libre no

responder. Las medidas o indicadores que responden a estas interrogantes se conocen como **valores predictivos.**

El **valor predictivo de una prueba positiva** equivale a la probabilidad condicional de que los individuos con una prueba positiva realmente respondan:

VP(+) = P(Resp/T+)

El **valor predictivo de una prueba negativa** es la probabilidad condicional de que los individuos con una prueba negativa realmente no respondan:

VP(-) = P(No Resp/T-)

Mediante la tabla de 2×2 que se introdujo antes se puede ilustrar también como se estiman los valores predictivos (suponiendo que esta tabla se conforma seleccionando una muestra al azar de tamaño N de la población, y luego se clasifiquen los sujetos de la muestra en los cuatro grupos posibles según la prueba diagnóstica y el criterio de verdad) a través de:

$$Valor\ predictivo\ positivo = \frac{Verdaderos\ positivos}{Total\ de\ positivos} = \frac{VP}{VP + FP}$$

$$Valor\ predictivo\ negativo = \frac{Verdaderos\ negativos}{Total\ de\ negativos} = \frac{VN}{VN + FN}$$

3.4.3. Curva ROC

Para la elección entre dos o más modelos, se recurre a las curvas ROC, ya que es una medida global e independiente del punto de corte (o umbral).

Tradicionalmente cuando se tiene un test cuantitativo, se escoge el cut-off o punto de corte más adecuado, que combine mejor la sensibilidad y especificidad del test (es decir, mayor rendimiento). Habitualmente deberían estar con sensibilidad de 85 %, con especificidad de 74 % o cercanos a estos valores.

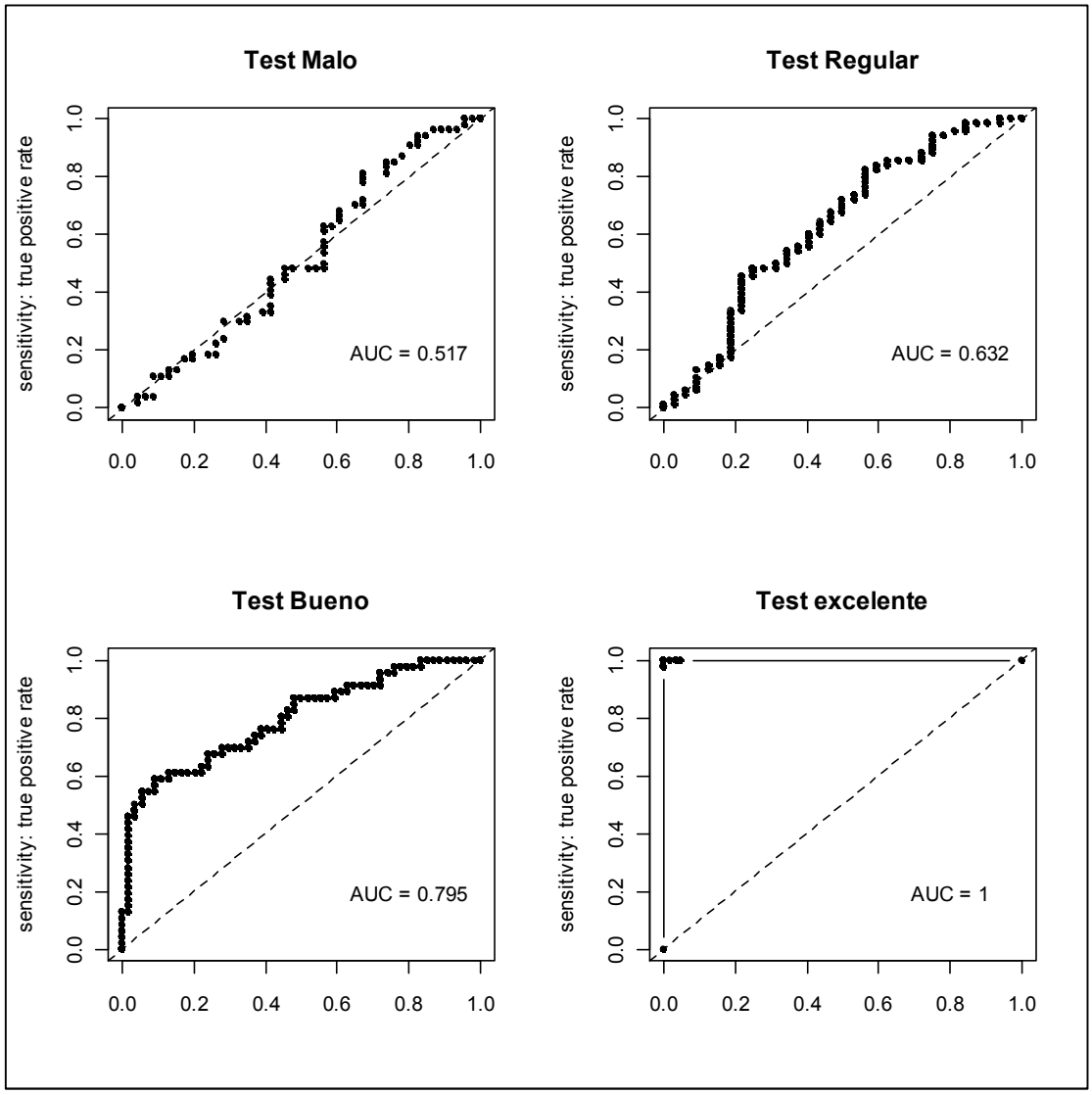
La elección se realiza mediante la comparación del área bajo la curva (AUC, de su acrónimo en inglés Area Under the Curve) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Por ejemplo, si el AUC

para una prueba diagnóstica médica es 0,8 significa que existe un 80% de probabilidad de que el diagnóstico realizado a un enfermo sea más correcto que el de una persona sana escogida al azar. Por esto, siempre se elige la prueba diagnóstica que presente una mayor área bajo la curva.

A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC:

- [0,5 - 6): Test malo.
- [0,6 - 0,75): Test regular.
- [0,75 - 0,9): Test bueno.
- [0,9 - 0,97): Test muy bueno.
- [0,97 - 1] Test excelente.

Gráfico 10. Diferentes curvas ROC



Fuente: Elaboración propia

3.5. Análisis de Elegibilidad

Para modelar la probabilidad de que una persona conteste la encuesta de la III EME dado que pertenece a una unidad elegible, se analiza primeramente la operacionalización de la variable “Código de disposición de la última visita al hogar” reportado por el encuestador en la hoja de ruta.

3.5.1. Operacionalización de variables

Cuadro 26. Distribución de personas clasificadas según el código de disposición de la última visita al hogar

Código de disposición de la última visita al hogar	Frecuencia	Porcentaje	Elegible	La Persona Responde
11. Entrevista lograda	6.758	89,50%	Sí	Sí
12. Entrevista lograda de forma parcial	7	0,090%	Sí	Si
21. Se rechazó la entrevista	118	1,54%	Sí	No
22. Vivienda ocupada sin moradores presentes	174	2,27%	Sí	No
23. Informante inubicable por cambio de domicilio o fuera del país	339	4,44%	Sí	No
24. Muerte del informante	5	0,06%	Sí	No
25. Informante con dificultad física, mental o cognitiva para contestar	19	0,24%	Sí	No
26. Problemas del Idioma	0	0,00%	Sí	No
27. Fuera de muestra	189	2,47%	No	-
31. Se impidió acceso a la vivienda(administrador, conserje o junta de vigilancia niega acceso)	0	0,00%	Sí	No
32. No fue posible localizar la dirección	2	0,02%	No	-
33. Área de difícil acceso o peligrosa	3	0,03%	No	-
41. Inmueble para uso no habitacional (empresa, oficina, vivienda colectiva, institución pública, etc.)	1	0,01%	No	-
42. Vivienda en demolición, incendiada, destruida o erradicada	0	0,00%	No	-
43. Vivienda particular desocupada (en arriendo, en venta, otro.)	16	0,20%	No	-
44. Vivienda de uso temporal (vacaciones, descanso, etc.)	1	0,01%	No	-
Total	7.632	100,00%	7420	6765

Fuente: Elaboración propia

En base a este cuadro se dividen las unidades elegibles (7.420) de las que no (212) y dentro de las unidades elegibles clasificamos las personas que responden (6.765) versus las que no (655).

3.5.2. Análisis Descriptivo

En esta sección se realiza un estudio descriptivo exploratorio para ver la relación de forma empírica entre algunas variables que pueden ingresar a nuestro modelo y ver su influencia en la variable de interés. Dada la característica de nuestra variable de interés (la persona que pertenece a una unidad elegible, responde sí o no), se

realizan principalmente cruces con variables socio-demográficas. En este sentido se inspeccionarán las distribuciones relativas y marginales (perfil fila y columna) de las siguientes variables: Nivel Educacional, Estado Conyugal y Cantidad de Visitas Colapsado.

Para simplificar el análisis de las distribuciones marginales, se divide la muestra en las personas que responde versus las que no de manera independiente. Digamos las personas que no responden pertenecen al Grupo 1 y las personas que responden al Grupo 2.

La variable "**Nivel educacional Colapsado**" corresponde a una simplificación de la variable nivel educacional, en donde la categoría Básica, incluye aquellas personas que declararon su nivel educacional con los códigos 000, 01, 02, 03 (Nunca asistió, Sala Cuna/Jardín Infantil, Kínder/Pre-Kínder, Básica o primaria) respectivamente. La categoría Media comprende los códigos 04, 05, 06 (Media común, Media Técnico Profesional, Humanidades) respectivamente. Finalmente, en la categoría Superior se encuentran los códigos 07, 08, 09, 10, 11, 12, 14 (Centro de formación técnica, Instituto Profesional, Universidad, Postítulo, Magíster, Doctorado y Normalista) respectivamente.

Cuadro 27. Distribución de personas que responden según nivel educacional colapsado y sexo.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	351	115	466
	Media	128	54	182
	Superior	3	4	7
Total No		**482**	**173**	**655**
Sí	Básica	3270	2116	5386
	Media	829	491	1320
	Superior	44	15	59
Total Sí		**4143**	**2622**	**6765**
Total general		**4625**	**2795**	**7420**

Fuente: Elaboración propia

El cuadro N°27 muestra cómo se distribuyen los casos muestrales donde por simple inspección se puede apreciar diferencias entre las personas que responden o no, respecto al nivel educacional y sexo.

Con esto se construye la distribución porcentual relativa según nivel educacional y sexo.

Cuadro 28. Distribución porcentual relativa de personas que responden según nivel educacional colapsado y sexo.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	4,7%	1,5%	6,3%
	Media	1,7%	0,7%	2,5%
	Superior	0,0%	0,1%	0,1%
Total No		**6,5%**	**2,3%**	**8,8%**
Sí	Básica	44,1%	28,5%	72,6%
	Media	11,2%	6,6%	17,8%
	Superior	0,6%	0,2%	0,8%
Total Sí		**55,8%**	**35,3%**	**91,2%**
Total general		**62,3%**	**37,7%**	**100,0%**

Fuente: Elaboración propia

En este caso se analiza el aporte de casos, distribuidos en las personas que **Responde**, **Nivel Educacional** y **Sexo**, respecto al total de casos. Donde por ejemplo el 4,7% de los casos que no respondieron pertenecen al nivel educacional básica y son hombres versus 44,1% de las personas que responden en la misma categoría.

Cuadro 29. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Nivel Educacional con respecto al sexo.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	75,3%	24,7%	100,0%
	Media	70,3%	29,7%	100,0%
	Superior	42,9%	57,1%	100,0%
Total No		**73,6%**	**26,4%**	**100,0%**
Sí	Básica	60,7%	39,3%	100,0%
	Media	62,8%	37,2%	100,0%
	Superior	74,6%	25,4%	100,0%
Total Sí		**61,2%**	**38,8%**	**100,0%**

Fuente: Elaboración propia

En este caso para la distribución marginal Sexo respecto al nivel educacional, se puede decir que dentro de todas las personas que no responden, dado que poseen un nivel educacional básico, el 75,3% de los casos pertenecen al sexo Hombre y el 24,7% Mujer. Para los que pertenecen al nivel educacional medio 70,3% son hombres y 29,7% son mujeres. Finalmente del nivel educacional superior el 42,9%

son hombres y el 57,1% son mujeres. De igual forma, en el caso de todas las personas que responden, se tiene que; los que pertenecen al nivel educacional básico el 60,7% son hombres y el 39,3% son mujeres. En Media el 62,8% son hombres y el 37,2% son mujeres. En el caso del nivel educacional superior el 74,6% son hombres y el 25,4% mujeres.

Cuadro 30. Análisis de perfil columna separando la distribución porcentual de personas que responden (si o no). Fijando Sexo con respecto al Nivel Educacional.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	72,8%	66,5%	71,1%
	Media	26,6%	31,2%	27,8%
	Superior	0,6%	2,3%	1,1%
Total No		**100,0%**	**100,0%**	**100%**
Sí	Básica	70,7%	80,7%	79,6%
	Media	17,9%	18,7%	19,5%
	Superior	1,0%	0,6%	0,9%
Total Sí		**100,0%**	**100,0%**	**100,0%**

Fuente: Elaboración propia

Dentro de todas las personas que no responden, las personas que pertenecen al sexo Hombre, el porcentaje que pertenece a educación Básica es de 72,8%, el que pertenece a la educación Media es 26,6% y a la educación superior es 0,6%. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 66.5% pertenece a la educación Básica, el 31,2% a la Media y el 2,3% al nivel Superior. Dentro de todas las personas que responden, se puede observar que dado que son hombres; el 70,7% de los casos pertenece al nivel educacional básico, el 17,9% Media y el 1% a nivel Superior. En este mismo sentido, en el caso de las mujeres 80,7% pertenece al nivel educacional Básica, 18,7% Media y 0,6% al nivel Superior.

En este contexto, se puede apreciar que el mayor aporte en contestar la encuesta son mujeres que pertenecen al nivel educacional básico. (70,7% versus 80,7% hombres y mujeres respectivamente).

Para la variable "**Estado Conyugal Colapsado**" se realizó una simplificación de la variable Estado Conyugal, en donde la categoría Casado(a) – Conviviente se encuentran los códigos 1 y 2 (Casado y Conviviente) respectivamente. En la categoría otros se encuentran los códigos 3, 4, 5 y 6 (Soltero(a), Viudo(a), separado(a) de hecho anulado(a) y Divorciado(a)) respectivamente. Se constata que

las personas que responden la III EME tienen una relación directa con el estado "Casado(a)-conviviente".

Cuadro 31. Distribución de personas que responden según estado conyugal colapsado y sexo.

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	328	90	418
	Otros	154	83	237
Total No		**482**	**173**	**655**
Sí	Casado(a) - Conviviente	3016	1470	4486
	Otros	1127	1152	2279
Total Sí		**4143**	**2622**	**6765**
Total general		**4625**	**2795**	**7420**

Fuente: Elaboración propia

En base al cuadro anterior se puede construir la frecuencia porcentual relativa de personas que responden según nivel educacional y sexo.

Cuadro 32. Distribución porcentual relativa de personas que responden según nivel educacional colapsado y sexo.

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	4,4%	1,2%	5,6%
	Otros	2,1%	1,1%	3,2%
Total No		**6,5%**	**2,3%**	**8,8%**
Sí	Casado(a) - Conviviente	40,6%	19,8%	60,5%
	Otros	15,2%	15,5%	30,7%
Total Sí		**55,8%**	**35,3%**	**91,2%**
Total general		**62,3%**	**37,7%**	**100,0%**

Fuente: Elaboración propia

Dentro de las personas que no responden el 4,4% de los hombres y el 1,2% de mujeres, pertenecen a estado conyugal Casado-conviviente representado el 5,6% de los casos. En cambio dentro de las personas que responden, el 40,6% hombres y 19,8% son mujeres, respecto a la misma categoría.

Por otro lado si analizamos las distribuciones marginales del estado conyugal con respecto al sexo, se pueden observar pequeñas diferencias porcentuales.

Cuadro 33. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Estado Conyugal con respecto al sexo.

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	78,5%	21,5%	100,0%
	Otros	65,0%	35,0%	100,0%
Total No		**73,6%**	**26,4%**	**100,0%**
Sí	Casado(a) - Conviviente	67,2%	32,8%	100,0%
	Otros	49,5%	50,5%	100,0%
Total Sí		**61,2%**	**38,8%**	**100,0%**

Fuente: Elaboración propia

Dentro de todas las personas que no responden, se puede decir que dado que pertenecen a la categoría Casado - Conviviente, el 78,5% de los casos pertenecen al sexo Hombre y el 21,5% Mujer. Para los que pertenecen a la categoría Otros, el 65% son hombres y el 35% son mujeres.

Por otro lado, en el caso de todas las personas que responden, se puede decir que dado que pertenecen a la categoría Casado - Conviviente, el 67,2% de los casos pertenecen al sexo Hombre y el 32,8% Mujer. Para los que pertenecen a la categoría Otros, el 49,5% son hombres y el 50,5% son mujeres.

De la misma forma, si analizamos la distribución marginal del estado conyugal fijando el sexo se tiene que:

Cuadro 34. Análisis de perfil columna separando la distribución porcentual de personas que responden (si o no). Fijando Sexo con respecto al Estado conyugal

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	68,0%	52,0%	63,8%
	Otros	32,0%	48,0%	36,2%
Total No		**100,0%**	**100,0%**	**100,0%**
Sí	Casado(a) - Conviviente	72,8%	56,1%	66,3%
	Otros	27,2%	43,9%	33,7%
Total Sí		**100,0%**	**100,0%**	**100,0%**

Fuente: Elaboración propia

Dentro de todas las personas que no responden, las personas que pertenecen al sexo Hombre, el porcentaje de estos que pertenecen al estado conyugal Casado – Conviviente es de 68%, el 32% pertenece a Otros. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 52% pertenece a Casado - Conviviente, y 48% a Otros. Para las personas que responden, las personas que pertenecen al sexo Hombre, el porcentaje de estos que pertenecen al estado conyugal Casado – Conviviente es de 72,8% y el 27,2% pertenece a Otros. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 56,1% pertenece a Casado - Conviviente, y 43,9% a Otros.

Finalmente, se analiza la variable **cantidad de visitas** que tiene un recorrido de 1 a 12 visitas, para esto se simplificó en tres categorías “1-3”, “4-6” y”7 y más”. Se observa que gran parte de las personas que respondieron la encuesta se encuentra dentro del tramo 1 a 3 visitas al hogar.

Cuadro 35. Distribución de personas que responden según cantidad de visitas Colapsado y sexo.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	240	97	337
	4-6	197	58	255
	7 y más	45	18	63
Total No		**482**	**173**	**655**
Sí	1-3	3725	2412	6137
	4-6	362	186	548
	7 y más	56	24	80
Total Sí		**4143**	**2622**	**6765**
Total general		**4625**	**2795**	**7420**

Fuente: Elaboración propia

En base a este cuadro se pueden obtener las siguientes frecuencias relativas:

Cuadro 36. Distribución porcentual relativa de personas que responden según Cantidad de Visitas colapsado y sexo.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	3,2%	1,3%	4,5%
	4-6	2,7%	0,8%	3,4%
	7 y más	0,6%	0,2%	0,8%
Total No		**6,5%**	**2,3%**	**8,8%**
Sí	1-3	50,2%	32,5%	82,7%
	4-6	4,9%	2,5%	7,4%
	7 y más	0,8%	0,3%	1,1%
Total Sí		**55,8%**	**35,3%**	**91,2%**
Total general		**62,3%**	**37,7%**	**100,0%**

Fuente: Elaboración propia

Se puede ver que el 50,2% de los casos se concentra en las personas que responden entre “1-3” visitas y son hombres.

Al analizar la distribución marginal de la cantidad de visitas se tiene que:

Cuadro 37. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Sexo con respecto a la cantidad de visitas.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	71,2%	28,8%	100,0%
	4-6	77,3%	22,7%	100,0%
	7 y más	71,4%	28,6%	100,0%
Total No		**73,6%**	**26,4%**	**100,0%**
Sí	1-3	60,7%	39,3%	100,0%
	4-6	66,1%	33,9%	100,0%
	7 y más	70,0%	30,0%	100,0%
Total Sí		**62,3%**	**37,7%**	**100,0%**

Fuente: Elaboración propia

De igual forma se puede obtener la distribución marginal del sexo

Cuadro 38. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Cantidad de visitas con respecto al sexo.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	49,8%	56,1%	51,5%
	4-6	40,9%	33,5%	38,9%
	7 y más	9,3%	10,4%	9,6%
Total No		**100,0%**	**100,0%**	**100,0%**
Sí	1-3	89,9%	92,0%	90,7%
	4-6	8,7%	7,1%	8,1%
	7 y más	1,4%	0,9%	1,2%
Total Sí		**100,0%**	**100,0%**	**100,0%**

Fuente: Elaboración propia

Del total de personas que responde, el 90,7% fue visitada entre 1-3 veces, mientras que sólo el 1,2% fue visitado en 7 o más oportunidades para lograr concretar la entrevista.
Sin embargo, existe una mayor probabilidad de entrevistar a las mujeres en al menos tres visitas, 92%.

Para la variable cantidad de visitas, se puede apreciar que el mayor aporte en contestar la encuesta son mujeres que pertenecen a la categoría entre “1-3”. (89,9% hombres versus 92% mujeres).

En resumen, se puede observar que existe una relación entre las personas que responden versus nivel educacional, siendo los niveles básico y medio los con mayor participación de personas, como igual a la cantidad de visitas.

3.6. Aplicación

El principio básico en la inclusión de variables está basado en un modelo simple con un número de variables restringido sobre el total de variables existentes. Se probaron varios modelos, sin embargo el que mejor cumple las condiciones, es el que contiene las siguientes variables explicativas; edad de la persona, macrozona de pertenencia del hogar, grupo ocupacional, área geográfica, número de visitas, proveedor principal y sexo de la persona. El cuadro 39, muestra los parámetros estimados para este modelo, de acuerdo a las categorías que son estadísticamente significativas (p-value) de cada variable explicativa.

Los **Odd Ratios** $e^{\widehat{\beta}_1}$ se pueden interpretar como el aumento estimado en la probabilidad de éxito asociado con un cambio unitario en el valor de la variable predictora. En general, el aumento estimado está asociado con un cambio de d unidades en la variable predictora, es decir, $e^{d\cdot\widehat{\beta}_1}$.

La interpretación de los coeficientes de regresión en el modelo de regresión logístico múltiple se parece al caso en el que el predictor lineal sólo contiene un regresor, que nos indica que la cantidad $e^{\widehat{\beta}_1}$ es el cociente de ventaja para la covariable x_j, suponiendo que las demás variables predictoras son constantes.

Cuadro 39. Parámetros estimados del modelo de regresión logística seleccionado para modelar la respuesta o no de una persona que pertenece a una unidad elegible.

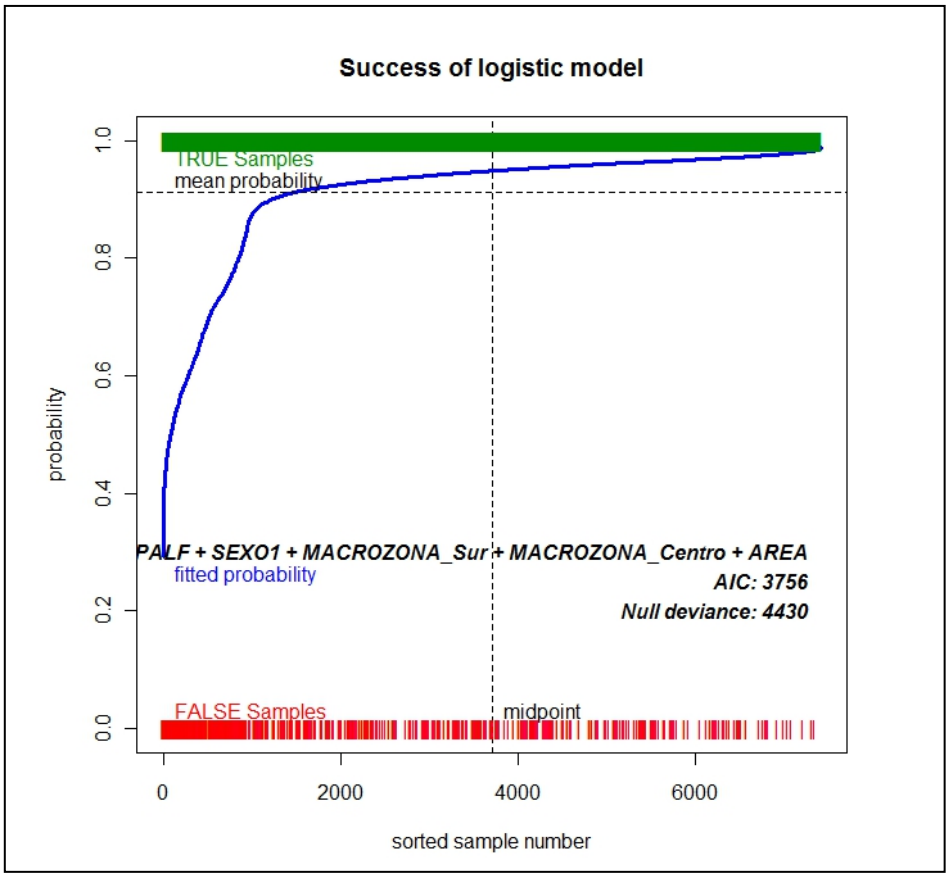
Variables	Estimación	Error Estándar	Valor z	Pr(>|z|)	Odd Ratios	Intervalo de Confianza 95%	
						Lim Inf	Lim Sup
Intercepto	2,579	0,215	12,004	0,000	13,190	8,682	20,161
Edad	0,013	0,004	3,757	0,000	1,013	1,006	1,020
Macrozona Sur	-0,268	0,125	-2,147	0,032	0,765	0,600	0,978
Macrozona Centro	-0,763	0,105	-7,274	0,000	0,466	0,379	0,572
CIUO_88_1_Digito7 (servicios Financieros)	0,189	0,109	1,737	0,082	1,208	0,979	1,498
Variables categóricas en el modelo							
Área							
Urbano	-	-	-	-	-	-	-
Rural	-0,293	0,114	-2,563	0,010	0,746	0,597	0,935
Cantidad de Visitas							
1-3	-	-	-	-	-	-	-
4-6	-2,224	0,102	-21,726	0,000	0,108	0,088	0,132
7 y más	-2,951	0,190	-15,530	0,000	0,052	0,036	0,076
Nivel Educacional							
Básica	-	-	-	-	-	-	-
Media	-0,007	0,109	-0,065	0,948	0,993	0,804	1,231
Superior	0,056	0,435	0,128	0,898	1,057	0,479	2,693
Proveedor Principal							
Sí	-	-	-	-	-	-	-
No	-0,211	0,102	-2,072	0,038	0,810	0,663	0,989
Sexo							
Hombre	-	-	-	-	-	-	-
Mujer	0,580	0,105	5,539	0,000	1,785	1,457	2,196

Fuente: Elaboración propia

3.6.1. Análisis de Resultados

El gráfico a continuación, presenta las probabilidades estimadas para cada persona que pertenece a una unidad elegible.

Gráfico 11 Probabilidad estimada de responder para cada una de las personas que pertenecen a la unidad elegible



Fuente: Elaboración propia

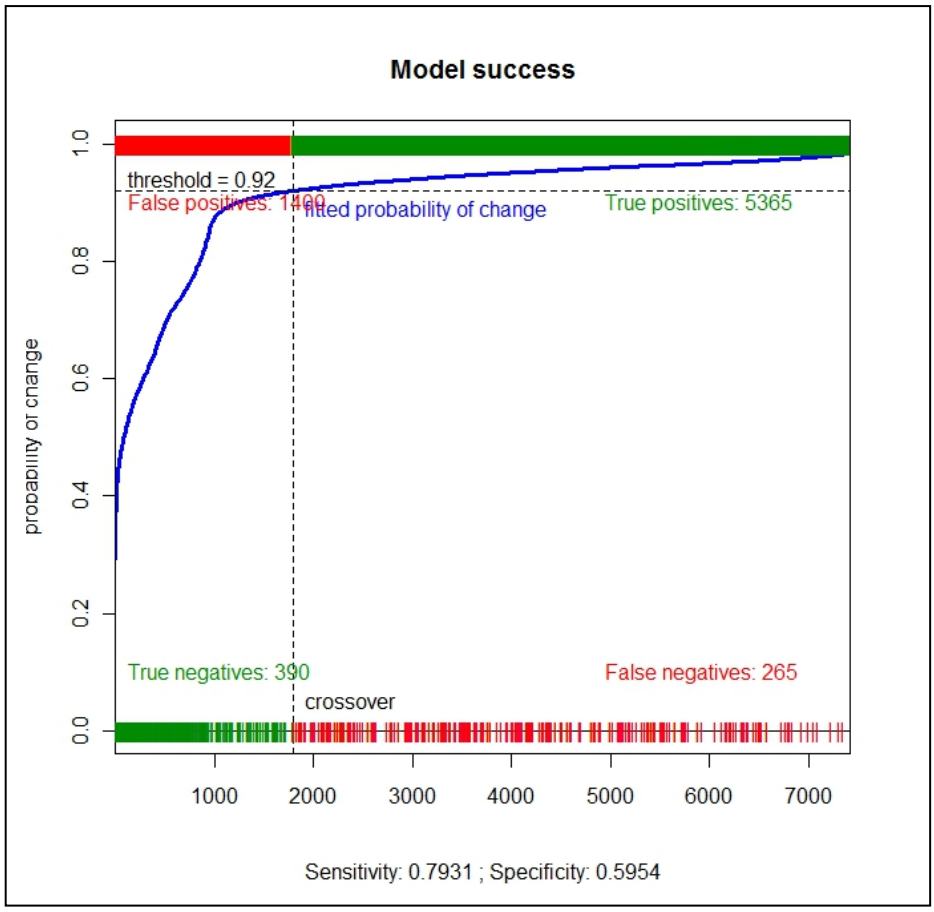
Cuadro 40. Estadísticas Descriptivas para la probabilidad estimada del modelo propuesto

Variable	n	Media	Desv. Estándar	Mediana	Mínimo	Máximo	Rango	Error Muestral
Prob. Estimada	7420	0,912	0,105	0,948	0,293	0,988	0,696	0,001

Fuente: Elaboración propia

Se puede observar que el modelo seleccionado ajusta una probabilidad de responder de una persona entre 30% y 99%, con una probabilidad media de 91%.

Gráfico 12. Clasificación de las personas que responden versus las que no responden con un umbral de 0,92



Fuente: Elaboración propia

Cuadro 41. Matriz de Confusión con un umbral = 0,92

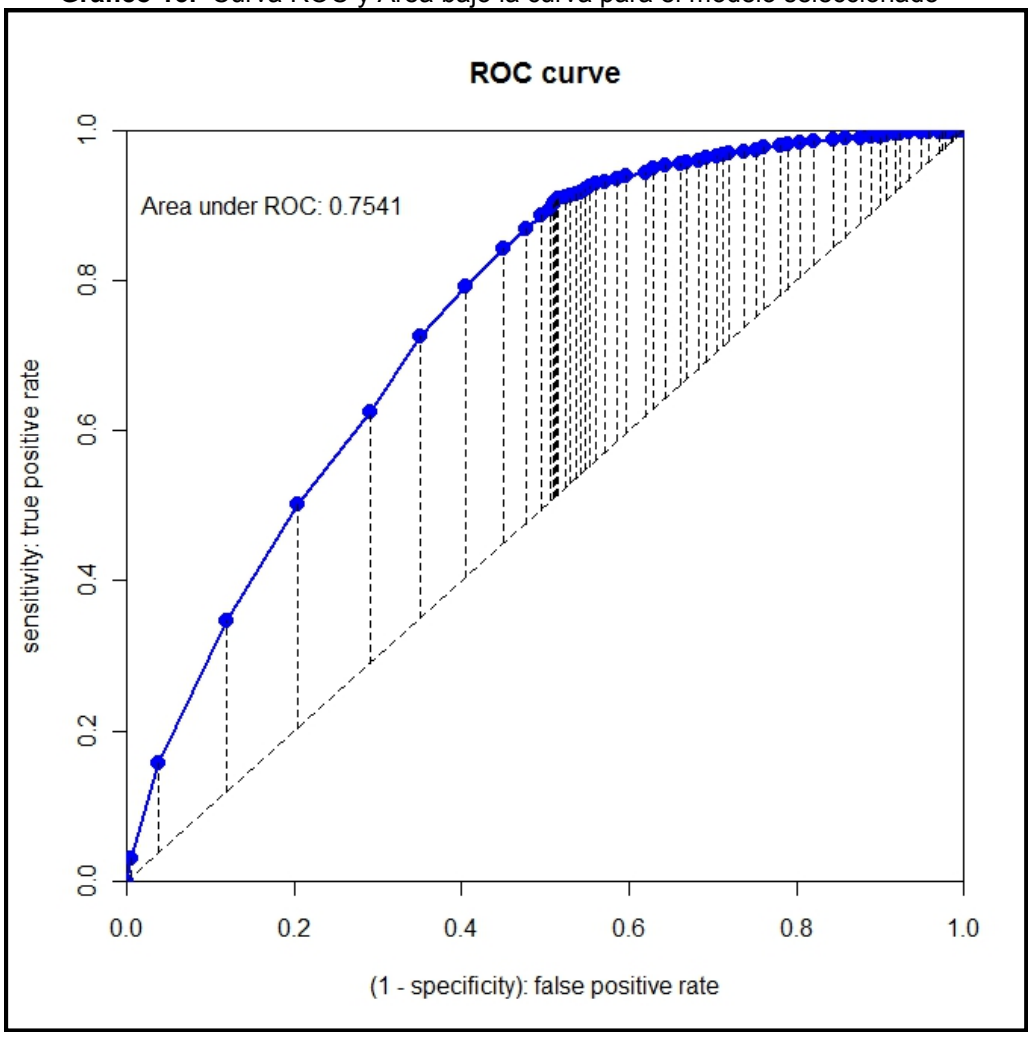
		Criterio de Verdad		
		Responden	**No responden**	**Total**
Prueba Diagnóstica	**Positivos**	5365	1400	6765
	Negativos	265	390	655
	Total	5630	1790	7420

Fuente: Elaboración propia

Finalmente, la sensibilidad y especificidad calculada es:

$$Sensibilidad = \frac{5365}{6765} = 0{,}793 \qquad Especificidad = \frac{390}{655} = 0{,}595$$

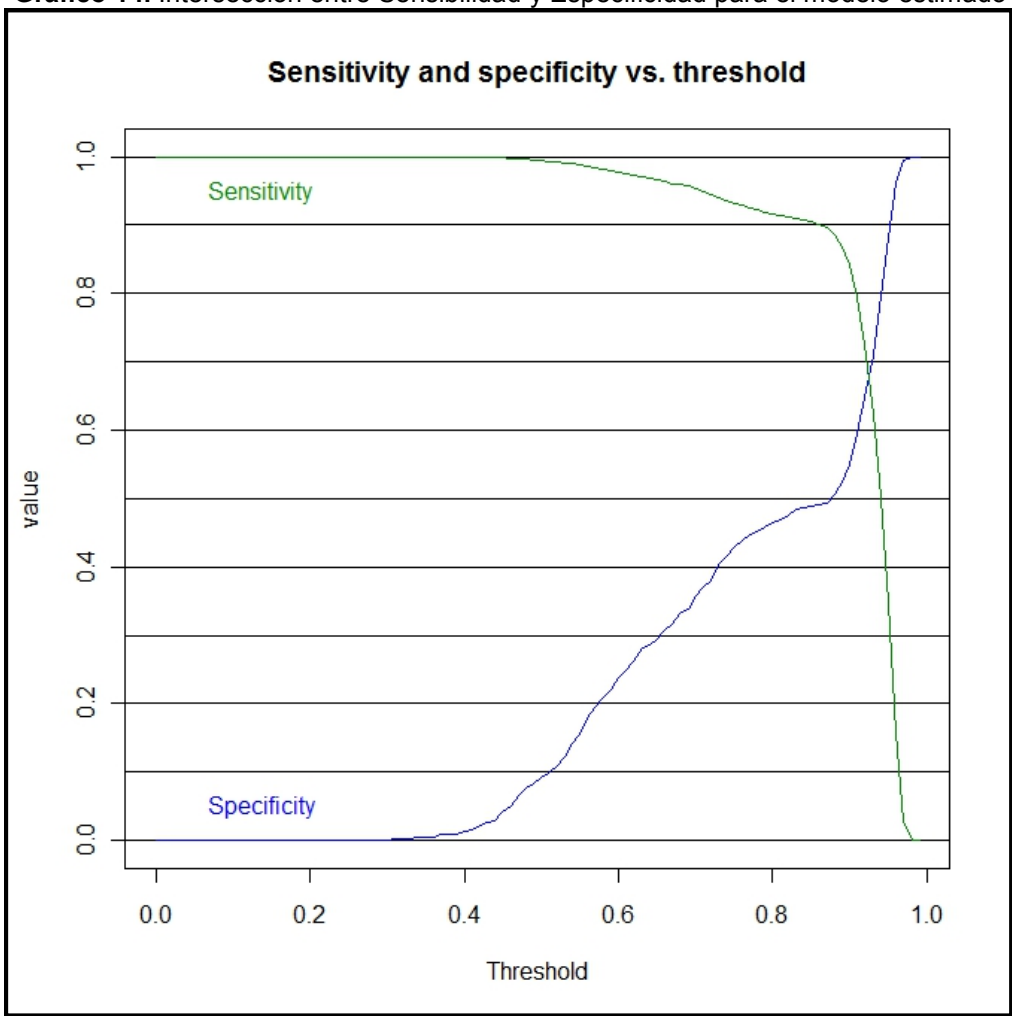
Gráfico 13. Curva ROC y Área bajo la curva para el modelo seleccionado



Fuente: Elaboración propia

En base al modelo estimado se puede observar que el área bajo la curva (AUC) es de 0,75, lo cual está dentro **de una categoría de "Test bueno".** O bien, se puede decir que el modelo tiene una capacidad de predicción del 75% de los casos.

Gráfico 14. Intersección entre Sensibilidad y Especificidad para el modelo estimado



Fuente: Elaboración propia

La mejor relación entre especificidad y sensibilidad que puede tener este modelo propuesto es cuando se utiliza un umbral de 0,92 para clasificar a las personas.

4. Anexo N°4. Estimación de varianzas

4.1 Creación de variables y determinación del diseño muestral en Spss

***Plan de muestreo**

```
CSPLAN ANALYSIS
  /PLAN FILE='planEME.csaplan'
  /PLANVARS ANALYSISWEIGHT=FACT_EME
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=pseudo_estrato CLUSTER=pseudo_conglomerado
  /ESTIMATOR TYPE=WR.
```

***Creación de variables**

```
* Rama reducida.
COMPUTE Rama_reducida =d15_rev3_1.
IF(d15_rev3_1=2 | d15_rev3_1=3 | d15_rev3_1=5) Rama_reducida=20.
IF (d15_rev3_1=8 | d15_rev3_1=10 | d15_rev3_1=12 | d15_rev3_1=13 | d15_rev3_1=14 |
d15_rev3_1=16) Rama_reducida=21.
EXECUTE.

VALUE LABELS Rama_reducida
20 'Sector Primario'
21 'Servicios'.
EXECUTE.
*Pegar etiqueta desde rama
APPLY DICTIONARY
  /FROM *
  /SOURCE VARIABLES=d15_Rev3_1
  /TARGET VARIABLES=Rama_reducida
  /FILEINFO
  /VARINFO ALIGNMENT FORMATS LEVEL ROLE MISSING VALLABELS=MERGE
ATTRIBUTES=MERGE VARLABEL WIDTH.
```

***Estimación de frecuencias en Spss**

```
CSTABULATE
  /PLAN FILE='J:\EME\2013\Varianza\Cálculo de varianzas\Final\planEME.csaplan'
  /TABLES VARIABLES=Rama_reducida CISE_EME
  /CELLS TABLEPCT
  /STATISTICS SE CV CIN(95) DEFF
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

4.2 Creación de variables y determinación del diseño muestral en Stata

***Plan de muestreo en Stata**

```
. svyset pseudo_conglomerado [pw=FACT_EME], strata(pseudo_estrato)
pweight: FACT_EME
        VCE: linearized
  Single unit: missing
     Strata 1: pseudo_estrato
         SU 1: pseudo_conglomerado
        FPC 1: <zero>
```

***Creación de variables**

```
. gen Rama_reducida:d15_Rev3_1=d15_Rev3_1
(388 missing values generated)
. replace Rama_reducida=20 if(d15_Rev3_1==2 | d15_Rev3_1==3 | d15_Rev3_1==5)
(159 real changes made)
. replace Rama_reducida=21 if (d15_Rev3_1==8 | d15_Rev3_1==10 | d15_Rev3_1==12 | d15_Rev3_1==13 | d15_Rev3_1==14 | d15_Rev3_1==16)
(463 real changes made)
```

***Estimación de frecuencias en Stata**

```
. svy: prop Rama_reducida
```