Encuesta de Microemprendimiento 2015

DISEÑO MUESTRAL

INSTITUTO NACIONAL DE ESTADÍSTICAS

Diciembre/ 2015

DEPARTAMENTO DE INVESTIGACIÓN Y DESARROLLO
DEPARTAMENTO DE ESTUDIOS LABORALES

Encuesta de Microemprendimiento 2015 - Diseño Muestral

Instituto Nacional de Estadísticas.

Diciembre / 2015.

Jefe Departamento de Investigación y Desarrollo: Charles Durán A.

Jefe Departamento de Estudios Laborales: David Niculcar.

Jefe de proyecto IV EME: David Niculcar

Coordinador Sección Estadísticas Sociales: Miguel Guerrero H.

Analista(s) Investigador(es): Bárbara Basáez Ch.

Marta Cisternas A.

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 Cuarta Encuesta de Microemprendimiento (IV 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 entrevistado a través de 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 IV EME fue seleccionada aleatoriamente a fin de representar tanto 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 cuarta 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] 2015 fueron utilizadas como marco de muestreo para la IV 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 con un listado de trabajadores independientes pertenecientes a las viviendas que no fueron descartadas. Posteriormente, la selección (pivote) de los trabajadores independientes se realizó de forma sistemática e independiente al interior de cada rama de actividad económica para cada región y bajo una estratificación implícita geográfica. Luego, se listaron todos los trabajadores independientes al interior del hogar, y si 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 2015.

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 IV 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 IV 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 2015); y segundo el marco utilizado para la selección de los "independientes" para la IV 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 IV 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 IV 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 2015

[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
	Aysén 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 IV 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 2015, 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 IV 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 vendedoras por catálogo 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 2015; 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 IV EME. En total, la depuración del marco corresponde a 10,9%[6] de casos descartados por ser potenciales unidades no elegibles[7], observándose los mayores cambios en la región de Aysén (15,0%) y los menores en la región de Magallanes y La Antártica Chilena con un 4,8%.

Cuadro 2. Distribución del total de independientes muestrales según ENE (MAM 2015) y según Marco EME

Macrozona	Región	Total Independientes ENE	Total Independientes EME
Total		**11.376**	**10.130**
Norte	Arica y Parinacota	470	422
	Tarapacá	422	389
	Antofagasta	263	229
	Atacama	267	237
	Coquimbo	794	725
Total Norte		**2.216**	**2.002**
Centro	Valparaíso	1.465	1.263
	Libertador General Bernardo O'Higgins	571	507
	Maule	744	681
	Biobío	1.264	1.127
Total Centro		**4.044**	**3.578**
Sur	La Araucanía	824	753
	Los Ríos	348	325
	Los Lagos	962	876
	Aysén del General Carlos Ibáñez del Campo	317	269
	Magallanes y La Antártica Chilena	124	118
Total Sur		**2.575**	**2.341**
Metropolitana	Metropolitana de Santiago	2.541	2.209
Total Metropolitana		**2.541**	**2.209**

[6] $\frac{11.376-10.130}{11.376} = 0{,}109$

[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.

Fuente: Elaboración propia

2.2. Estimación y Distribución del tamaño muestral

La IV 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:

$$R_{Y/X} = \frac{N° \; Trabajadores \; Cuenta \; Propia}{N° \; Trabajadores \; Independientes} = \frac{Y}{X} = \frac{\sum_{k \in U} y_k}{\sum_{k \in U} x_k} \qquad (1)$$

La variable pivote considerada fue la proporción entre trabajadores por cuenta propia y el total de trabajadores independientes.

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 2,52%.

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	83,5%	2,772	5.897	2,52%	3,02%

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 IV 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	83,5%	2,772	6.880

Fuente: Elaboración propia

[10] P corresponde a la razón entre el total de trabajadores por cuenta propia y el total de trabajadores independientes en el período de referencia. Este indicador se utiliza principalmente para poder comparar los errores teóricos provenientes de la primera fase (Encuesta Nacional de Empleo) versus los errores efectivos calculados con los resultados finales de la propia encuesta de Microemprendimiento, ya que para el cálculo del tamaño muestral de la encuesta se utilizó la variable "total de trabajadores independientes sobre el total de ocupados", proporción que no es posible de replicar en la segunda fase.

[11] P corresponde a la razón entre el total de trabajadores por cuenta propia y el total de trabajadores independientes en el período de referencia.

El tamaño de la muestra teórica12 es de 6.880 viviendas aproximadamente, sujeto a un nivel de estimación nacional y error absoluto fijo de 2,52% para la razón entre trabajadores por cuenta propia y trabajadores independientes. 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 y macrozona para el indicador principal. Las desagregaciones de la información que se hagan dentro de estas áreas de estimación pueden llevar a alta variabilidad de las estimaciones y a un menor grado de precisión estadística, por tanto siempre se deben analizar los errores de estimación antes de sacar una inferencia estadística sobre la población (ver anexo metodológico).

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 2015, se decidió utilizar como información auxiliar el total de independientes reportados en el trimestre móvil MAM 2014, 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 IV 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 2014. 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

[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.

iguales, cuando ello fuera posible, según el mes de levantamiento. Así, las viviendas a encuestar en el mes de mayo en la IV EME deberán ser aquellas viviendas que fueron entrevistadas en Marzo 2015 en la ENE.

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

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

Macrozona	Región	Mes Levantamiento IV EME			Total
		Mayo	Junio	Julio	
Total EME		**2.293**	**2.293**	**2.294**	**6.880**
Norte	Arica y Parinacota	97	98	97	292
	Tarapacá	87	87	86	260
	Antofagasta	57	59	58	174
	Atacama	60	56	60	176
	Coquimbo	154	155	154	463
Total Norte		**455**	**455**	**455**	**1.365**
Centro	Valparaíso	291	291	292	874
	Libertador General Bernardo O'Higgins	114	114	114	342
	Maule	153	153	152	458
	Biobío	265	266	266	797
Total Centro		**823**	**824**	**824**	**2.471**
Sur	La Araucanía	161	161	165	487
	Los Ríos	74	74	76	224
	Los Lagos	195	195	198	588
	Aysén del General Carlos Ibáñez del Campo	66	65	55	186
	Magallanes y La Antártica Chilena	26	26	28	80
Total Sur		**522**	**521**	**522**	**1.565**
Metropolitana	Metropolitana de Santiago	493	493	493	1.479
Total Metropolitana		**493**	**493**	**493**	**1.479**

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 2014. Sin embargo, el marco de muestreo desde el cual se seleccionó la IV 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 realizó la selección (pivote) de las personas. En el cuadro 6 se presenta la distribución del total de personas según región.

Cuadro 6. Total de personas Marco EME

Macrozona	Región	EME MAM 2015	Selección EME
		Total de Independientes	Número Independientes
Total EME		**10.130**	**7.543**
Norte	Región de Arica y Parinacota	422	316
	Región de Tarapacá	389	296
	Región de Antofagasta	229	189
	Región de Atacama	237	199
	Región de Coquimbo	725	512
Total Norte		**2.002**	**1.512**
Centro	Región de Valparaíso	1.263	952
	Región del Libertador Gral. Bernardo O'Higgins	507	368
	Región del Maule	681	502
	Región del Bío Bío	1.127	861
Total Centro		**3.578**	**2.683**
Sur	Región de La Araucanía	753	537
	Región de los Ríos	325	264
	Región de los Lagos	876	652
	Región De Aysén del Gral. Carlos Ibáñez del Campo	269	204
	Región de Magallanes y Antártica Chilena	118	85
Total Sur		**2.341**	**1.742**
Metropolitana	Región Metropolitana	2.209	1.606
Total Metropolitana		**2.209**	**1.606**

Fuente: Elaboración propia

En la segunda fase, la IV EME tiene cobertura del área urbana y rural del país, estratificada según rama de actividad económica en cada región para fines de distribución muestral de las unidades a levantar.

La selección (pivote) de los trabajadores independientes se realizó de forma sistemática e independiente al interior de cada rama de actividad económica para cada región. En aquellos hogares dentro de las viviendas con 2 o más trabajadores independientes se seleccionaron 2 o más informantes, según el número de actividades distintas que se identificaran. No obstante, si en un hogar se encontraban 2 o más trabajadores independientes con la misma rama, sólo se seleccionaba a uno de ellos. En el cuadro 7, se detalla la distribución según rama para la región de Valparaíso a modo de ejemplo.

Cuadro 7. Distribución de trabajadores independientes por rama de Actividad económica en la Región de Valparaíso

Rama Actividad	Número de Independientes	% de N Región	Tamaño	Disminución	Aumento	Personas a seleccionar INICIAL	Personas seleccionadas FINAL
			Teórico	15% + grandes	en menores		
1 Agricultura, ganadería, caza y silvicultura	31	6,90%	22		7	29	28
2 Pesca	5	1,10%	4		1	5	4
3 Explotación de minas y canteras			0			0	
4 Industrias manufactureras	62	**13,80%**	45	**7**		38	40
6 Construcción	58	**12,90%**	42	**6**		36	40
7 Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	145	**32,20%**	105	**16**		89	83
8 Hoteles y restaurantes	14	3,10%	10		4	14	14
9 Transporte, almacenamiento y comunicaciones	47	10,40%	34		4	38	37
10 Intermediación financiera	1	0,20%	1			1	1
11 Actividades inmobiliarias, empresariales y de alquiler	46	10,20%	33		4	37	37
12 Administración pública y defensa; planes de seguridad social de afiliación obligatoria						0	
13 Enseñanza	3	0,70%	2		1	3	3
14 Servicios sociales y de salud	12	2,70%	9		3	12	11
15 Otras actividades de servicios comunitarios, sociales y personales	26	5,80%	19		5	24	23
Total	**450**	**100%**	**326**	**29**	**29**	**326**	**321**

Fuente: Elaboración propia

Para disminuir la sobre representación, en términos del número de observaciones presentes en la muestra, que puedan tener algunas actividades económicas que a

priori se sabe son más recurrentes dentro de los trabajadores independientes como son los sectores de Comercio, Industria Manufacturera y Construcción, se disminuyó en un 15% la muestra de éstas tres ramas y se realizó un aumento en aquellas con menor presencia, para así tratar de disminuir los errores de estimación futuros. Cabe señalar, que esto sólo es una medida de mitigación, ya que la precisión estadística final de estimaciones está sujeta a los resultados encontrados en el trabajo de campo y los usuarios de la encuesta deben ser responsables del cálculo de los respectivos coeficientes de variación antes de sacar cualquier inferencia estadística sobre la población (ver anexo metodológico). En el diseño muestral de la encuesta se sugiere agrupar ciertas ramas de actividad económica para mejorar la precisión estadística.

.

3. FACTORES DE EXPANSIÓN

La muestra de la IV 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 periodo MAM 2015.
    - 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 IV).
2. Ajuste a un stock poblacional dado un periodo 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.
2. Tanto las probabilidades de selección de las viviendas y de las personas se extraerán directamente desde el Marco de la IV 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 IV EME.
4. Calibración al stock poblacional, el cual fue creado a partir de los datos recogidos en la ENE en el periodo de referencia donde se seleccionó la muestra (MAM 2015), pero ajustados al crecimiento poblacional estimado a partir de las proyecciones poblacionales de junio del 2015, 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 IV EME, las personas seleccionadas, corresponden a un subconjunto de personas que participaron durante el proceso de encuestaje del trimestre MAM 2015 de la ENE. Por lo tanto, uno de los insumos fundamentales del ponderador base, son los factores de expansión por vivienda de la ENE, que dan cuenta de la probabilidad de que una vivienda haya sido seleccionada 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 2015.

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 los trabajadores independientes. 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 periodo MAM 2015.

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

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 2015 en la ENE, está dado por:

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

Probabilidad de selección de los independientes

La selección de los independientes se realizó en dos etapas. Primero, observando la distribución para cada rama de actividad económica por región, y luego distribuyendo el total de independientes a seleccionar por esta estratificación (rama de actividad económica por región). Luego de esto se asociaron las viviendas a estos independientes seleccionados.

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. $R = 1, \dots, 15$.

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.

La probabilidad de seleccionar al independiente k al interior de la vivienda j, perteneciente a la rama de actividad económica R de la región R, dado que la vivienda fue seleccionada, puede ser aproximada por:

$$p_{RRjk}^{indep|v} = \frac{S_{RRj}^{indep}}{T_{RRj}^{indep}}$$

Donde:

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

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

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

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

En el cuadro 8 se observa que, en general, para el total de independientes seleccionados en la IV EME, la probabilidad de seleccionarlos oscila entre 33% y 100%.

Cuadro 8. Estadísticas descriptivas de la probabilidad de selección de las viviendas y personas, según rama de actividad económica reducida.

Probabilidad de selección de personas dentro de la actividad										
Rama de Actividad económica reducida (CIIU rev. 4)[14]										
	1	**4**	**6**	**7**	**9**	**11**	**15**	**18**	**19**	**Total**
Recuento	1118	959	929	1879	735	567	473	196	687	7543
Moda	1,0000	1,0000	1,0000	0,5649	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Mínimo	0,4783	0,4800	0,5455	0,4286	0,5625	0,5000	0,7273	0,3333	0,5000	0,3333
Percentil 05	0,4958	0,6129	0,6383	0,5467	0,6667	0,6200	0,7941	0,6857	0,8000	0,5500
Percentil 25	0,5686	0,6458	0,7143	0,5649	0,7872	0,6897	0,9535	0,8235	0,9286	0,6296
Mediana	0,6552	0,7200	0,8913	0,6080	0,9259	0,8333	0,9762	1,0000	0,9583	0,7500
Percentil 75	0,8182	0,8519	0,9512	0,7083	1,0000	1,0000	1,0000	1,0000	1,0000	0,9512
Percentil 95	1,0000	1,0000	1,0000	0,8409	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Percentil 99	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Máximo	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Media	0,6915	0,7479	0,8370	0,6452	0,8784	0,8476	0,9496	0,9096	0,9519	0,7806

Fuente: Elaboración propia

Ahora bien, si se observa el cuadro 9 se obtendrán las probabilidades de selección para cada macrozona. En la Región Metropolitana estas probabilidades se encuentran concentradas en torno a 0,76.

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.

[14] 1: Agricultura, ganadería, caza y silvicultura, 4: Industrias manufactureras, 6: Construcción, 7: Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres. 9: Transporte, almacenamiento y comunicaciones, 11: Actividades inmobiliarias, empresariales y de alquiler, 15: Otras actividades de servicios comunitarios, sociales y personales, 18: Sector Primario, 19, Servicios.

Cuadro 9. Estadísticas descriptivas de la probabilidad de selección de las viviendas y personas, según rama de actividad económica reducida.

Probabilidad de selección de personas dentro de la actividad					
	Macrozona				
	Norte	**Centro**	**Sur**	**Metropolitana**	**Total**
Recuento	1512	2683	1742	1606	7543
Moda	1,0000	1,0000	1,0000	0,5649	1,0000
Mínimo	0,4909	0,3333	0,4285	0,5645	0,3333
Percentil 05	0,5060	0,5724	0,5000	0,5645	0,5500
Percentil 25	0,6734	0,6296	0,6458	0,5893	0,6296
Mediana	0,8000	0,7500	0,7631	0,7157	0,7500
Percentil 75	1,0000	0,9285	0,9473	0,9534	0,9512
Percentil 95	1,0000	1,0000	1,0000	1,0000	1,0000
Percentil 99	1,0000	1,0000	1,0000	1,0000	1,0000
Máximo	1,0000	1,0000	1,0000	1,0000	1,0000
Media	0,7945	0,7814	0,7830	0,7633	0,7805

Fuente: Elaboración propia

3.1.2. 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 periodo MAM 2015.

Así, calculadas las probabilidades de selección y participación de una vivienda en la ENE en el trimestre MAM 2015 y la probabilidad de seleccionar un independiente desde la EME, el ponderador base se calcula como:

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

En el cuadro 10 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. 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 10. Estadísticas descriptivas del ponderador base según Macrozona.

Estadísticas descriptivas	Macrozona				Total País
	Norte	Centro	Sur	Región Metropolitana	
Recuento	1.512	2.683	1.742	1.606	7.543
Moda	35,7	78,9	155,9	234	78,9
Mínimo	4,8	9	8,9	14,9	4,8
Percentil 05	15,2	26,2	29	64,6	25,6
Percentil 25	44,3	60,6	59,7	133,4	63,8
Mediana	81,1	110,6	107,5	227,4	117,9
Percentil 75	138,7	197,7	179,3	427,2	219,7
Percentil 95	343,7	504,7	374,8	1.071,30	610,4
Percentil 99	604,7	848	662,6	2.094,70	1.178,00
Máximo	1.108,4	1.681,2	1.594,9	5.408,7	5.408,7
Media	114,6	165,2	143,4	362,2	192
Error típico de la media	3,1	3,3	3,2	10,3	2,9
Suma	173.219,0	443.348,5	249.740,9	581.734,8	1.448.043,1

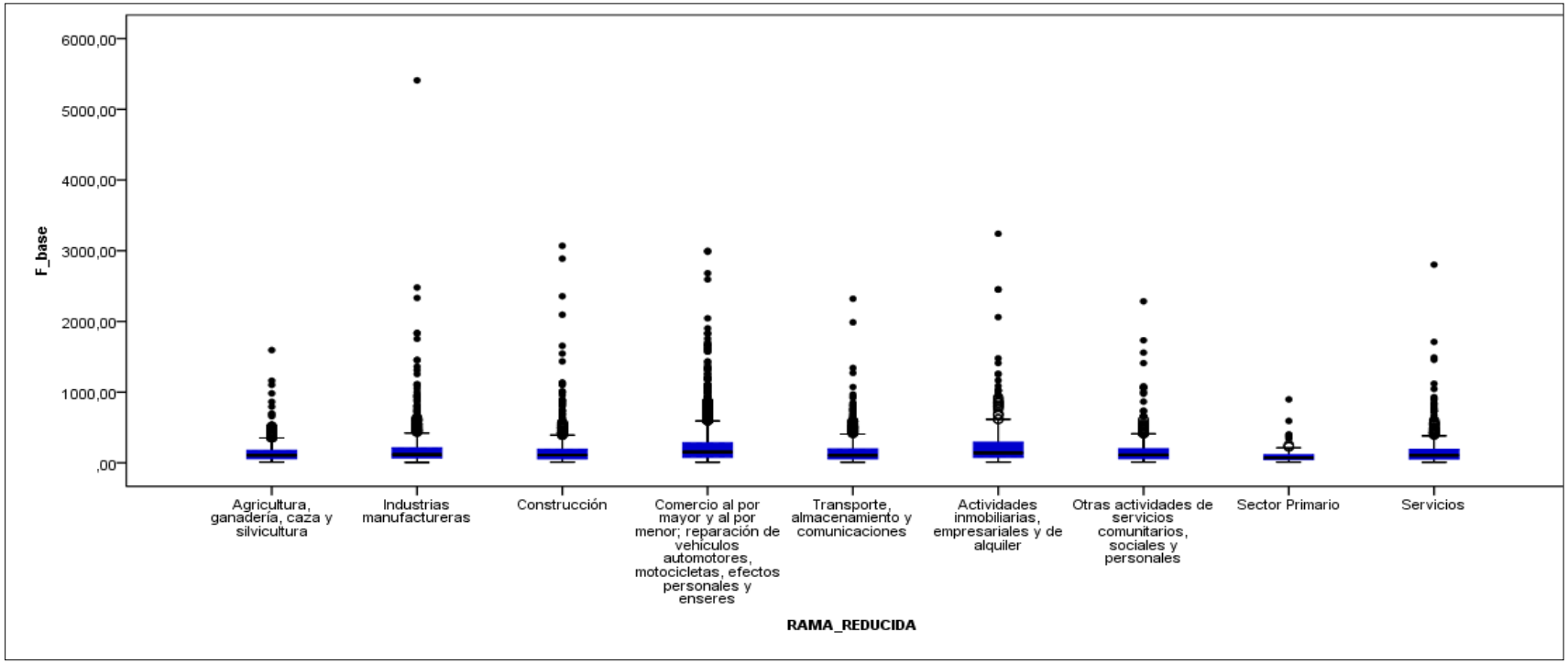
Fuente: Elaboración propia

Cuadro 11. Estadísticas descriptivas del ponderador base según Rama de actividad económica reducida.

Estadísticas Descriptivas	Rama de actividad económica reducida (CIIU rev. 4)[15]									
	1	**4**	**6**	**7**	**9**	**11**	**15**	**18**	**19**	**Total**
Recuento	1.118	959	929	1.879	735	567	473	196	687	7.543
Moda	78,9	110,9	40,2	234,0	58,0	168,6	39,3	16,1	50,1	78,9
Mínimo	8,9	4,8	9,1	7,3	8,7	8,9	10,2	10,5	7,0	4,8
Percentil 05	21,3	27,1	25,4	31,2	24,9	31,1	22,6	18,3	19,5	25,6
Percentil 25	59,3	70,1	58,0	78,1	57,0	76,0	59,5	45,9	54,9	63,8
Mediana	106,7	117,2	110,3	153,2	105,6	139,5	113,1	73,0	105,3	117,9
Percentil 75	178,8	213,7	193,4	284,7	198,6	293,0	200,8	114,5	191,5	219,7
Percentil 95	359,5	618,5	541,9	810,2	538,4	688,4	520,5	224,5	477,5	610,4
Percentil 99	537,8	1.313,1	1.110,8	1.574,6	919,3	1.410,7	1.082,8	592,4	929,9	1.178,0
Máximo	1.594,9	5.408,7	3.069,9	2.997,6	2.319,7	3.240,3	2.284,3	897,8	2.803,3	5.408,7
Media	139,8	197,8	177,2	244,8	173,2	234,7	178,7	94,5	166,0	192,0
Error Típ. de la media	3,8	9,5	8,2	6,7	7,6	12,4	10,4	6,8	8,0	2,9
Suma	156.318,9	189.649,2	164.649,1	459.981,7	127.270,9	133.086,3	84.526,6	18.521,3	114.039,1	1.448.043,1

[15] 1: Agricultura, ganadería, caza y silvicultura, 4: Industrias manufactureras, 6: Construcción, 7: Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres. 9 : Transporte, almacenamiento y comunicaciones, 11: Actividades inmobiliarias, empresariales y de alquiler, 15: Otras actividades de servicios comunitarios, sociales y personales, 18: Sector Primario, 19, Servicios.

Gráfico 1. Ponderador base según Rama de actividad económica reducida.



Fuente: Elaboración propia

También existen valores extremos en cada rama de actividad económica. Sin embargo, los casos más preocupantes son los considerados “casos influyentes”, pues las características de un individuo pueden representar hasta 5.408 personas, es decir al 0,92% de la población estimada (suma del ponderador) de la Región Metropolitana. Para minimizar el efecto en las estimaciones de ponderadores de ésta 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.

3.1.3. Suavizamiento de Ponderador Base

En la etapa de construcción del ponderador base, se observó 1 caso con valor mayor a las 5.000 unidades, los que en conjunto representan a un 0,06% de los independientes de la Región Metropolitana. Las restantes observaciones poseen ponderadores inferiores o iguales a 3.240. 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: un caso extremo (los que se encuentran al interior de la elipse continua) que superan las 5.000 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.

Gráfico 2. Dispersión del Factor de expansión base o inicial

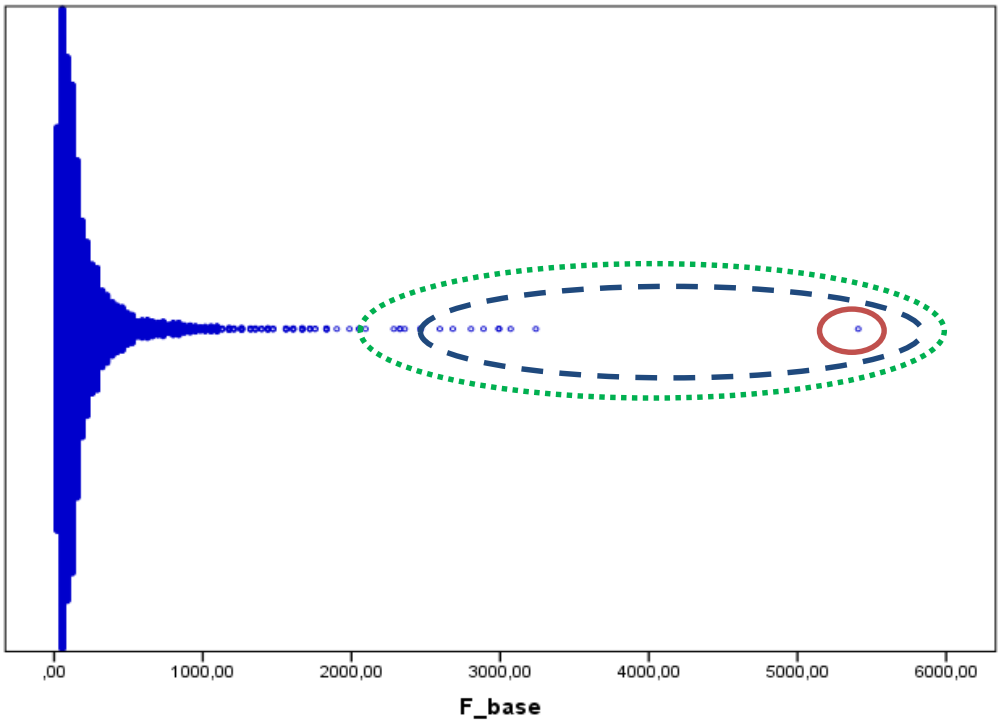
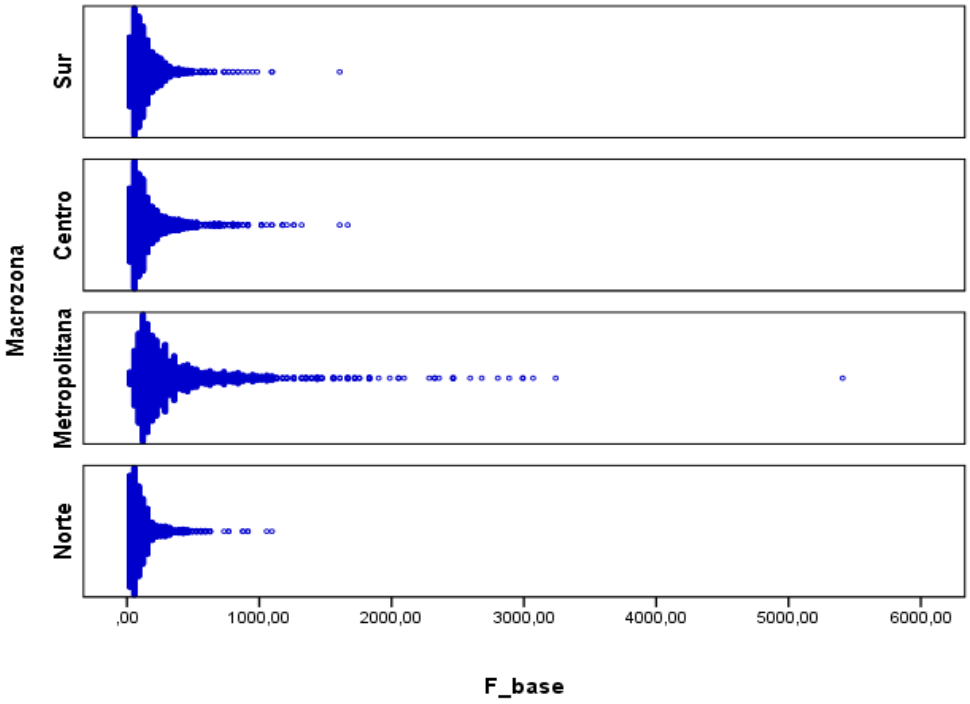
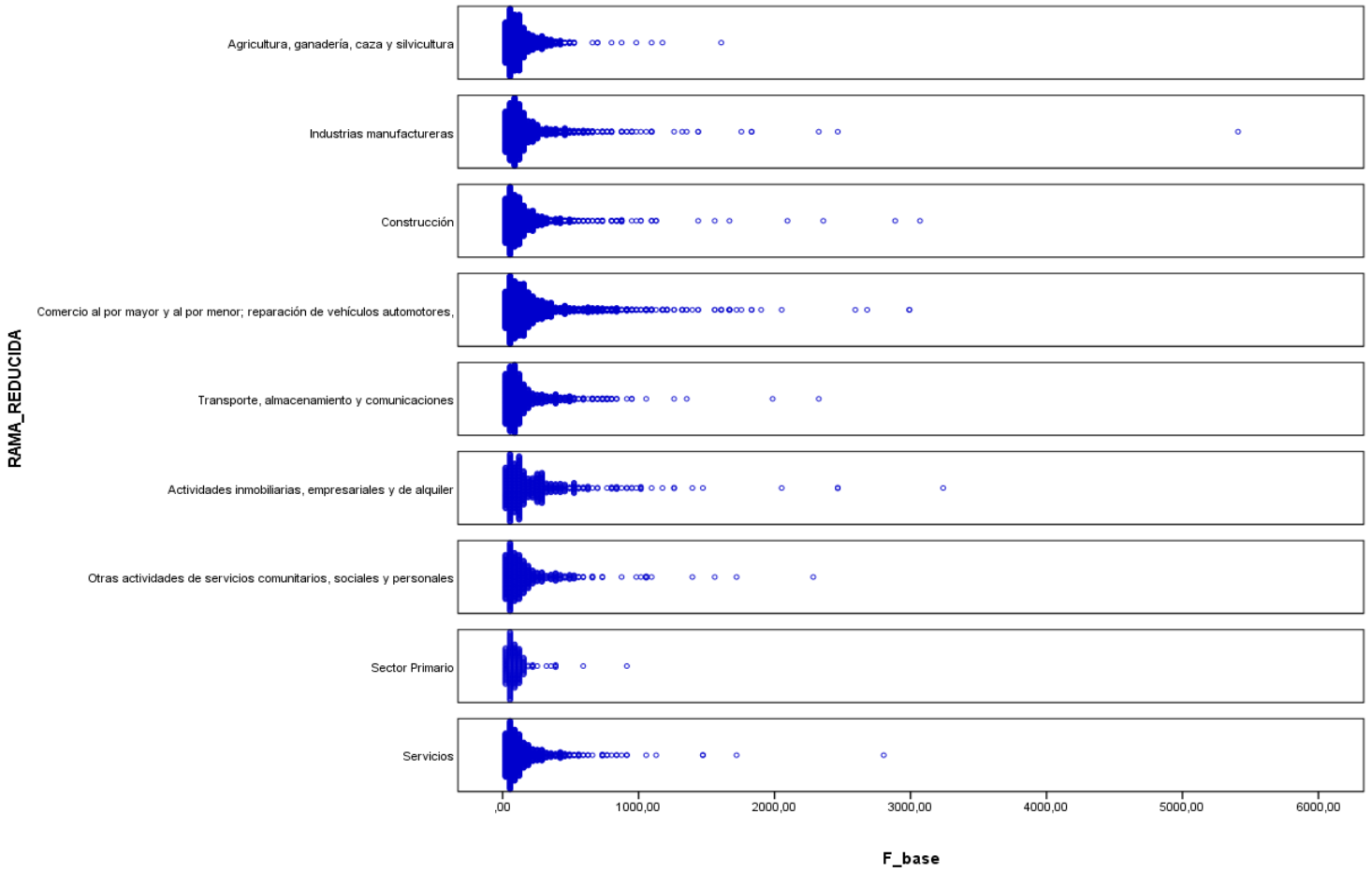


Gráfico 3. Dispersión del Factor de expansión base o inicial, según macrozona



Fuente: Elaboración propia

Gráfico 4. Dispersión del Factor de expansión base o inicial, según rama de actividad económica



Para inspeccionar la existencia de valores extremos, se utilizaron dos estrategias: (¡) 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 rama de actividad económica reducida[16].

Considerando lo anterior, se analizaron 7 puntos de cortes distintos definidos como sigue:

$$\beta_r = r * \overline{F} \ , \quad con \ \overline{F} = \ \bar{F}^{base}_{Rjk,g} \ , \ con \ r \ = \ 4,5,6,7,8,9,10$$

De otra forma, los 7 puntos de corte son:

$$\beta_r = \begin{cases} \beta_4 \ = \ 4 * \overline{F} \\ \beta_5 \ = \ 5 * \overline{F} \\ \beta_6 \ = \ 6 * \overline{F} \\ \beta_7 \ = \ 7 * \overline{F} \\ \beta_8 \ = \ 8 * \overline{F} \\ \beta_9 \ = \ 9 * \overline{F} \\ \beta_{10} = \ 10 * \overline{F} \end{cases}$$

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

$$T_{Rjk,g} = \begin{cases} F^{base}_{Rjk} & si \ \ F^{base}_{Rjk} \leq \beta_r \\ \beta_r & si \ F^{base}_{Rjk} > \beta_r \end{cases}$$

Donde,

g: Subíndice de la Rama de Actividad Económica Reducida, de procedencia de los independientes.

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

$T_{Rjk,g}$: Es el ponderador base truncado de la Región R, vivienda j persona k, perteneciente a 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 debe distribuir la diferencia faltante en el resto de los ponderadores que no fueron truncados. Los pesos fueron distribuidos al interior de cada Rama g de la siguiente forma:

[16] Al chequear gráfico 4 se observó que el comportamiento de los ponderadores es distinto al interior de cada rama de actividad económica, por lo tanto se determinó realizar el suavizamiento de forma independiente al interior de cada uno de estos.

$$F_{Rjk,g}^{Sr} = \begin{cases} F_{Rjk}^{base} \cdot \dfrac{\left(\sum_{k \in g} F_{Rjk}^{base} - \sum_{\in g \cap F_{Rjk}^{base} > \beta_r} \beta_r\right)}{\sum_{k \in g \cap F_{Rjk}^{base} \leq \beta_r} F_{Rjk}^{base}} & , \; si \; F_{Rjk}^{base} \leq \beta_r \\ \beta_r & , \quad si \; F_{Rjk}^{base} > \beta_r \end{cases}$$

Donde $F_{Rjk,g}^{SR}$ es el factor suavizado del individuo k de la vivienda j en la región R de la Rama de Actividad Económica Reducida g.

Esto es, aquellos ponderadores identificados como valores extremos son truncados al valor máximo establecido ($\beta_r = r * \overline{F}$), mientras que el peso "sobrante" de los ponderadores truncados es distribuido sobre el resto de los ponderadores.

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.).

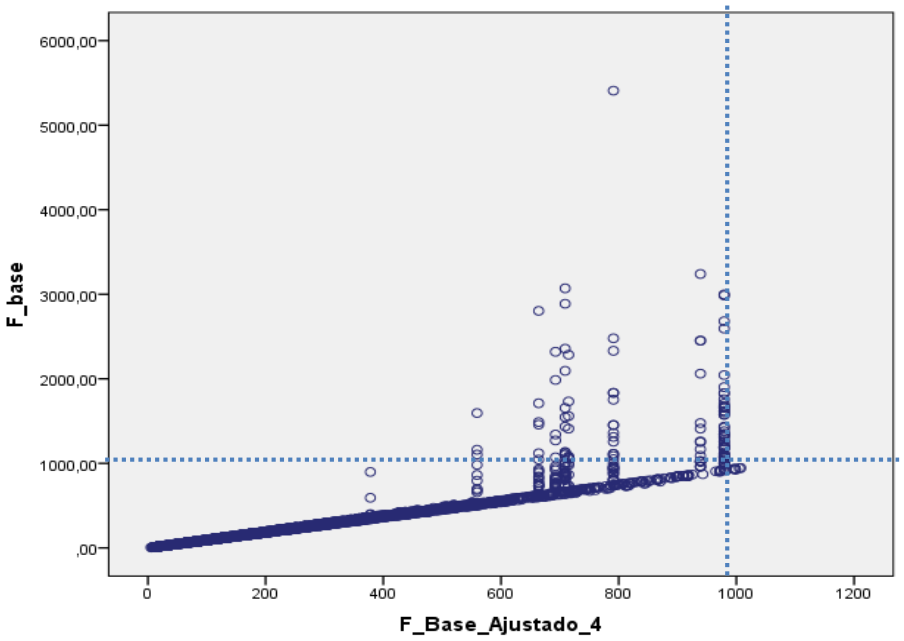
Cuadro 12. Estadísticas descriptivas del ponderador base y ponderador base suavizados en distintos puntos de corte

Estadístico		Ponderador base	Ponderador Suavizado k4	Ponderador Suavizado k5	Ponderador Suavizado k6	Ponderador Suavizado k7	Ponderador Suavizado k8	Ponderador Suavizado k9	Ponderador Suavizado k10
Rango		5.404	1.002	1.262	1.464	1.723	1.953	2.198	2.443
Mínimo		5	5	5	5	5	5	5	5
Máximo		5.409	1.008	1.268	1.469	1.728	1.958	2.203	2.448
Suma		1.448.043	1.448.043	1.448.043	1.448.043	1.448.043	1.448.043	1.448.043	1.448.043
Media	Estimación	192	192	192	192	192	192	192	192
	Error Típico	3	2	2	2	3	3	3	3
Desviación estándar		249	190	203	212	219	224	229	233
Varianza		61.849	35.974	41.282	45.026	48.025	50.284	52.355	54.110
Asimetría	Estimación	5	2	2	3	3	3	4	4
	Error Típico	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
Curtosis	Estimación	54	4	7	9	12	15	17	21
	Error Típico	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06

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 5, 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.

Gráfico 5. Dispersión del ponderador base versus ponderador suavizado en corte igual a K_4.



Fuente: Elaboración propia

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[17]) y sobre estas categorías se calculó el sesgo utilizando la siguiente fórmula:

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

[17] Ver más detalles en capítulo 4.2

Tras 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.

Cuadro 13. Estimación del sesgo de la estructura de la rama de actividad económica.

	Sesgo						
Rama Actividad Reducida	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
A. Agricultura, Ganadería, caza y silvicultura	0,000	0,007	0,011	0,008	0,001	0,021	0,006
D. Industrias manufactureras	0,033	0,001	0,000	0,009	0,013	0,009	0,001
F. Construcción	0,004	0,024	0,001	0,016	0,028	0,003	0,001
G. Comercio al por mayor y al por menor; reparación de vehículos automotores	0,003	0,003	0,000	0,010	0,009	0,005	0,011
I. Transporte, almacenamiento y comunicaciones	0,000	0,000	0,003	0,003	0,000	0,011	0,009
K. Actividades inmobiliarias, empresariales y de alquiler	0,001	0,005	0,009	0,005	0,000	0,016	0,004
O. Otras actividades de servicios comunitarios, sociales y personales	0,042	0,001	0,000	0,011	0,016	0,010	0,003
Sector Primario	0,005	0,028	0,002	0,018	0,024	0,003	0,001
Servicios	0,002	0,004	0,000	0,011	0,011	0,006	0,008
Total	**0,001**	**0,000**	**0,002**	**0,003**	**0,000**	**0,012**	**0,010**

Fuente: Elaboración Propia.

Finalmente, se calcula el ECM para cada categoría como:

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

Al revisar el cuadro 13, se observa que en términos del ECM no existen diferencias entre los puntos de suavizamiento.

Cuadro 14. Estimación del ECM de la estructura de la rama de actividad económica.

	ECM						
Rama Actividad Reducida	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
A Agricultura, ganadería, caza y silvicultura	0,00075	0,007258	0,011241	0,007914	0,001019	0,020761	0,006114
D Industrias manufactureras	0,033365	0,002858	0,000933	0,008939	0,013287	0,009428	0,001782
F Construcción	0,004739	0,024829	0,001401	0,018131	0,028491	0,003501	0,001864
G Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	0,003458	0,003357	0,00051	0,011096	0,009214	0,007237	0,01166
I Transporte, almacenamiento y comunicaciones	0,000804	0,00044	0,003267	0,003293	0,000559	0,011775	0,009572
K Actividades inmobiliarias, empresariales y de alquiler	0,001424	0,005243	0,009119	0,005231	0,000869	0,016629	0,00409
O Otras actividades de servicios comunitarios, sociales y personales	0,04203	0,003625	0,000607	0,011285	0,016824	0,010833	0,003549
Sector primario	0,005438	0,028679	0,002067	0,021343	0,024265	0,003344	0,001559
Servicios	0,002483	0,004045	0,000646	0,012346	0,01133	0,00868	0,008782
Mediana	0,003458	0,004045	0,001401	0,011096	0,01133	0,009428	0,00409
Promedio	0,010499	0,008926	0,00331	0,011064	0,011762	0,010243	0,005441

Fuente: Elaboración Propia.

En el cuadro 15 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 5.409 disminuye a 4.469 unidades.

Cuadro 15. Estadísticas descriptivas del ponderador base y ponderador suavizado.

Estadísticas descriptivas	MACROZONA									
	Norte		Centro		Sur		Metropolitana		Total	
	Ponderador base	Ponderador K6	Ponderador base	Ponderador K6	Ponderador base	Ponderador K6	Ponderador base	Ponderador K6	Ponderador base	Ponderador K6
Recuento	1.512	1.512	2.683	2.683	1.742	1.742	1.606	1.606	7.543	7.543
Moda	36	36	79	80	156	158	234	1.469	79	1.469
Mínimo	5	5	9	9	9	9	15	15	5	5
Percentil 05	15	15	26	27	29	30	65	66	26	27
Percentil 25	44	46	61	63	60	61	133	139	64	66
Mediana	81	84	111	115	108	111	227	236	118	122
Percentil 75	139	143	198	205	179	187	427	443	220	227
Percentil 95	344	352	505	518	375	381	1.071	1.061	610	635
Percentil 99	605	610	848	881	663	693	2.095	1.469	1.178	1.072
Máximo	1.108	1.132	1.681	1.469	1.595	1.174	5.409	1.469	5.409	1.469
Media	115	118	165	170	143	147	362	347	192	192
Error Típ. de la media	3	3	3	3	3	3	10	8	3	2
Suma	173.219	178.343	443.348	456.231	249.741	256.104	581.735	557.365	1.448.043	1.448.043

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 IV 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 IV 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 IV EME, de un total de 6.880 viviendas seleccionadas, se seleccionaron 7.543 trabajadores independientes. De éstos, 7.319 fueron clasificados como elegibles (97,03%), de los cuales 6.488 respondieron la encuesta[18]. Por lo tanto, la tasa de respuesta de la EME, ajustada por elegibilidad, es de 88,65%.

[18] De los cuales 6.485 respondieron el cuestionario de forma completa y 3 de forma parcial. Para mayor detalle ver Anexo N° 2

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 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[19], este método consiste en modelar la probabilidad de respuesta en la IV 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[20] (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 IV 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 IV EME, el error asociado a la estimación de

[19] Mayor detalle ver Valliant, R. Drever, J. Kreuter, F. (2013, section 13.5) "Practical Tools for Designing and Weighting Survey Samples", New York. Springer.
[20] 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.

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

De acuerdo a lo anterior y según el comportamiento de los datos de la IV 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[21].

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,

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

[21] Mayor detalle ver Anexo N°2

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 8 se presentan las tasas de respuesta para cada una de estas celdas, así como también el factor de ajuste por no respuesta ($\hat{R}_c^{NR}$). Se observa que el grupo 5 presenta menor tasa de respuesta, por lo que cada factor base se incrementó en un 14% aproximadamente.

Cuadro 16. Total unidades elegibles, que responde y tasa de respuesta.

Celda Ajuste	Total Responde	Total Elegibles	Tasa de Respuesta	$\hat{R}_c^{NR}$
Total	6.488	7.319	88,65%	1,13
1	1.117	1.219	91,60%	1,09
2	1.096	1.220	89,80%	1,11
3	1.091	1.220	89,40%	1,12
4	1.085	1.220	88,90%	1,12
5	1.068	1.220	87,50%	1,14

Fuente: Elaboración Propia

En el cuadro 16 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 13,6%, observándose en la rama Otras actividades de servicios comunitarios, sociales y personales el mayor crecimiento promedio de los ponderadores (17%). Por otro lado, el mayor ponderador se encuentra en la Rama Actividades inmobiliarias, empresariales y de alquiler el cual no supera las 1.700 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 17. Estadísticas descriptivas del ponderador ajustado por falta de respuesta.

Estadísticas descriptivas			Recuento	Moda	Mínimo	Percentil 05	Percentil 25	Mediana	Percentil 75	Percentil 95	Percentil 99	Máximo	Media	Error Típ. media	Suma
Rama de actividad económica reducida[22]	1	F_{Rjl}^{Tr}	1.068	80	9	22	60	108	180	363	543	839	140	4	149.138
		F_{Rjl}^{NR}	978	166	10	27	67	120	195	383	627	912	152	4	148.695
	4	F_{Rjl}^{Tr}	934	1.187	5	29	73	124	224	643	1.187	1.187	196	7	183.004
		F_{Rjl}^{NR}	855	124	5	32	80	136	255	710	1.253	1.306	218	8	186.061
	6	F_{Rjl}^{Tr}	907	1.063	10	27	61	116	203	571	1.063	1.063	176	6	159.755
		F_{Rjl}^{NR}	782	1.184	12	31	67	129	227	656	1.159	1.188	197	8	154.356
	7	F_{Rjl}^{Tr}	1.827	1.469	7	32	78	156	291	830	1.469	1.469	245	6	447.177
		F_{Rjl}^{NR}	1.654	1.639	8	34	86	172	328	929	1.586	1.639	273	7	451.827
	9	F_{Rjl}^{Tr}	717	1.039	9	25	58	108	205	561	941	1.039	174	7	125.047
		F_{Rjl}^{NR}	605	488	11	29	68	125	241	660	1.118	1.224	204	9	123.434
	11	F_{Rjl}^{Tr}	546	1.408	9	32	76	145	298	788	1.408	1.408	236	11	128.586
		F_{Rjl}^{NR}	457	1.659	12	38	92	175	344	903	1.503	1.659	276	13	126.354
	15	F_{Rjl}^{Tr}	463	1.072	11	23	62	119	209	529	1.072	1.103	178	9	82.528
		F_{Rjl}^{NR}	403	1.253	12	25	70	132	242	638	1.253	1.253	209	12	84.150
	18	F_{Rjl}^{Tr}	186	16	11	19	49	76	117	218	567	567	96	6	17.850
		F_{Rjl}^{NR}	167	74	12	20	53	82	129	265	618	659	110	8	18.331
	19	F_{Rjl}^{Tr}	671	996	7	20	57	109	204	509	962	996	167	7	111.886
		F_{Rjl}^{NR}	587	1.142	9	22	62	127	233	588	1.076	1.142	190	8	111.763
	Total	F_{Rjl}^{Tr}	7.319	1.469	5	26	66	122	226	635	1.072	1.469	192	2	1.404.969
		F_{Rjl}^{NR}	6.488	1.639	5	30	74	135	257	714	1.241	1.659	217	3	1.404.969

[22] 1: Agricultura, ganadería y silvicultura, 4: Industrias manufactureras, 6: Construcción, 7: Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres. 9: Transporte, almacenamiento y comunicaciones, 11: Actividades inmobiliarias, empresariales y de alquiler, 15: Otras actividades de servicios comunitarios, sociales y personales, 18: Sector Primario, 19: Servicios.

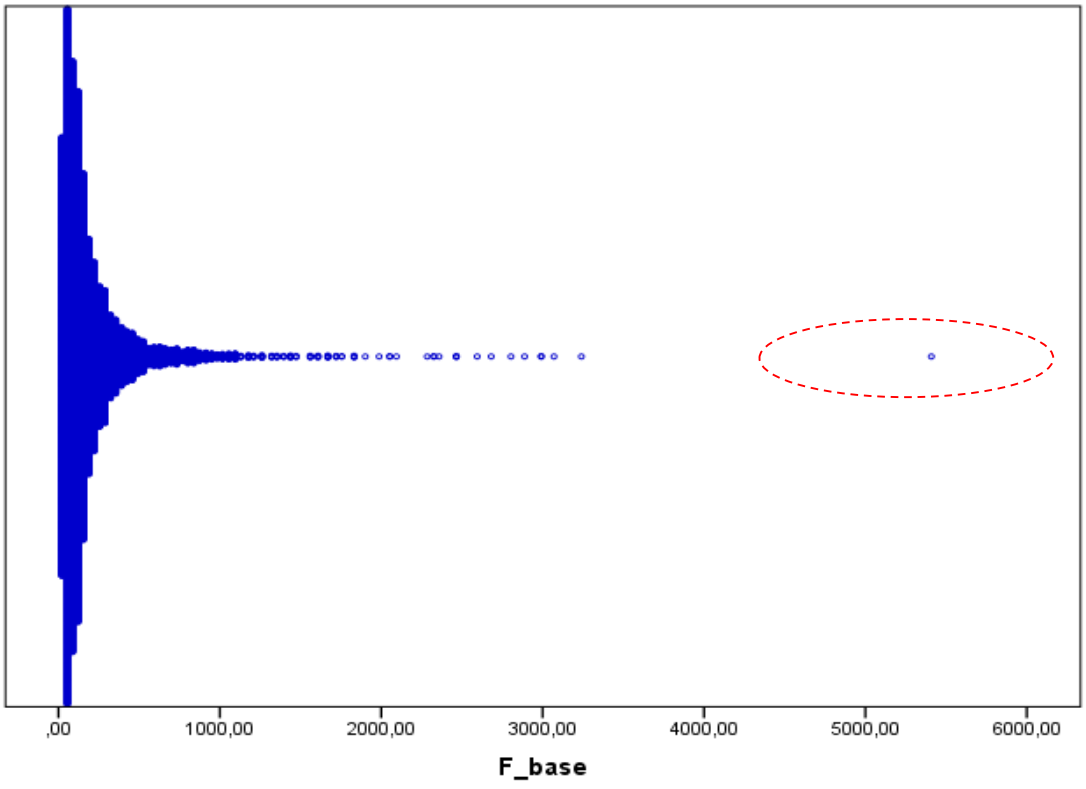
3.2.1. Suavizamiento del Ponderador ajustado por falta de respuesta

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 K_6.

En el gráfico 6 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 17 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[23]. En consecuencia, al interior de cada rama de actividad económica no fue necesario realizar suavizamiento.

Gráfico 6. Distribución de Factor ajustado por falta de respuesta.



Fuente: Elaboración propia

[23] Según Rama de Actividad Económica reducida: 1= 0,16; 4=0,16; 6=0,16; 7=0,16; 7:0,16; 9=0,16; 11=0,16; 15=0,16; 18=0,16; 19=0,13.

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[24] 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 periodo de levantamiento.

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 IV EME está anclada a la población de referencia del trimestre MAM 2015 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 2015 de la ENE actualizada al período del trabajo de campo de la IV 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 2015. 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 2015.
2. Segundo, se calcula un nuevo factor de expansión, considerando las proyecciones de población a Junio del 2015.

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

En el período MAM 2015, 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 ENEh.

P_{hs}^{4}: Total de población del sexo s, del estrato ENEh, proyectado al 15 de abril de 2015.

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 2015. 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 10, se presenta el total de independientes estimado a partir de la publicación de la ENE, periodo MAM 2015; y según total de personas estimado con la información levantada en MAM 2015, pero con proyecciones actualizadas a la fecha de levantamiento de la EME (en adelante I_{gs}.)

Como se observa en el cuadro 18, el total de “trabajadores independientes” estimados y publicados oficialmente son 1.950.422 personas. Sin embargo, al actualizar las proyecciones de población este total asciende a 1.955.539, lo que equivale a un incremento del 0,26% a nivel nacional, un 0,34% en la macrozona Norte, un 0,34% en el Centro, un 0,29% en el Sur un 0,34% y un 0,16% en la Región Metropolitana.

Cuadro 18. Total de independientes estimado a partir de la ENE- Período MAM 2015

Macrozona	Sexo	Total Independientes	
		Factor Expansión Oficial ENE - MAM	Factor Expansión Información ENE - ajustado Junio
Total	Hombre	1.204.861	1.208.132
	Mujer	745.561	747.407
	Total	1.950.422	1.955.539
Norte	Hombre	150.681	151.213
	Mujer	86.913	87.211
	Total	237.593	238.424
Centro	Hombre	360.781	361.905
	Mujer	231.003	231.649
	Total	591.784	593.553
Sur	Hombre	240.438	241.280
	Mujer	121.121	121.538
	Total	361.560	362.818
Metropolitana	Hombre	452.961	453.734
	Mujer	306.524	307.010
	Total	759.485	760.744

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 2015. 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 2015, actualizado al crecimiento poblacional de junio 2015 - mes central de levantamiento de EME.

En el cuadro 19 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 1.639 personas, sin embargo al ajustar según sexo y macrozona, esta persona representa 2.400 individuos. Cabe señalar, que en el caso de un hombre de la misma macrozona su ponderador cambió de 1659 a 2.234, lo cual puede ser revisado en el cuadro 19.

Cuadro 19. 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	4.113	4.113	2.375	2.375	6.488	6.488
Moda	74	105	1.639	193	1.639	105
Mínimo	9	12	5	7	5	7
Percentil 05	31	43	28	39	30	41
Percentil 25	75	104	71	100	74	102
Mediana	136	187	133	191	135	188
Percentil 75	258	350	255	368	257	359
Percentil 95	696	953	752	1.080	714	993
Percentil 99	1.184	1.599	1.279	1.855	1.241	1.737
Máximo	1.659	2.234	1.639	2.400	1.659	2.400
Media	215	294	219	315	217	301
Error típico de la media	4	5	5	8	3	4
Suma	885.756	1.208.132	519.213	747.407	1.404.969	1.955.539

Fuente: Elaboración propia

Cuadro 20. 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	1.292	1.292	2.322	2.322	1.539	1.539	1.335	1.335	6.488	6.488
Moda	60	85	84	113	166	105	1.639	1.595	1.639	105
Mínimo	5	7	11	15	12	17	18	24	5	7
Percentil 05	18	25	31	41	34	50	75	103	30	41
Percentil 25	52	71	69	94	68	100	163	229	74	102
Mediana	94	131	126	171	122	175	284	386	135	188
Percentil 75	157	217	229	308	205	298	540	750	257	359
Percentil 95	403	552	578	783	411	602	1.199	1.687	714	993
Percentil 99	707	964	954	1.283	823	1.165	1.599	2.234	1.241	1.737
Máximo	1.309	1.834	1.639	2.268	1.306	1.928	1.659	2.400	1.659	2.400
Media	133	185	189	256	161	236	409	570	217	301
Error típico de la media	4	5	4	5	4	5	10	14	3	4
Suma	171.959	238.424	437.937	593.553	248.455	362.818	546.618	760.744	1.404.969	1.955.539

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 4 veces o más el factor promedio de la macrozona), 190 observaciones debería ser truncadas, mientras que bajo el criterio utilizado 48 observaciones deberían ser truncadas, como se ilustra en el cuadro 20.

Cuadro 21. Número de observaciones a truncar según criterio o punto de corte

	Valor extremo		
Punto corte	NO	Si	Total
k_4	7353	190	7543
k_5	7453	90	7543
k_6	7495	48	7543
k_7	7516	27	7543
k_8	7533	10	7543
k_9	7537	6	7543
k_{10}	7540	3	7543

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.

4. ESTIMACIÓN DE VARIANZA

De acuerdo a lo descrito en los apartados anteriores, el diseño muestral de la IV 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 IV 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 IV 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 IV 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 IV 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[25], 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

[25] 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 IV EME se seleccionaron independientes desde 159 estratos, de los cuales 4 de ellos contienen independientes seleccionadas en solo un conglomerado, además uno de estos cuatro contiene a un solo independiente. Sin embargo, existen 65 estratos con 30 o menos independientes. Así el total de pseudo-estratos creados de acuerdo a los criterios anteriores desciende a 109 unidades.

Cuadro 22. Total de estratos y de Pseudo-estratos, según macrozona.

Macrozona	Estratos	Pseudo-estrato
Total	159	109
Norte	25	20
Centro	62	45
Sur	25	19
Metropolitana	47	25

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 2.791 de estos conglomerados se seleccionaron trabajadores independientes, los que se transformaron en 473 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 23 se expone el total de conglomerados y pseudo-conglomerados según macrozona.

Cuadro 23. Total de conglomerados y de pseudo-conglomerados, según macrozona

Macrozona	Conglomerados	Pseudo-Conglomerados
Total	2.791	473
Norte	507	93
Centro	1094	174
Sur	585	104
Metropolitana	605	102

Fuente: Elaboración propia

4.2. Estimación de variables y varianzas en SPSS

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.

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 IV 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 tabla 1.

Tabla 1. 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.

Cuadro 24. Estructura de la Actividad económica en la cual se desenvuelven los trabajadores independientes- estimación realizada en SPSS

Rama de actividad económica	Estimación	Error estándar	95% de intervalo de confianza		Coeficiente de variación
			Inferior	Superior	
1 Agricultura, ganadería, caza y silvicultura	10,7%	,5%	9,7%	11,8%	,050
4 Industrias manufactureras	13,3%	,6%	12,2%	14,6%	,044
6 Construcción	10,8%	,5%	9,8%	11,8%	,047
7 Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres	32,2%	,9%	30,5%	34,1%	,029
9 Transporte, almacenamiento y comunicaciones	8,7%	,5%	7,7%	9,8%	,060
11 Actividades inmobiliarias, empresariales y de alquiler	8,9%	,6%	7,7%	10,2%	,073
15 Otras actividades de servicios comunitarios, sociales y personales	6,0%	,5%	5,2%	7,0%	,078
18 Sector Primario	1,3%	,2%	1,0%	1,7%	,140
19 Servicios	8,1%	,5%	7,1%	9,1%	,064
Total	**100,0%**	**0,0%**	**100,0%**	**100,0%**	**0,000**

Fuente: Elaboración Propia

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, Construcción y Agricultura, actividades que en conjunto reúnen a más del 57% de los trabajadores independientes.

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

Código de disposición de la última visita a la vivienda	Frecuencia	Porcentaje	Tipo de Legibilidad
11 Entrevista lograda	6485	85,97	Elegibles
12 Entrevista lograda de forma parcial	3	0,04	Elegibles
21 Se rechazó la entrevista	143	1,90	Elegibles
22 Vivienda ocupada sin moradores presentes	254	3,37	Elegibles
23 Informante inubicable por cambio de domicilio, fuera del país o motivos laborales	402	5,33	Elegibles
24 Muerte del informante	4	0,05	Elegibles
25 Informante con dificultad física, mental o cognitiva para contestar	28	0,37	Elegibles
28 Fuera de Marco	15	0,20	No Elegibles
27 Fuera de Muestra	130	1,72	No Elegibles
31 Se impidió acceso a la vivienda (administrador, conserje o junta de vigilancia niega acceso)	4	0,05	Elegibilidad Desconocida
32 No fue posible localizar la dirección	5	0,07	Elegibilidad Desconocida
33 Área de difícil acceso o peligrosa	40	0,53	Elegibilidad Desconocida
41 Inmueble para uso no habitacional (empresa, oficina, vivienda colectiva, institución pública, etc.)	2	0,03	No Elegibles
42 Vivienda en demolición, incendiada, destruida o erradicada	2	0,03	No Elegibles
43 Vivienda particular desocupada (en arriendo, en venta, otro.)	25	0,33	No Elegibles
44 Vivienda de uso temporal (vacaciones, descanso, etc.)	1	0,01	No Elegibles
Total	**7543**	**100,00**	**7319**

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 IV 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. Regresión Logística

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 relación es de tipo no lineal. Para resolver esto, se necesita una función g que relacione la respuesta

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

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.

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

---

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 $f(y;\pi) = \exp[a(y)b(\pi) + c(\pi) + d(y)]$, 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 $b(\pi_i) = \log(\pi_i/1-\pi_i)$ 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)]

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

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:

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

3.2.2. 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].

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

[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)]

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

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

3.2.3. 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.

3.2.4. 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:

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

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 IV 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 $HQIC = -2\log L\left(\hat{\theta}, \boldsymbol{y}\right) + 2\ln(\ln(n))p$ 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.

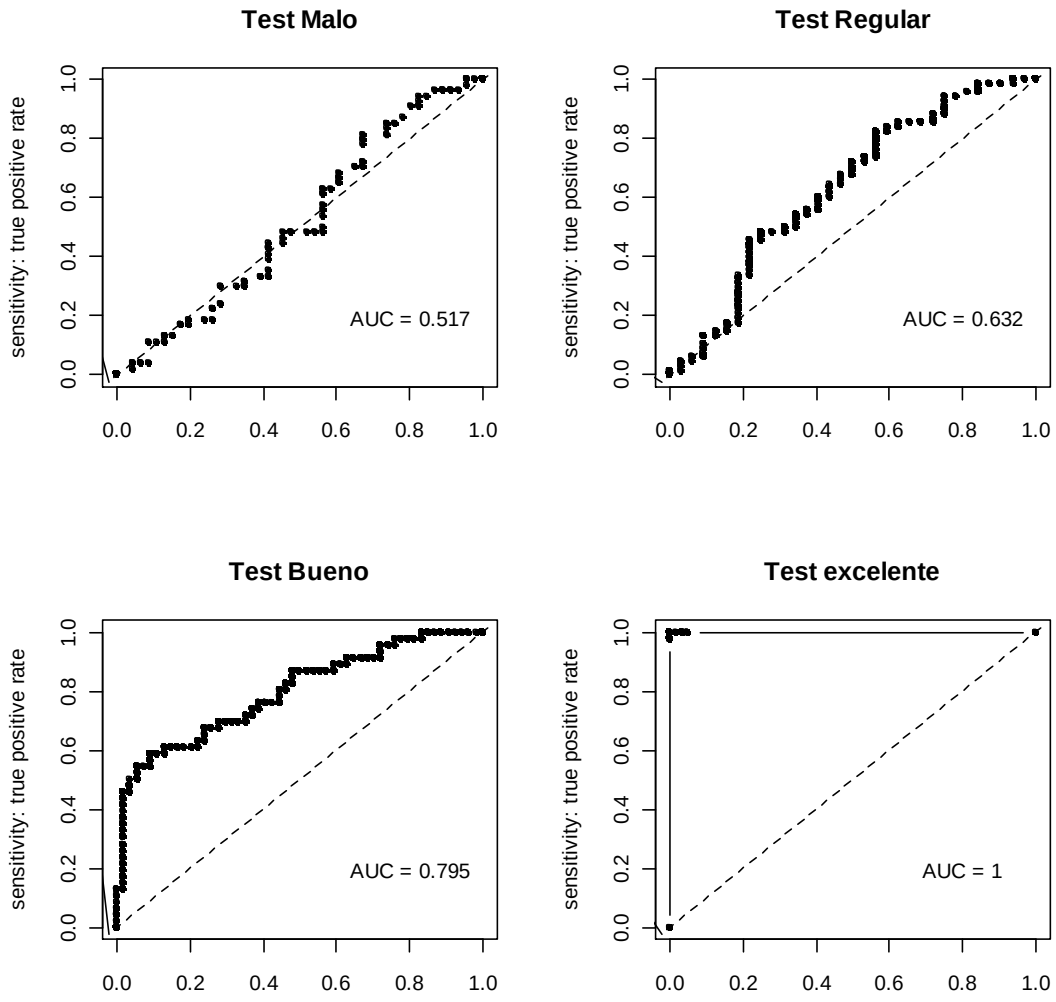
La elección se realiza mediante la comparación del área bajo la curva (AUC, de su acrónimo en inglés Area Under the Curve) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. 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 7. Diferentes curvas ROC



Fuente: Elaboración propia

3.5. Análisis de Elegibilidad

Para modelar la probabilidad de que una persona conteste la encuesta de la IV 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 27. 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	%	Elegible	La Persona responde
11. Entrevista Lograda	**6485**	**85,97%**	**Sí**	**Sí**
12. Entrevista lograda de forma parcial	**3**	**0,04%**	**Sí**	**Sí**
21. Se rechazó la entrevista	143	1,90%	Sí	No
22. Vivienda ocupada sin moradores presentes	254	3,37%	Sí	No
23. Informante inubicable por cambio de domicilio, fuera del país o motivos laborales	402	5,33%	Sí	No
24. Muerte del informante	4	0,05%	Sí	No
25. Informante con dificultad física, mental o cognitiva para contestar	28	0,37%	Sí	No
27. Fuera de Muestra	130	1,72%	No	-
28. Fuera de Marco	15	0,00%	No	-
31. Se impidió acceso a la vivienda (administrador, conserje o junta de vigilancia niega acceso)	4	0,05%	No	-
32. No fue posible localizar la dirección	5	0,07%	No	-
33. Área de difícil acceso o peligrosa	40	0,53%	No	-
41. Inmueble para uso no habitacional (empresa, oficina, vivienda colectiva, institución pública, etc)	2	0,03%	No	-
42. Vivienda en demolición, incendiada, destruida o erradicada	2	0,03%	No	-
43. Vivienda particular desocupada (en arriendo, en venta, otro.)	25	0,33%	No	-
44. Vivienda de uso temporal (vacaciones, descanso, etc.)	1	0,01%	No	-
Total	**7543**	**100%**	**7319**	**6488**

Fuente: Elaboración propia

En base a este cuadro se dividen las unidades elegibles (7.319) de las que no (224) y dentro de las unidades elegibles clasificamos las personas que responden (6.488) versus las que no (831).

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

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	400	128	528
	Media	202	89	291
	Superior	8	4	12
Total No		**610**	**221**	**831**
Sí	Básica	3.139	1.826	4.965
	Media	938	535	1.473
	Superior	33	17	50
Total Sí		**4.110**	**2.378**	**6.488**
Total general		**4.720**	**2.599**	**7.319**

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

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
	Básica	5,5%	1,7%	7,2%
	Media	2,8%	1,2%	4,0%
No	Superior	0,1%	0,1%	0,2%
Total No		**8,3%**	**3,0%**	**11,4%**
	Básica	42,9%	24,9%	67,8%
	Media	12,8%	7,3%	20,1%
Sí	Superior	0,5%	0,2%	0,7%
Total Sí		**56,2%**	**32,5%**	**88,6%**
Total general		**64,5%**	**35,5%**	**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 5,5% de los casos que no respondieron pertenecen al nivel educacional básica y son hombres versus 42,9% de las personas que responden en la misma categoría.

Cuadro 30. 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**
	Básica	75,76%	24,24%	100,00%
No	Media	69,42%	30,58%	100,00%
	Superior	66,67%	33,33%	100,00%
Total No		73,41%	26,59%	100,00%
	Básica	63,22%	36,78%	100,00%
Sí	Media	63,68%	36,32%	100,00%
	Superior	66,00%	34,00%	100,00%
Total Sí		63,35%	36,65%	100,00%
Total general		64,49%	35,51%	100,00%

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 75,76% de los casos pertenecen al sexo Hombre y el 24,24% Mujer. Para los que pertenecen al nivel educacional medio 69,42% son

hombres y 30,58% son mujeres. Finalmente del nivel educacional superior el 66,67% son hombres y el 33,33% 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 63,22% son hombres y el 36,78% son mujeres. En Media el 63,68% son hombres y el 36,32% son mujeres. En el caso del nivel educacional superior el 66% son hombres y el 34% mujeres.

Cuadro 31. 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	65,6%	57,9%	63,5%
	Media	33,1%	40,3%	35,0%
	Superior	1,3%	1,8%	1,4%
Total No		**100,0%**	**100,0%**	**100,0%**
Sí	Básica	76,4%	76,8%	76,5%
	Media	22,8%	22,5%	22,7%
	Superior	0,8%	0,7%	0,8%
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 65,6%, el que pertenece a la educación Media es 33,1% y a la educación superior es 1,3%. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 57,9% pertenece a la educación Básica, el 40,3% a la Media y el 1,8% al nivel Superior. Dentro de todas las personas que responden, se puede observar que dado que son hombres; el 76,4% de los casos pertenece al nivel educacional básico, el 22,8% Media y el 0,8% a nivel Superior. En este mismo sentido, en el caso de las mujeres 76,8% pertenece al nivel educacional Básica, 22,5% Media y 0,7% 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. (76,4% versus 76,8% 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 IV EME tienen una relación directa con el estado "Casado(a)-conviviente".

Cuadro 32. Distribución de personas que responden según estado conyugal colapsado y sexo.

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Casado(a) - Conviviente	403	114	517
	Otros	207	107	314
Total No		**610**	**221**	**831**
Sí	Casado(a) - Conviviente	3009	1318	4327
	Otros	1101	1060	2161
Total Sí		**4110**	**2378**	**6488**
Total general		**4720**	**2599**	**7319**

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

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Casado(a) - Conviviente	5,5%	1,6%	7,1%
	Otros	2,8%	1,5%	4,3%
Total No		**8,3%**	**3,0%**	**11,4%**
Sí	Casado(a) - Conviviente	41,1%	18,0%	59,1%
	Otros	15,0%	14,5%	29,5%
Total Sí		**56,2%**	**32,5%**	**88,6%**
Total general		**64,5%**	**35,5%**	**100,0%**

Fuente: Elaboración propia

Dentro de las personas que no responden el 5,5% de los hombres y el 1,6% de mujeres, pertenecen a estado conyugal Casado-conviviente representado el 7,1% de los casos. En cambio dentro de las personas que responden, el 41,1% hombres y 18% 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 34. 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 general**
No	Casado(a) - Conviviente	77,9%	22,1%	100,0%
	Otros	65,9%	34,1%	100,0%
Total No		**73,4%**	**26,6%**	**100,0%**
Sí	Casado(a) - Conviviente	69,5%	30,5%	100,0%
	Otros	50,9%	49,1%	100,0%
Total Sí		**63,3%**	**36,7%**	**100,0%**
Total general		**64,5%**	**35,5%**	**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 77,9% de los casos pertenecen al sexo Hombre y el 22,1% Mujer. Para los que pertenecen a la categoría Otros, el 65,9% son hombres y el 34,1% 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 69,5% de los casos pertenecen al sexo Hombre y el 30,5% Mujer. Para los que pertenecen a la categoría Otros, el 50,9% son hombres y el 49,1% son mujeres.

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

Cuadro 35. 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 general**
No	Casado(a) - Conviviente	66,1%	51,6%	62,2%
	Otros	33,9%	48,4%	37,8%
Total No		**100,0%**	**100,0%**	**100,0%**
Sí	Casado(a) - Conviviente	73,2%	55,4%	66,7%
	Otros	26,8%	44,6%	33,3%
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 66,1%, el 33,9% pertenece a Otros. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 51,6% pertenece a Casado - Conviviente, y 48,4% 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 73,2% y el 26,8% pertenece a Otros. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 55,4% pertenece a Casado - Conviviente, y 44,6% 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 36. Distribución de personas que responden según cantidad de visitas Colapsado y sexo.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	231	91	322
	4-6	296	95	391
	7 y más	83	35	118
Total No		**610**	**221**	**831**
Sí	1-3	3596	2165	5761
	4-6	454	188	642
	7 y más	60	25	85
Total Sí		**4110**	**2378**	**6488**
Total general		**4720**	**2599**	**7319**

Fuente: Elaboración propia

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

Cuadro 37. 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%	1,2%	4,4%
	4-6	4,0%	1,3%	5,3%
	7 y más	1,1%	0,5%	1,6%
Total No		**8,3%**	**3,0%**	**11,4%**
Sí	1-3	49,1%	29,6%	78,7%
	4-6	6,2%	2,6%	8,8%
	7 y más	0,8%	0,3%	1,2%
Total Sí		56,2%	32,5%	88,6%
Total general		64,5%	35,5%	100,0%

Fuente: Elaboración propia

Se puede ver que el 49,1% 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 38. 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**	71,7%	28,3%	100,0%
	4-6	75,7%	24,3%	100,0%
	7 y más	70,3%	29,7%	100,0%
Total No		73,4%	26,6%	100,0%
Sí	**1-3**	62,4%	37,6%	100,0%
	4-6	70,7%	29,3%	100,0%
	7 y más	70,6%	29,4%	100,0%
Total Sí		63,3%	36,7%	100,0%
Total general		64,5%	35,5%	100,0%

Fuente: Elaboración propia

De igual forma se puede obtener la distribución marginal del sexo

Cuadro 39. 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	37,9%	41,2%	38,7%
	4-6	48,5%	43,0%	47,1%
	7 y más	13,6%	15,8%	14,2%
Total No		100,0%	100,0%	100,0%
Sí	1-3	87,5%	91,0%	88,8%
	4-6	11,0%	7,9%	9,9%
	7 y más	1,5%	1,1%	1,3%
Total Sí		100,0%	100,0%	100,0%

Fuente: Elaboración propia

Del total de personas que responde, el 87,9% fue visitada entre 1-3 veces, mientras que sólo el 1,5% 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, 91%.

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”. (87,5% hombres versus 91% 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 Regresión logística

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 5, 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 40. Parámetros estimados del modelo de regresión logística seleccionado para modelar la respuesta de una persona que pertenece a una unidad elegible.

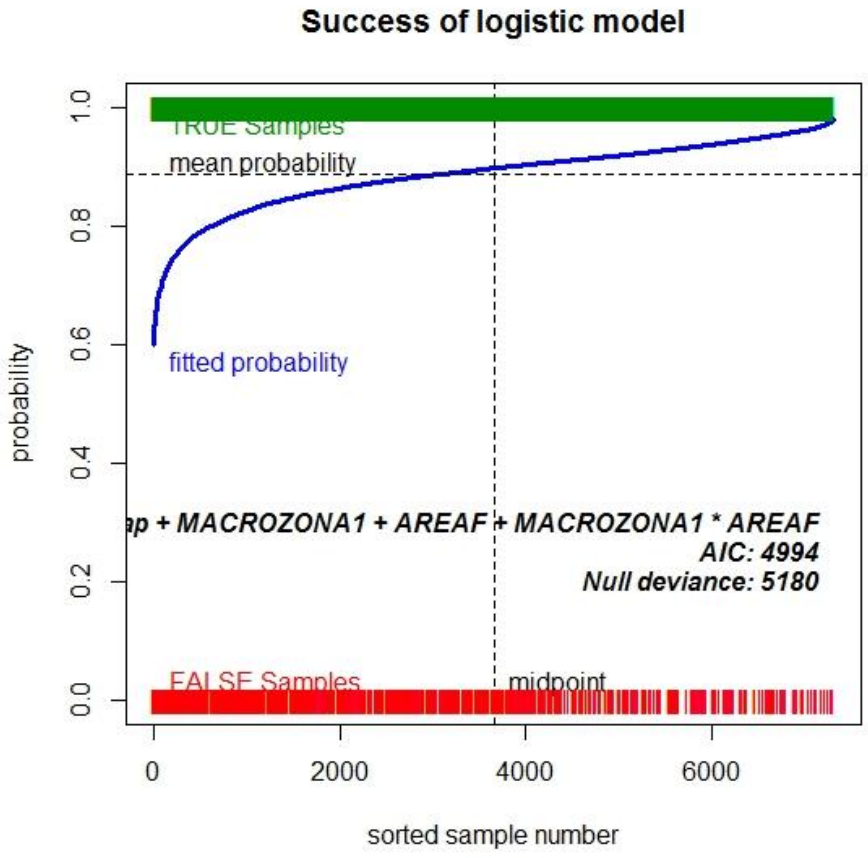
Variables	Estimación	Error Estándar	Valor z	Pr(>|z|)	Odd Ratios	Intervalo de Confianza 95%	
						Lim Inf	Lim Sup
Intercepto	2,177	0,265	8,217	0,000	8,821	5,302	15
EDAD	0,015	0,003	4,944	0,000	1,015	1,009	1,02
Macrozona Metropolitana	-1,099	0,36	-3,05	0,002	0,333	0,167	0,695
Macrozona Norte	-0,764	0,239	-3,204	0,001	0,466	0,289	0,74
Macrozona Sur	0,052	0,249	0,21	0,834	1,054	0,644	1,713
CIUO_88_1_DIGITOF4	-0,341	0,465	-0,732	0,464	0,711	0,306	1,95
CIUO_88_1_DIGITOF1	-0,275	0,207	-1,325	0,185	0,76	0,507	1,144
CIUO_88_1_DIGITOF7	0,036	0,158	0,226	0,821	1,036	0,758	1,407
CIUO_88_1_DIGITOF8	-0,331	0,176	-1,88	0,06	0,718	0,507	1,013
CIUO_88_1_DIGITOF2	-0,557	0,201	-2,77	0,006	0,573	0,385	0,849
CIUO_88_1_DIGITOF3	-0,303	0,185	-1,636	0,102	0,738	0,513	1,061
CIUO_88_1_DIGITOF5	-0,03	0,172	-0,174	0,862	0,97	0,691	1,356
CIUO_88_1_DIGITOF9	-0,177	0,18	-0,984	0,325	0,838	0,588	1,192
Variables categóricas en el modelo							
Area							
Urbano	-0,763	0,201	-3,8	0,000	0,466	0,31	0,681
Rural	-	-	-	-	-	-	-
Nivel Educacional							
Básica	-0,227	0,101	-2,26	0,024	0,797	0,655	0,972
Media	-0,373	0,347	-1,073	0,283	0,689	0,36	1,418
Superior	-	-	-	-	-	-	-
Proveedor Principal							
Sí	-	-	-	-	-	-	-
No	-0,271	0,086	-3,163	0,002	0,763	0,645	0,902
Sexo							
Mujer	0,592	0,094	6,292	0,000	1,808	1,505	2,177
Hombre							
Estado Conyugal Colapsado							
Casado(a) - Conviviente	-	-	-	-	-	-	-
Otros	-0,159	0,081	-1,976	0,048	0,853	0,729	0,999

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



Fuente: Elaboración propia

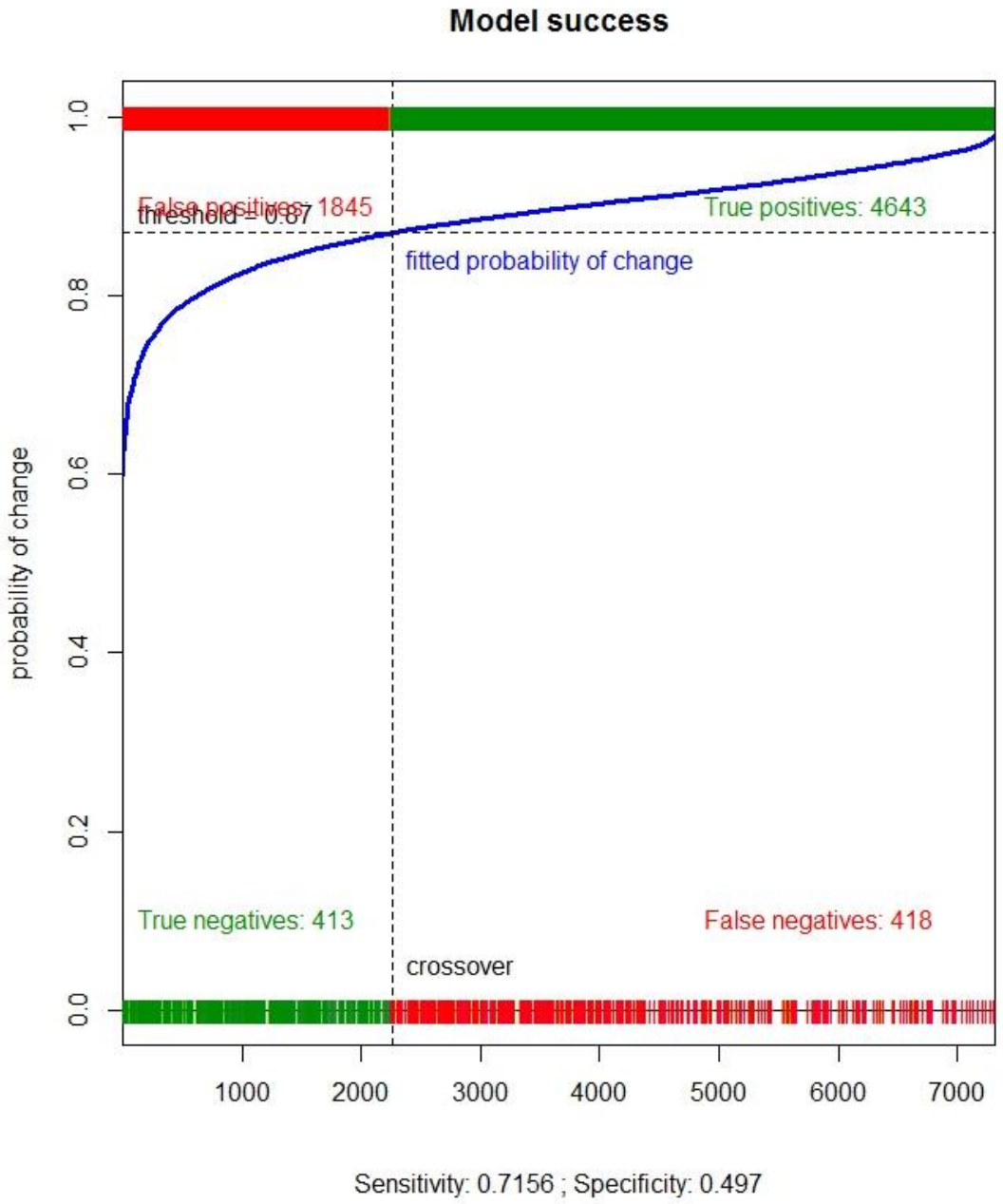
Cuadro 41. Estadísticas Descriptivas para la probabilidad estimada del modelo propuesto

Variable	n	Media	Desv. Estándar	Mediana	Mínimo	Máximo	Rango	Error Muestral
Prob. Estimada	7319	0,886	0,057	0,9	0,599	0,98	0,38	0,0067

Fuente: Elaboración propia

Se puede observar que el modelo seleccionado ajusta una probabilidad de responder de una persona entre 59.92% y 98,03%, con una probabilidad media de 88,65%.

Gráfico 9. Clasificación de las personas que responden versus las que no responden con un umbral de 0,87



Fuente: Elaboración propia

Cuadro 42. Matriz de Confusión con un umbral = 0,87

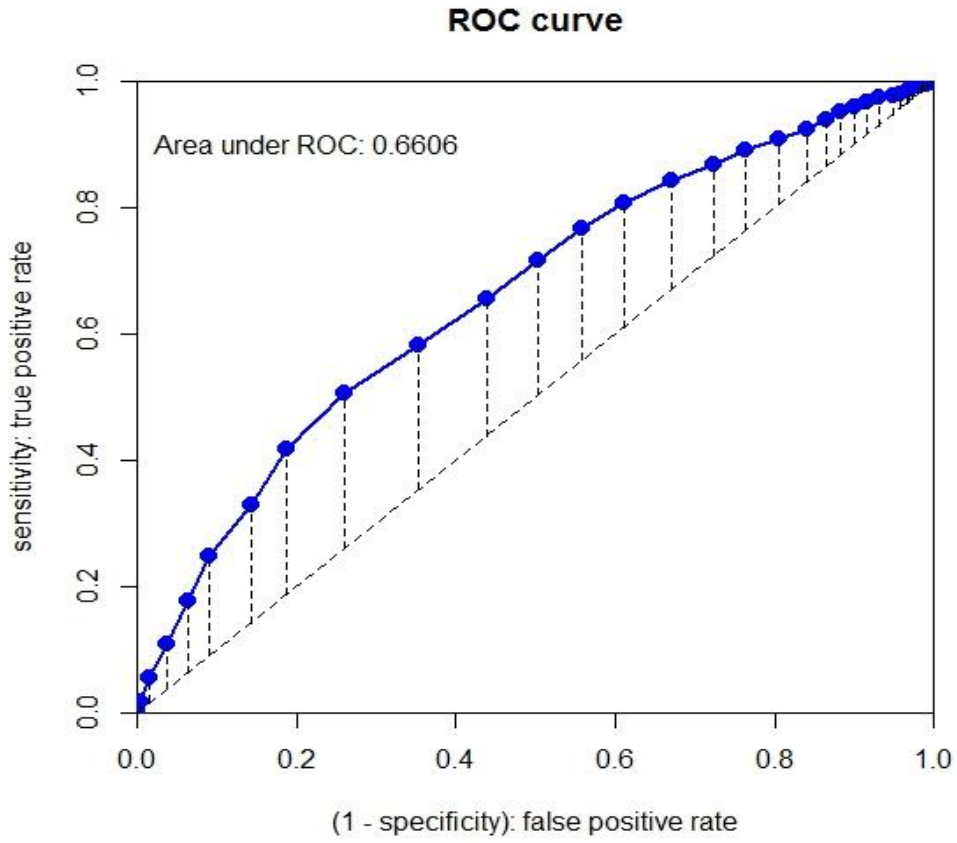
		Criterio de Verdad		
		Responden	**No responden**	**Total**
Prueba Diagnóstica	**Positivos**	4643	1845	6488
	Negativos	418	413	831
	Total	5061	2258	7319

Fuente: Elaboración propia

Finalmente, la sensibilidad y especificidad calculada es:

$$Sensibilidad = \frac{4643}{6488} = 0{,}716 \qquad Especificidad = \frac{413}{831} = 0{,}497$$

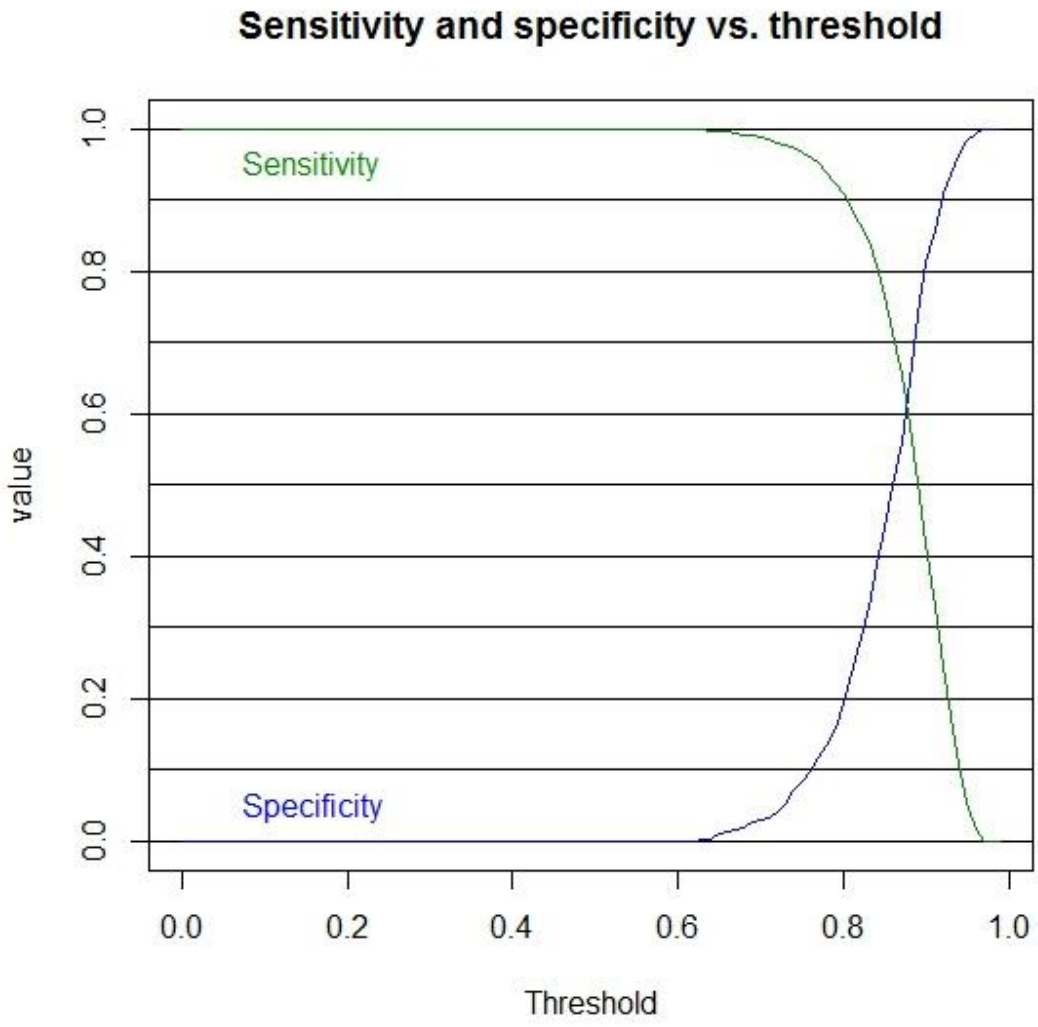
Gráfico 10. Curva ROC y Área bajo la curva para el modelo seleccionado.



Fuente: Elaboración propia

En base al modelo estimado se puede observar que el área bajo la curva (AUC) es de 0,66, lo cual está dentro **de una categoría de "Test regular".** O bien, se puede decir que el modelo tiene una capacidad de predicción del 66% de los casos.

Gráfico 11. 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 0,87 para clasificar a las personas.

4. Anexo N°4. Estimación de varianzas

4.1.1. 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 =d17_rev3_1.
IF(d17_rev3_1=2 | d17_rev3_1=3 | d17_rev3_1=5) Rama_reducida=20.
IF (d17_rev3_1=8 | d17_rev3_1=10 | d17_rev3_1=12 | d17_rev3_1=13 | d17_rev3_1=14 |
d17_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=d17_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.
```