ESTUDIO Y AJUSTE DE LA NO RESPUESTA EN LAS ENCUESTAS A HOGARES

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

Presentación

En este documento se presenta el trabajo desarrollado por Eustat durante los años 2006 y 2007 en relación al estudio y tratamiento de la no respuesta.

El trabajo realizado se enmarca en una de las operaciones del Plan Vasco de Estadística 2005-2008, relativa a I+D+i de métodos estadísticos, y cuya finalidad se dirige a investigar y aplicar nuevas metodologías estadístico-matemáticas en las operaciones estadísticas. El espíritu que anima estos estudios se encuadra en una gestión de la excelencia, que está definido por un proceso de mejora continua.

La inclusión del estudio de la no respuesta en las operaciones del Plan es un reflejo de la preocupación que EUSTAT comparte con otras oficinas estadísticas del ámbito internacional por este tema. La falta de respuesta es un fenómeno que se está produciendo en diferente medida en muchas operaciones estadísticas, que reduce siempre la precisión de las estimaciones de los resultados y, además, introduce sesgos.

El estudio y seguimiento de la falta de respuesta en las encuestas a hogares, se ha venido realizando desde siempre en EUSTAT. Sin embargo, se ha planteado la necesidad de establecer unos estándares o criterios comunes para su medición, basados en la metodología europea publicada para tal fin por el Institute for Social & Economic Research, de la Universidad de Essex (Reino Unido).

Además, es un objetivo también de esta investigación estudiar cómo mejorar las estimaciones en las encuestas hogares, en situaciones de falta de respuesta, a través de información auxiliar, información normalmente externa a la encuesta.

Espero que esta difusión sea de utilidad para todos los interesados en este ámbito de la estadística.

Vitoria-Gasteiz, septiembre de 2008

Josu Iradi Arrieta

Director General.

RESUMEN

El presente documento[1] está dividido en los siguientes capítulos:

En el primer capítulo se realiza una introducción y se mencionan los objetivos que han marcado la elaboración de este cuaderno técnico.

En el segundo capítulo se expone la estandarización del cálculo de la tasa de respuesta y sus resultados en la Encuesta de Población en Relación con la Actividad (en adelante PRA). Dicha estandarización se ha tratado con más detalle en el Cuaderno de Trabajo dedicado a este tema, publicado por Eustat.

El objetivo del tercer capítulo es exponer el análisis de la información auxiliar, que es la información disponible, tanto para respondientes como para no respondientes, y que tenga relación con la falta de respuesta. Este análisis se centrará en seleccionar las variables auxiliares más adecuadas para reducir el efecto de la falta de respuesta en la encuesta. Para ello se ha aplicado la metodología propuesta por Statistics Sweden.

En el cuarto capítulo, se abordará la calibración en base a las variables auxiliares determinadas utilizando el estimador de regresión generalizado. Este estimador es explicado ampliamente en este capítulo, así como el programa CLAN[2]

En el capítulo quinto se presentan las conclusiones a las que se ha llegado y el último punto hace referencia a la bibliografía utilizada.

[1] EUSTAT quiere agradecer el excelente trabajo de investigación desarrollado por Susana Sanz Abrego, en el marco de la Beca de Investigación de Metodología Estadístico-Matemática que promueve EUSTAT, concretamente en el campo de la No Respuesta.

[2] CLAN es una macro desarrollada por Statistics Sweden que funciona en entorno SAS

Indice

Indice de tablas

Indice de gráficos

Capítulo 1

Introducción

Introducción y objetivos

En estos últimos años, las oficinas estadísticas han mostrando cierta preocupación por la falta de respuesta producida en la respuesta a las encuestas oficiales.

La importancia de la falta de respuesta estriba principalmente en dos motivos. El primero, la introducción de un sesgo o error de medida, debido a que los no respondientes pueden diferir de los respondientes en características importantes. Y, el segundo, en que la no respuesta reduce la precisión de las estimaciones, ya que habrá menos casos disponibles para el análisis.

Ante la posibilidad de que se produzca falta de respuesta, se puede actuar de dos modos, no excluyentes. En primer lugar, las oficinas pueden actuar antes, y tratar de reducirla durante el trabajo de campo. Para esto, se revisan todo lo relativo a la recogida de la información, desde los cuestionarios hasta los incentivos, los modos de contactar con la población, la recogida telefónica, por Internet, etc.

En segundo lugar, se puede actuar después, cuando el fenómeno ya se ha producido y entonces las oficinas estadísticas se plantean cómo puede ser la estimación en situaciones de no respuesta. En líneas generales, hay dos procedimientos para tratar la falta de información, una vez recogida. En primer lugar, existe el procedimiento de la imputación, que consiste en sustituir los datos en blanco por datos aceptables conocidos (Puerta 2002). Este procedimiento suele considerarse en casos de falta de respuesta parcial, es decir, cuando el individuo ha dado alguna información.

El segundo procedimiento sería el de la calibración o ajuste con información auxiliar y se suele utilizar en casos de no respuesta total. Se lleva a cabo en el momento de la estimación, cuando se calculan los pesos de las respuestas de los individuos en las encuestas. Las oficinas estadísticas tratan de estudiar entonces, cuál es la información auxiliar disponible para respondientes y no respondientes, que permita mejorar la estimación que se obtendría aplicando los métodos de estimación adecuados cuando no se produce falta de respuesta, o cuando esta es ignorable.

Eustat plantea el estudio de la no respuesta en dos sentidos principalmente. En primer lugar, se trata de medir la falta de respuesta atendiendo a criterios estandarizados, que permita comparar tasas entre encuestas de distintas oficinas estadísticas y también de modo temporal.

En cuanto a este aspecto, en este Cuaderno Técnico se muestran la definición de la tasa de respuesta y los resultados que se obtienen para la Encuesta de Población en Relación con la Actividad (PRA), desde el trimestre 4º de 2004 hasta el 3º del 2006.

En realidad, éste es un resumen del Cuaderno de Trabajo Estandarización y sistematización del cálculo de las tasas de respuesta (Eustat 2007) donde se explica con más detalle la metodología que se aplica (Lynn et al. 2001), así como el conjunto de las tasas de respuesta que se aplican (tasa de cooperación, contacto, rechazo y elegibilidad).

En segundo lugar, se trata de estudiar el método de la calibración o la reponderación para la falta de respuesta, que permita mejorar las estimaciones de las encuestas, refiriéndonos siempre a las dirigidas a hogares. Este estudio tiene dos aspectos a su vez muy relacionados como:

- lo relativo a los estimadores, en el sentido de que algunos puedan tener propiedades más adecuadas para el objeto de estudio

- lo relativo a la información auxiliar, en el sentido de que hay criterios para seleccionar información más eficiente para combinar con los estimadores anteriores

Este Cuaderno Técnico se centra en el segundo objetivo y dedica un capítulo al análisis de la información auxiliar disponible. Se parte de la metodología que propone Statistics Sweden, con una serie de condiciones que debe satisfacer la información que se dispone para respondientes y no respondientes. Las pruebas realizadas, en base a esta metodología, se aplican de nuevo a la PRA.

En un capítulo posterior, se procede a la calibración de dicha encuesta utilizando esta información auxiliar. Para la calibración se utiliza un software determinado, la macro CLAN, desarrollada en entorno SAS por Statistics Sweden. Previamente, se expone también el estimador que se va a utilizar, el lineal generalizado (GREG), para situaciones de falta de respuesta. Y por último, se dedica una apartado a la exposición de los resultados con la nueva calibración, que no son muy diferentes de los ya publicados con el método de calibración general, lo cuál se valorará en las conclusiones.

Capítulo 2

Estandarización y sistematización de las tasas de respuesta

El estudio y seguimiento de la falta de respuesta en las encuestas a hogares se ha venido realizando desde siempre en EUSTAT. Sin embargo, se ha planteado la necesidad de establecer unos estándares o criterios comunes para su medición y se ha optado por seguir la metodología europea publicada para tal fin por el Institute for Social & Economic Research, de la Universidad de Essex (Reino Unido) (Lynn et al. 2001).

Tomando como referencia dicha metodología, se han codificado las incidencias que se obtienen en EUSTAT para encuestas a hogares. En concreto, el estudio de la no respuesta se ha realizado para la Encuesta de Población en Relación con la Actividad.

2.1 Notación ISER utilizada y definición de la tasa de respuesta

Una vez codificadas las incidencias con la clasificación ISER (Lynn et al. 2001), como ya se muestra en detalle en el Cuaderno de Trabajo (Eustat, 2007), se aplican las siguientes definiciones estándar que serán utilizadas para el cálculo de las tasas. El número entre paréntesis se refiere a la clasificación ISER.

I = Entrevista completa (1)

P = Entrevista parcial (2)

NC = No contacto (3)

R = Negativa (4)

O = Otro tipo de no respuesta (5)

UC = Elegibilidad dudosa, contacto (641,651,661 y parte de 67).

UN = Elegibilidad dudosa, no contacto (61,62,63,642,652,662, 68 y resto de 67).

NE = No elegible (7)

Ec = Proporción de los casos contactados de elegibilidad dudosa que son elegibles

En = Proporción de los casos no contactados de elegibilidad dudosa que son elegibles

La definición de la tasa de respuesta es la siguiente:

Tasa de respuesta indica la proporción de entrevistas realizadas de todos los casos elegibles. Es decir, relaciona el número de encuestas realizadas, ya sean completas o parciales, sobre el total de encuestas realizadas, junto con las negativas, no contactadas, las clasificadas como otro tipo de no respuesta y, en una proporción determinada, aquellas posibles elegibles, sean contactadas o no.

La fórmula es la siguiente:

$$RR_O = \frac{I + P}{(I + P) + (R + NC + O) + e_C UC + e_N UN)}$$

En el mencionado Cuaderno de Trabajo se expone el resto de las tasas (cooperación, contacto, rechazos o negativas y elegibilidad), con los resultados obtenidos.

2.2 Resultados obtenidos

Las anteriores definiciones se han aplicado sobre la muestra de la Encuesta de Población en Relación con la Actividad (PRA). La finalidad de la encuesta es, en líneas generales, conocer la relación con la actividad de la población de 16 y más años.

En cuanto a su diseño muestral, la PRA es una encuesta longitudinal que se realiza cada tres meses. La muestra está compuesta por 5.088 viviendas en cada uno de los distintos trimestres de encuestación. Cada vivienda permanece en el panel durante 8 turnos de rotación, lo que es equivalente a 2 años, y una vez transcurrido este tiempo la vivienda abandona el panel. De la misma forma, cada trimestre se renueva la muestra en un 1/8.

Entre los trimestres 4º de 2004 y 3º de 2006, las tasas de respuesta por trimestre que se obtienen son superiores al 80%. En el trimestre 4º de 2004, cuando se renueva el panel completamente, la tasa de respuesta es del 83,9%. Esta tasa desciende ligeramente durante los 3 trimestres posteriores y a partir de 2006 comienza a aumentar en todos los trimestres hasta alcanzar un 86,9% en el tercer trimestre de 2006.

Sin hacer una comparación sistemática de estos resultados, parece que las tasas de respuesta podrían considerarse aceptables.

T.1 Tasas de respuesta por trimestre

Trimestre/Año	Tasa de respuesta
4/2004	83,9
1/2005	82,1
2/2005	82,6
3/2005	82,8
4/2005	83,8
1/2006	85,1
2/2006	86,0
3/2006	86,9
MEDIA	84,2

Fuente: Eustat. PRA

Capítulo
3

Análisis de la información auxiliar para la reponderación

Como ya se decía en el capítulo de introducción, hay dos modos principales de tratar la falta respuesta en las encuestas. Un modo es el de la imputación, por el cual se asignan por métodos estadísticos valores a la información que falta. Y el otro modo es el de la reponderación. Este consiste en calcular nuevos pesos, teniendo en cuenta la información auxiliar, al conjunto de la muestra que ha respondido a la encuesta. El éxito de esta reponderación consiste en disponer de una información auxiliar eficiente y que tenga en cuenta de alguna forma el mecanismo de la falta de respuesta.

El objetivo en este capítulo es precisamente determinar la información auxiliar que va a intervenir en la reponderación. Se trata de utilizarla en el cálculo de los pesos de los individuos que han contestado, de modo que se reduzca el sesgo de la no respuesta en la encuesta. Para ello se ha aplicado la metodología propuesta por Statistics Sweden, en su publicación "Estimation in presence of Nonresponse and frame imperfections" (Lundström y Särndal 2002).

En el manual mencionado las condiciones a cumplir para llegar a determinar las variables auxiliares a considerar son los siguientes:

1. Tienen que explicar la variación de las probabilidades de respuesta

2. Tienen que explicar la variación de las principales variables de estudio

3. Tienen que identificar los dominios más importantes

3.1 Información auxiliar y tasas de respuesta

En primer lugar, se ha analizado una serie de variables auxiliares para ver si cumplen el primer principio. Las variables auxiliares disponibles para el estudio han sido: sexo, edad, territorio, tamaño del hogar, número de personas activas en el hogar, tamaño de municipio y edad de la persona mayor del hogar. Posteriormente se han calculado las tasas de respuesta por categoría para estas variables, y también la combinación de las variables auxiliares tamaño del hogar y edad de la persona mayor del hogar.

A continuación se muestran las tasas de respuesta de las modalidades de las variables auxiliares que en principio presentan diferencias :

3.1.1 Tasa de respuesta de la combinación de las variables "tamaño del hogar" y "edad de la persona mayor del hogar"

Con el objetivo de ver más claramente las diferencias entre las categorías de las variables combinado tamaño del hogar y edad de la persona más mayor del hogar. Realizado el estudio correspondiente, se decidió dividir en cuatro categorías.

Las categorías seleccionadas son las siguientes:

1. Hogares con una persona en el hogar donde la persona mayor del hogar es menor de 45 años

2. Hogares con una persona en el hogar donde la persona mayor del hogar es mayor de 45 años

3. Hogares con 2 ó más personas en el hogar donde la persona mayor del hogar es menor de 45 años.

4. Hogares con 2 ó más personas en el hogar donde la persona mayor del hogar es mayor de 45 años.

A continuación se muestra la tabla correspondiente.

T.2.Tasa de respuesta por tamaño del hogar y edad de la persona mayor del hogar

		20044	20051	20052	20053	20054	20061	20062	20063
Total	0 a 44 años	85,3	84,2	83,5	84,0	85,2	87,3	87,6	87,6
	45 y más	83,6	81,5	82,6	82,9	84,0	85,2	86,5	88,0
	Total	84,0	82,1	82,8	83,2	84,3	85,7	86,8	87,9
1 Persona	0 a 44 años	65,9	64,5	65,9	67,7	68,3	71,1	71,0	73,4
	45 y más	79,7	78,9	79,8	80,8	81,9	83,3	83,7	84,9
	Total	76,4	75,5	76,5	77,9	79,0	80,9	81,2	82,4
2 ó mas	0 a 44 años	86,7	85,8	85,0	85,3	86,5	88,4	88,7	88,7
	45 y más	83,9	81,7	82,8	83,1	84,2	85,4	86,8	88,2
	Total	84,6	82,6	83,3	83,6	84,7	86,0	87,2	88,3

Fuente: Eustat.PRA

Analizando las tasas de respuesta en las nuevas categorías se observa lo siguiente:

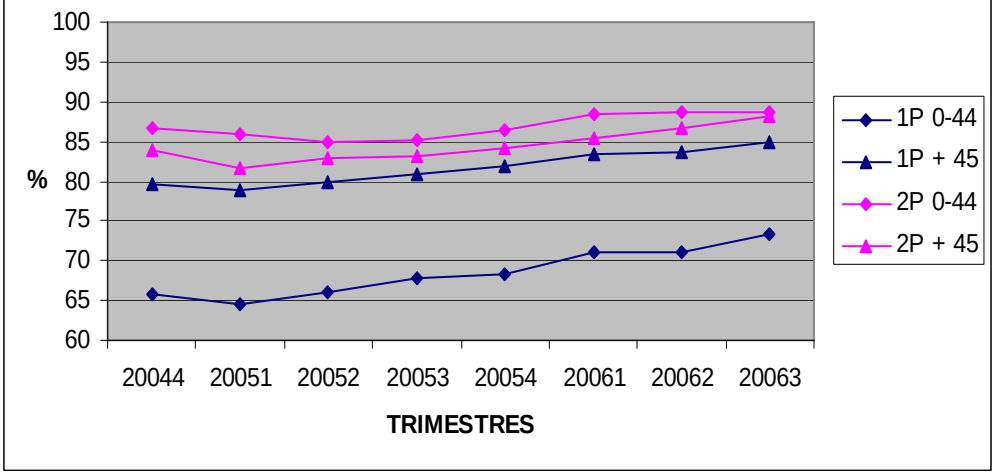
1. Existe una gran diferencia entre la tasa de respuesta de los hogares con una persona de 0-44 años y el resto. Esta diferencia es de aproximadamente unos 15 puntos con respecto al resto de categorías, por lo que podría afirmarse que los hogares con una sola persona menor de 45 años tienen una tasa de respuesta mucho más baja

que la del resto. Esta tasa oscila entre un 66% y un 73%, correspondientes al primer y último trimestre de encuestación respectivamente.

2. Por otro lado, estarían las 3 restantes categorías con una tasa de respuesta mucho más elevada. De entre éstas, la más baja corresponde a los hogares de una persona mayor de 45 años, seguida de los hogares con 2 ó más personas y edad de la persona mayor de 0 a 44 años, y por último, hogares de 2 ó más personas mayores de 45 años.

A continuación se muestra el gráfico correspondiente de la tasa de respuesta por tamaño del hogar y edad de la persona más mayor.

G.1. TASA DE RESPUESTA POR TAMAÑO DEL HOGAR Y EDAD DE LA PERSONA MAYOR



3.1.2 Tasa de respuesta de la variable "tamaño de municipio"

La siguiente tabla muestra la evolución de las tasas de respuesta por tamaño de municipio para cada uno de los trimestres.

Observando las 3 categorías, la tabla muestra que hay diferencias entre las 3 modalidades. Las tasas de respuesta más altas corresponden a los municipios con menos de 5000 habitantes, y por el contrario, las tasas menores, corresponden a los municipios con mayor número de habitantes.

Por otro lado, se observa que las tasas de los municipios con menor tamaño se mantienen prácticamente estables, mientras que la tasa de respuesta de los municipios con más población aumenta 6 puntos.

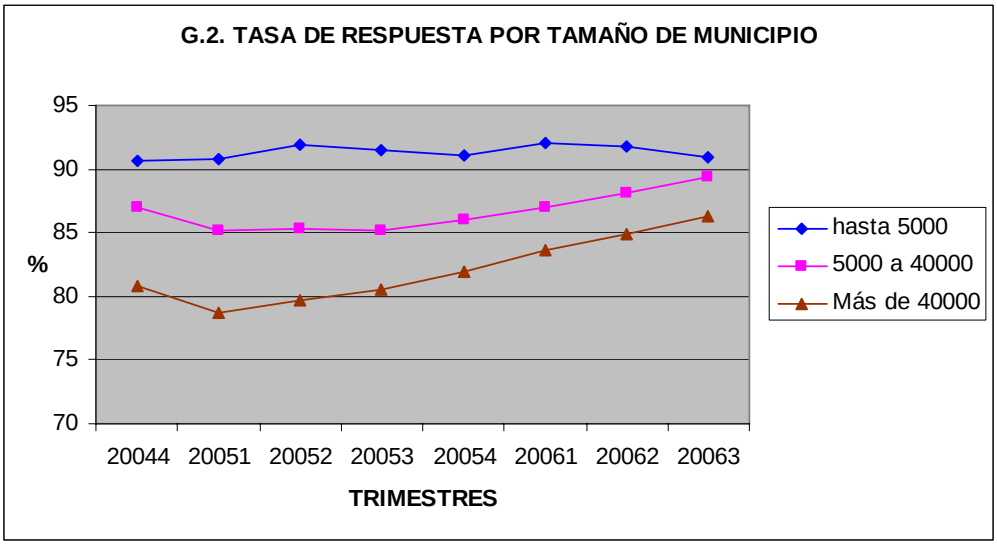
Esta variable será candidata a ser seleccionada, debido a que existen diferencias entre las distintas modalidades , aunque las diferencias se han reducido en los dos últimos trimestres.

T.3. Tasa de respuesta por tamaño de municipio

Trimestre	Total	Hasta 5000	5000 a 40000	Más de 40000
20044	84,0	90,7	87,1	80,9
20051	82,1	90,7	85,2	78,7
20052	82,8	91,9	85,3	79,7
20053	83,2	91,5	85,1	80,5
20054	84,3	91,1	86,0	81,9
20061	85,7	92,1	87,1	83,6
20062	86,8	91,8	88,1	84,9
20063	87,9	91,0	89,3	86,3

Fuente: Eustat.PRA

En el siguiente gráfico se muestra la evolución de las tasas de respuesta por tamaño de municipio en cada uno de los trimestres.



3.2 Información auxiliar y variables de estudio

Una vez seleccionadas las variables que cumplen el primer principio, la siguiente fase dentro de la selección de la información auxiliar, es la de ver su efecto en las variables en estudio. Siguiendo el manual ya citado, habría que comprobar si las variables seleccionadas con el criterio anterior además cumplen la condición de que tienen alguna relación con las variables en estudio.

La variable de estudio seleccionada en este caso es la variable “relación con la actividad” codificada en 3 categorías: Ocupado, Parado e Inactivo..

A continuación se muestran los resultados obtenidos para las variables seleccionadas con el primer criterio.

3.2.1 Variable auxiliar “tamaño de municipio” por actividad

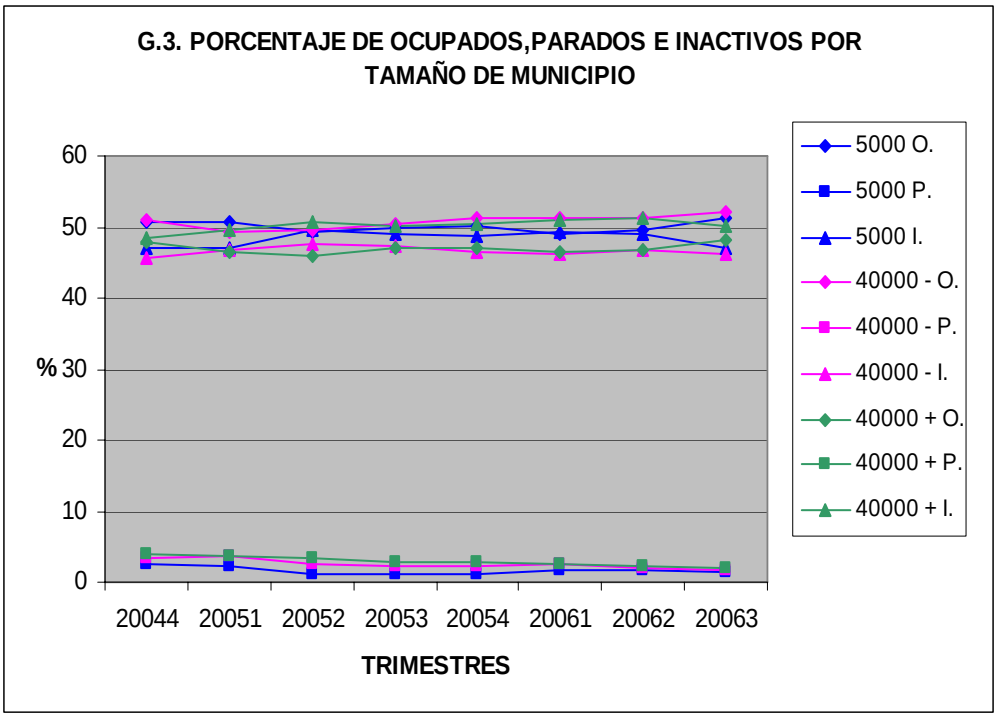
Analizando la variable “tamaño de municipio” se observa que la distribución porcentual de ocupados, parados e inactivos en los municipios de menos de 5.000 habitantes es similar a la de los municipios de 5.000-40.000 habitantes. En los municipios de más de 40000 habitantes estas diferencias son algo mayores, pero no resultan significativas.

T.4. Porcentaje de ocupados, parados e inactivos de más de 16 años por tamaño de municipio

		20044	20051	20052	20053	20054	20061	20062	20063
Total	Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Ocupado	48,9	47,8	47,4	48,3	48,6	48,2	48,4	49,6
	Parado	3,6	3,6	2,9	2,5	2,4	2,5	2,0	1,8
	Inactivo	47,5	48,6	49,7	49,2	49,0	49,3	49,6	48,6
Hasta 5000	Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Ocupado	50,8	50,7	49,4	49,8	50,0	49,1	49,5	51,4
	Parado	2,5	2,2	1,0	1,1	1,3	1,7	1,6	1,5
	Inactivo	47,1	47,0	49,6	49,1	48,7	49,2	49,0	47,1
5000-40000	Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Ocupado	51,0	49,4	49,6	50,4	51,4	51,3	51,3	52,1
	Parado	3,4	3,7	2,6	2,2	2,2	2,5	2,0	1,8
	Inactivo	45,6	46,9	47,7	47,4	46,4	46,3	46,7	46,1
más 40000	Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Ocupado	47,8	46,6	46,0	47,1	47,0	46,5	46,7	48,0
	Parado	3,8	3,7	3,3	2,7	2,7	2,6	2,1	1,9
	Inactivo	48,4	49,7	50,7	50,1	50,3	50,9	51,2	50,1

Fuente: Eustat.PRA

A continuación se muestra el gráfico correspondiente de los porcentajes de ocupados, parados e inactivos por tamaño de municipio.



3.2.2 Combinación de "tamaño del hogar" y "edad de la persona mayor del hogar" por actividad

Ahora se analizarán los resultados de la combinación de las variables auxiliares tamaño del hogar y edad de la persona mayor del hogar por actividad.

A continuación se muestran los resultados por la variable actividad. Las tablas y los gráficos siguientes están desagregados por cada una de las 3 categorías de la variable actividad: ocupados, inactivos y parados.

3.2.2.1 Porcentaje de ocupados de 16 y más años por tamaño del hogar y persona mayor del hogar

En el siguiente gráfico y tabla se observa lo siguiente:

1. Con respecto a los hogares con una persona, el porcentaje de ocupados de las personas de 0-44 años, es muchísimo mayor que el de las personas mayores de 45 años. La proporción de los primeros fluctúa entre un 83% y un 89%, mientras que el porcentaje de los segundos oscila entre un 15% y un 17%.

2. En los hogares con 2 personas ó más, esta diferencia es también significativa, pero no tanto como en el caso anterior. Por un lado, el porcentaje de ocupados de los menores de 44 años, varía entre un 77% y un 81%, y el de los mayores de 45 es de un 44%.

3. Examinando ambos resultados, se observa que en la categoría de 0-44 años existen diferencias de aproximadamente 6 ó 8 puntos entre los hogares con una sola persona en el hogar y los hogares con 2 ó más personas en el hogar. Sin embargo, en la categoría de más de 45 años estas diferencias son muy significativas, siendo mayor el porcentaje de ocupados de las viviendas con 2 personas ó más que los de 1 sola persona.

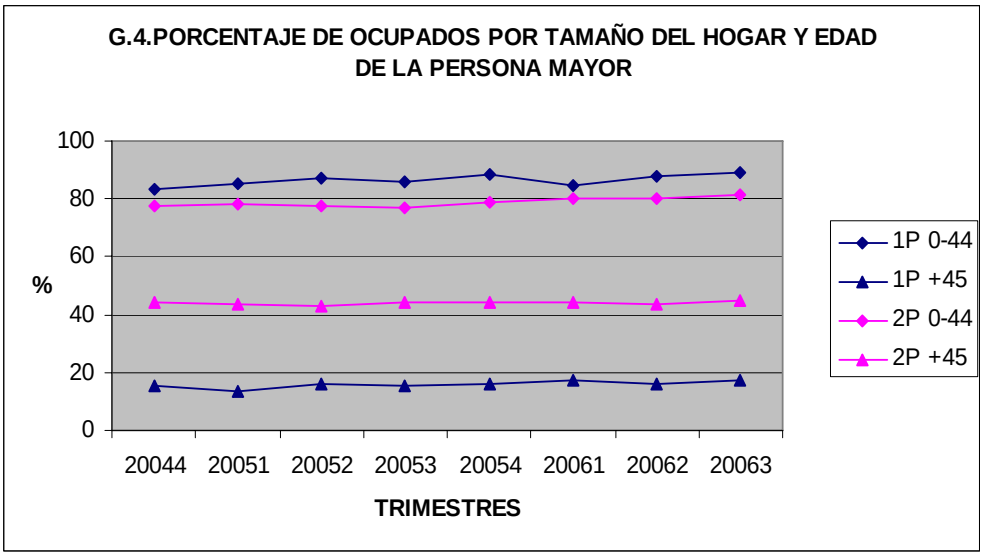
4. A modo de resumen cabe destacar que dentro de los mayores de 45 años existen diferencias muy significativas entre el porcentaje de ocupados de las viviendas con 2 personas ó más y el porcentaje de las viviendas con una sola persona en el hogar. En la categoría de los menores de 45 años esta diferencia no es tan grande como en el caso anterior, pero si que es significativa.

T.5. Porcentaje de ocupados de 16 y más años por tamaño del hogar y persona mayor del hogar

		20044	20051	20052	20053	20054	20061	20062	20063
Total	0 a 44 años	77,9	79,1	78,7	77,8	79,9	80,4	80,9	81,9
	45 y más	42,5	41,5	41,1	42,3	42,0	41,8	41,5	42,5
	Total	48,9	47,8	47,4	48,3	48,6	48,2	48,4	49,6
1 Persona	0 a 44 años	83,6	85,1	87,4	85,7	88,4	84,8	88,0	89,2
	45 y más	15,2	13,3	15,8	15,4	16,3	17,2	16,3	17,0
	Total	28,4	26,5	29,8	28,4	29,8	28,9	28,9	31,1
2 ó mas	0 a 44 años	77,4	78,5	77,8	77,1	79,0	80,0	80,2	81,1
	45 y más	44,5	43,7	43,1	44,5	44,3	44,0	43,8	44,9
	Total	50,5	49,5	48,8	50,1	50,3	49,9	50,1	51,3

Fuente: Eustat.PRA

En el siguiente gráfico se muestran las diferencias entre las 4 categorías:



3.2.2.2 Porcentaje de parados de 16 y más años por tamaño del hogar y persona mayor del hogar

La siguiente tabla y gráfico correspondiente al porcentaje de parados muestra lo siguiente:

1. Por un lado, se observa que los porcentaje de parados dentro los hogares donde vive una sola persona, es diferente para los menores de 44 años que para los mayores de 44 años. Los primeros tienen una tasa que fluctúa entre un 3% y un 7%, mientras que la proporción de los segundos oscila entre un 0,25% y un 1,3%

2. Por otro lado, examinando los hogares de 2 ó más personas, se concluye que las diferencias entre estas dos categorías apenas son perceptibles. Ambas categorías varían entre un 2% y un 4%.

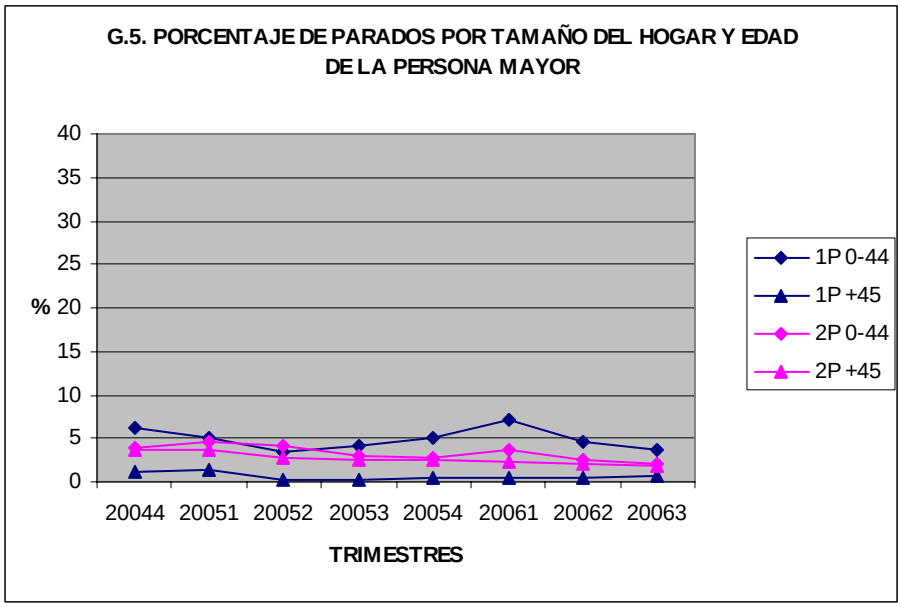
3. Por lo tanto, comparando los dos resultados, es obvio que en los hogares con una persona, la tasa de paro es mucho más extrema, presentando los menores de 44 años una tasa de paro bastante más elevada de lo normal y los mayores de 45 una tasa muy baja.

T.6. Porcentaje de parados de 16 y más años por tamaño del hogar y persona mayor del hogar

		20044	20051	20052	20053	20054	20061	20062	20063
Total	0 A 44 años	4,1	4,6	4,0	3,2	2,9	4,0	2,7	2,2
	45 y más	3,4	3,4	2,7	2,3	2,3	2,2	1,9	1,7
	Total	3,6	3,6	2,9	2,5	2,4	2,5	2,0	1,8
1 Persona	0 A 44 años	6,2	5,2	3,5	4,1	5,0	7,0	4,5	3,7
	45 y más	1,2	1,3	0,3	0,3	0,4	0,4	0,4	0,6
	Total	2,2	2,0	1,0	1,0	1,2	1,5	1,2	1,2
2 ó mas	0 A 44 años	4,0	4,5	4,1	3,1	2,7	3,7	2,5	2,1
	45 y más	3,6	3,6	2,9	2,5	2,5	2,4	2,0	1,9
	Total	3,7	3,7	3,1	2,6	2,5	2,6	2,1	1,9

Fuente: Eustat.PRA

En el siguiente gráfico se muestra el porcentaje de parados por tamaño del hogar y edad de la persona mayor en cada uno de los trimestres.



3.2.2.3 Porcentaje de inactivos de 16 y más años por tamaño del hogar y persona mayor del hogar

Por último se muestran los resultados de la proporción de inactivos:

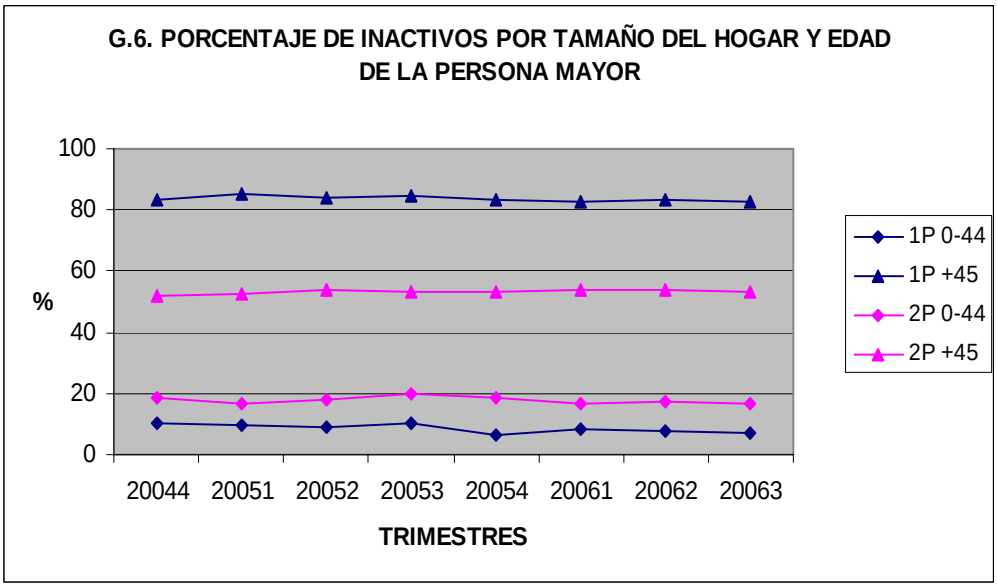
1. En primer lugar, examinando los hogares donde únicamente reside una persona, se observa que hay una gran diferencia entre los mayores de 44 años y los menores de 45. La proporción de inactivos de los primeros varía entre un 82% y un 85%, mientras que la de los segundos oscila entre un 7% y un 10%.

2. Por otro lado, analizando los resultados de los hogares con 2 ó más personas, también existen diferencias significativas entre los mayores y menores de 45 años, pero estas diferencias no son tan extremas. El porcentaje de mayores de 45 años inactivos, fluctúa entre un 51% y un 54%, mientras que la proporción de inactivos menores de 45 años es de aproximadamente oscila entre un 16% y un 19%.

T.7. Porcentaje de inactivos de 16 y más años por tamaño del hogar y persona mayor del hogar

		20044	20051	20052	20053	20054	20061	20062	20063
Total	0 A 44 años	18,0	16,4	17,3	19,1	17,3	15,6	16,4	15,9
	45 y más	54,1	55,1	56,2	55,4	55,6	56,0	56,6	55,7
	Total	47,5	48,6	49,7	49,2	49,0	49,3	49,6	48,6
1 Persona	0 A 44 años	10,3	9,7	9,1	10,2	6,6	8,2	7,5	7,1
	45 y más	83,6	85,4	83,8	84,3	83,4	82,4	83,3	82,4
	Total	69,4	71,4	69,3	70,7	69,0	69,6	69,9	67,7
2 ó mas	0 A 44 años	18,6	17,0	18,1	19,9	18,3	16,4	17,3	16,8
	45 y más	51,9	52,8	54,0	53,0	53,2	53,6	54,1	53,3
	Total	45,8	46,8	48,1	47,4	47,2	47,5	47,8	46,8

Fuente: Eustat.PRA



Como conclusión de estos últimos gráficos y tablas que hemos comentado, cabe destacar que **los porcentajes tanto de parados, ocupados e inactivos son diferentes en los hogares con una persona en el hogar, que en los hogares con 2 personas ó más**.

En los tres casos, los hogares con una persona en el hogar presentan porcentajes extremos, y en muchos de ellos, las diferencias con respeto a la misma categoría de edad en los hogares con 2 personas ó más, es muy considerable.

Cabe destacar que las diferencias existentes en las tablas de ocupados e inactivos, con respecto a la categoría de mayores de 45 años, entre los hogares con una persona ó 2 personas ó más, son muy significativas.

Un ejemplo de ello, es que la proporción de ocupados en la categoría de mayores de 45 años, es mucho mayor en los hogares con 2 ó más personas que en las de uno. Por el contrario, la proporción de inactivos mayores de 45 años es mucho más elevada en los hogares con una sola persona.

Por todo ello, se ha decidido **seleccionar la combinación de las variables tamaño del hogar por edad de la persona mayor del hogar**, y descartar la variable tamaño de municipio debido a que las diferencias en las distintas modalidades no son significativas.

3.3 Dominios

Tras seleccionar la combinación de las variables tamaño del hogar por edad de la persona mayor del hogar debido a que cumple también el segundo principio, se pasaría a la siguiente fase dentro de la selección de la información auxiliar.

En el tercer paso del manual sueco, se señala que las variables auxiliares seleccionadas tienen que identificar los dominios más importantes. Por ello, se ha añadido como información auxiliar la población por TH, sexo y edad, debido a que estas

variables son utilizadas hasta ahora para calcular los pesos de calibración y por su importancia para la publicación de resultados.

Capítulo
4

Aplicación de la calibración para la no respuesta

En este capítulo se van a exponer los resultados obtenidos después de realizar la calibración con la información auxiliar analizada. Esta calibración se va a realizar con la macro CLAN, de Statistics Sweden, que utiliza un estimador lineal generalizado (GREG).

Una vez determinadas las variables auxiliares que pueden intervenir en el cálculo de los pesos de los individuos que han contestado, que en nuestro caso son:

- por un lado, la combinación de las variables tamaño del hogar y edad de la persona mayor del hogar, y
- por otro, la combinación de TH, sexo y edad,

La macro CLAN se utilizó para calcular las estimaciones puntuales correspondientes y las desviaciones estándar de interés en la PRA. CLAN es un programa escrito en lenguaje macro de SAS y está diseñado por la oficina estadística sueca (Andersson y Nordberg 2007).

La macro CLAN puede calcular las estimaciones puntuales basadas en el estimador de Horvitz-Thompson (H.T), o en el estimador de calibración ó de regresión generalizado (GREG). Este último es un método de estimación que utiliza información auxiliar en la etapa de estimación, y es el que hemos utilizado en las distintas pruebas que se han llevado a cabo. La idea de utilizar información auxiliar está basada en la correlación de las variables auxiliares con la variable de estudio.

Se utiliza la información auxiliar para:

- reducir los errores de muestreo
- para reducir el sesgo y la varianza debidos a la falta de respuesta

Las anteriores definiciones se han aplicado sobre la muestra de la Encuesta de Población en Relación con la Actividad (PRA).

4.1 El estimador de regresión generalizado (GREG)

La estimación de regresión significa que para el elemento *k* en la muestra, el par (y_k, $\mathbf{x}_k$) es observado, donde y_k es el valor observado de *y* (la variable de interés), mientras $\mathbf{x}_k$ es el vector de información auxiliar. La metodología también requiere que el total poblacional del vector ***x*** sea conocido.

Para una descripción más detallada de los estimadores de regresión, Särndal C, Swensson B and Wretman (1992). A continuación se ofrece una breve formalización del estimador GREG en situación de falta de respuesta.

Una muestra aleatoria *s* de tamaño $n_{s,}$ es extraída de una población U consistente en N individuos, de acuerdo con el diseño de muestreo *p(.)*, donde todos los individuos tienen una probabilidad >0 de ser incluidos en la muestra. Debido a la no respuesta, los datos de la variable *y* pueden ser recogidos solo para un subconjunto *r* de tamaño m_r. El diseño de muestro *p(.)* en la PRA implica, por ejemplo, que la población es dividida en H estratos, donde el estrato *h* contiene N_h viviendas. En cada estrato *h*, se extrae una muestra aleatoria de tamaño n_h por lo que todas las viviendas tienen la misma probabilidad de ser incluidos en la muestra.

El **estimador de regresión, en general** para un total $t_y = \sum_U y_k$ sería:

$$\hat{t}_y = \sum_r w_k y_k$$

y_k = el valor de la variable *y* para el elemento *k.*

Si lo aplicamos a las situaciones de no respuesta, la notación es la siguiente:

$w_k = g_k \times d_k$ = el peso depende del diseño de muestreo, el vector auxiliar $\boldsymbol{x}_k$ y el modelo utilizado para el ajuste de no respuesta.

$$d_k = 1/\left(\pi_k \hat{\theta}_k\right)$$

π_k = la probabilidad de inclusión del individuo *k*. En la PRA $\pi_k = \frac{n_h}{N_h}$ para todos los individuos que pertenecen al estrato *h.*

$\hat{\theta}_k$ = La probabilidad de respuesta estimada para el individuo *k,* $\hat{\theta}_k = \frac{m_h}{n_h}$.

En realidad, lo que hemos aplicado es $\hat{\theta}_k = \frac{m_{hg}}{n_{hg}}$, asumiendo que todos los individuos del estrato responden de modo independiente y con la misma probabilidad.

Los estratos se han hecho teniendo en cuenta los Grupos de Respuesta Homogéneo (RHG), que se han obtenido del estudio de la información auxiliar, es decir, los grupos formados por la combinación de las variables tamaño del hogar y edad de la persona más mayor del hogar.

Esta estimación está basada en la asunción de que los individuos responden independientemente uno de otro y con la misma probabilidad en el estrato h. (en nuestro caso, dentro del estrato h y de cada RHG)

$$g_k = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)' \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \mathbf{x}_k q_k$$

g_k, es un factor de corrección que refleja la contribución de la información auxiliar para reducir el sesgo debido a la no respuesta y el error de muestreo.

$\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots x_{Jk})'$, es un vector de extensión J, donde J es el nº de variables auxiliares..

$q_{k,}$ es una constante conocida.

$\mathbf{t}_x = (t_{x1}, \ldots, t_{xj}.., t_{xJ})$, es un vector de extensión J, que contiene los totales conocidos del registro.

$\hat{\mathbf{t}}_x = (\hat{t}_{x1}, \ldots, \hat{t}_{xj}.., \hat{t}_{xJ})$, contiene las estimaciones de los totales. $\hat{t}_x = \sum_r d_k x_k$

La **varianza, en situaciones de falta de respuesta,** para $\hat{t}_y$, se estima como:

$$\hat{V}(\hat{t}_{yGREG}) = \sum\sum \frac{\pi_{kl}\hat{\theta}_{kl} - \pi_k \hat{\theta}_k \pi_l \hat{\theta}_l}{\pi_{kl}\hat{\theta}_{kl}} w_k e_k w_l e_l$$

π_{kl}, es la probabilidad de inclusión de segundo orden.

$\hat{\theta}_{kl} = \frac{m_h}{n_h} \frac{m_h - 1}{n_h - 1}$, es la probabilidad estimada de que *k* y *l* pertenezcan a *r* (los respondientes)

$$e_k = y_k - \mathbf{B}'\mathbf{x}_k \qquad \mathbf{B} = \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \sum_r \frac{\mathbf{x}_k y_k q_k}{\pi_k \hat{\theta}_k}$$

4.2 Resultados obtenidos con la macro CLAN

Para el cálculo de estimador GREG se han utilizado Grupos de Respuesta Homogéneos (RHG). El cálculo de los RHGs está basado en la información auxiliar seleccionada en el estudio de la no respuesta, es decir, la combinación de las variables tamaño del hogar y edad de la persona mayor del hogar. De esta forma se han obtenido 4 grupos de respuesta homogéneos:

1. Tamaño del hogar= 1 persona y edad de la persona mayor<45 años.

2. Tamaño del hogar= 1 persona y edad de la persona mayor>=45 años.

3. Tamaño del hogar= Más de una persona y edad de la persona mayor<45 años.

4. Tamaño del hogar= Más de una persona y edad de la persona mayor>=45 años.

Utilizando RHGs, en nuestro caso, cada estrato (territorio o provincia) se divide en los 4 grupos de respuesta homogéneos. Con este modelo de no respuesta, se asume, que las viviendas de cada grupo responden de manera independiente y con la misma probabilidad de respuesta. Como información auxiliar adicional para el estimador de regresión se tomaron las proyecciones de población por territorio, sexo y edad.

Se han obtenido resultados para toda la serie de trimestres. Se incluyen aquí los correspondientes para el primer trimestre de 2005 (20051) y para el tercer trimestre de 2006 (20063) con este método, los resultados publicados (sin tratamiento de no respuesta), y las diferencias entre ambos por territorio histórico y sexo.

Los datos que se muestran son los relativos al total de ocupados, total de parados y a la tasa de paro.

T.8. Población ocupada, parada y tasa de paro por TH y sexo. Resultados obtenidos, resultados publicados y diferencias entre ambos.

			Clan			Publicado			Diferencias Para-Clan		
Año/Trim	Territorio	Sexo	Ocupados	Parados	Tasa paro	Ocupados	Parados	Tasa paro	Ocupados	Parados	Tasa paro
20051	Total	Varon	547,0	34,4	5,9	547,0	33,8	5,8	0,0	-0,6	-0,1
20051	Total	Mujer	392,7	36,3	8,5	389,1	36,9	8,7	-3,6	0,6	0,2
20051	Total	Total	939,7	70,7	7,0	936,1	70,7	7,0	-3,6	0,0	0,0
20051	Alava	Varon	82,5	1,8	2,1	82,7	2,0	2,4	0,2	0,2	0,3
20051	Alava	Mujer	55,9	2,5	4,3	56,6	2,8	4,7	0,7	0,3	0,4
20051	Alava	Total	138,4	4,3	3,0	139,3	4,8	3,3	0,9	0,5	0,3
20051	Gipuzkoa	Varon	179,8	10,6	5,6	177,9	10,4	5,5	-1,9	-0,2	-0,1
20051	Gipuzkoa	Mujer	130,1	10,3	7,4	131,6	10,0	7,1	1,5	-0,3	-0,3
20051	Gipuzkoa	Total	309,9	21,0	6,3	309,5	20,4	6,2	-0,4	-0,6	-0,1
20051	Bizkaia	Varon	284,7	22,0	7,2	286,3	21,4	7,0	1,6	-0,6	-0,2
20051	Bizkaia	Mujer	206,7	23,4	10,2	200,8	24,1	10,7	-5,9	0,7	0,5
20051	Bizkaia	Total	491,4	45,4	8,5	487,2	45,5	8,5	-4,2	0,1	0,0
20063	Total	Varon	556,7	19,0	3,3	556,1	18,6	3,2	-0,6	-0,4	-0,1
20063	Total	Mujer	408,1	16,5	3,9	403,6	17,6	4,2	-4,5	1,1	0,3
20063	Total	Total	964,8	35,5	3,6	959,7	36,2	3,6	-5,1	0,7	0,0
20063	Alava	Varon	83,3	1,8	2,1	83,9	2,1	2,4	0,6	0,3	0,3
20063	Alava	Mujer	60,7	2,5	3,9	59,3	2,5	4,0	-1,4	0,0	0,1
20063	Alava	Total	144,0	4,3	2,9	143,3	4,6	3,1	-0,7	0,3	0,2
20063	Gipuzkoa	Varon	187,2	4,9	2,6	183,9	5,1	2,7	-3,3	0,2	0,1
20063	Gipuzkoa	Mujer	134,7	3,8	2,7	135,9	3,6	2,6	1,2	-0,2	-0,1
20063	Gipuzkoa	Total	322,0	8,8	2,6	319,8	8,7	2,6	-2,2	-0,1	0,0
20063	Bizkaia	Varon	286,2	12,2	4,1	288,3	11,4	3,8	2,1	-0,8	-0,3
20063	Bizkaia	Mujer	212,7	10,3	4,6	208,4	11,4	5,2	-4,3	1,1	0,6
20063	Bizkaia	Total	498,9	22,5	4,3	496,6	22,8	4,4	-2,3	0,3	0,1

Fuente: Eustat.PRA. (Nota: Los totales de ocupados y parados están reflejados en miles, y la tasa de paro en %)

Las diferencias entre ambos métodos son en general pequeñas, por lo que no puede decirse que el ajuste de no respuesta utilizando este tratamiento con las variables auxiliares seleccionadas influya en los resultados que se obtienen.

Capítulo
5

Conclusiones

En este Cuaderno Técnico se han tratado algunos aspectos relacionados con la falta de respuesta en las encuestas a hogares, concretamente. En primer lugar, se ha tratado el tema de cómo medir la falta de respuesta y en segundo lugar, se ha tratado el tema del ajuste o la calibración cuando hay una falta de respuesta en una encuesta a hogares.

En relación al primer punto, cómo medir la falta de respuesta, se ha aplicado el estandar propuesto por el Institute for Social & Economic Research (Lynn et al. 2001) a la Encuesta de Población en Relación con la Actividad (PRA), que es un panel dirigido a hogares.

Una de las conclusiones principales en este punto es la ventaja de utilizar para las encuestas a hogares un modo común o estándar de clasificar las incidencias, o resultados de campo, y también para el cálculo de las tasas. Su uso en encuestas distintas permite la comparación entre ellas, siempre y cuando los diseños sean similares y las diferencias en las tasas de respuesta no sean atribuibles, en su totalidad, a sus diferentes características.

Como consecuencia de lo anterior, actualmente en Eustat se está trabajando en el modo de extender esto a otras encuestas dirigidas a hogares. Este trabajo tiene dos aspectos: uno, el relacionado con los procedimientos informatizados para el cálculo de estas tasas a partir de las incidencias; y otro, el relacionado con la documentación y difusión del método. Con esta finalidad, se ha realizado un manual de incidencias adaptando la propuesta de categorización del ISER. Dado el uso cada vez más generalizado de métodos mixtos de recogida de información, se ha aprovechado la ocasión para extender el manual de incidencias a las entrevistas telefónicas.

En relación al segundo punto, la calibración o el ajuste para la falta de respuesta en las encuestas, se han tratado dos aspectos. En primer lugar, se ha expuesto el método de Statistics Sweden para la selección de información auxiliar, disponible para respondientes y no respondientes de una encuesta, de modo que permita mejorar las estimaciones en situación de falta de respuesta. Este estudio se ha hecho para las variables disponibles en la PRA y se han seleccionado las que mejor cumplen las condiciones, que, muy resumidamente, son: tener relación con la falta de respuesta y además con las variables en estudio.

En segundo lugar, se ha aplicado el estimador de regresión generalizado con esta información auxiliar, y se han vuelto a estimar los resultados. Ello se ha hecho a través de la macro CLAN, de Statistics Sweden, que produce estimaciones, con sus errores de muestreo. Esto supone un paso más respecto a la medición del nivel de falta de respuesta.

Los resultados que se han obtenido para la PRA, con estas nuevas estimaciones, no son en realidad muy diferentes a los ya publicados. Esta semejanza en los resultados podría interpretarse de dos formas, principalmente. Aunque no hay una correspondencia directa entre falta de respuesta y sesgo, podría entenderse que con una tasa de respuesta superior al 80% en el panel de la PRA, en el periodo estudiado (4º trimestre de 2004 hasta 3er trimestre de 2006), quizá el sesgo que se comete no es

muy elevado. Otra interpretación podría ser que la información auxiliar que se utiliza en el estudio, la más adecuada de entre la disponible, según el método, no es suficiente para introducir diferencias en los resultados.

La conclusión principal de esta fase del estudio es que este método que ahora se ha aplicado ofrece muchas posibilidades. En particular, en un futuro inmediato, en el que la información administrativa está siendo más accesible y se está utilizando más en el proceso estadístico. De hecho, ya se están dando pasos para intentar aplicarlo con otras variables auxiliares a esta misma encuesta.

Capítulo 6

Bibliografía

CLAES ANDERSSON AND LENNART NORDBERG

A user´s guide to clan 97. Statistics Sweden. (1998)

CLAES ANDERSSON AND LENNART NORDBERG

Supplement to "a user´s guide to clan 97". . Statistics Sweden. (2006)

EUSTAT

Estandarización y sistematización del cálculo de las tasas de respuesta. Cuaderno de trabajo. (2007) http://www.eustat.es/document/datos/ct_16_c.pdf

SIXTEN LUNDSTRÖM AND CARL-ERIK SÄRNDAL

Estimation in the presence of nonresponse and frame imperfections . . Statistics Sweden. (2001)

SIXTEN LUNDSTRÖM AND CARL-ERIK SÄRNDAL

Estimation in surveys with nonresponse. Wiley. (2005)

PETER LYNN, ROELAND BEERTEN, JOHANNA LAIHO AND JEAN MARTIN.

Recommended Standard Final Outcome Categories and Standard Definitions of Response Rate for Social Surveys.

ISER Working Papers Number 2001-23.

AITOR PUERTA GOICOECHEA

Imputación basada en árboles de clasificación. Cuaderno técnico, eustat. (2002) http://www.eustat.es/document/datos/ct_04_c.pdf

CARL-ERIK SÄRNDAL, BENG SWENSSON, JAN WRETMAN

Model assisted survey sampling. Springer-verlag new york, inc. (1992)