Islas Canarias
Del 15 al 19 de noviembre de 2021

MODELIZACION DE LOS DATOS DE LA ENCUESTA ESIE

Jorge Aramendi
Eustat
j-aramendi@eustat.eus

Marina Ayestaran
Eustat
m-ayestaranarregi@eustst.eus

Introducción

Las encuestas están diseñadas para producir datos a un nivel de agregación determinado, para dominios/áreas, bien geográficas, bien socio-económicas, de la población. Producimos estimaciones de calidad a un nivel de agregación determinado, pero existen demandas de información a un nivel más desagregado. Entre las posibles soluciones podemos:

- aumentar los tamaños muestrales en las áreas en las que es necesario (aumento del coste).
- redistribuir la muestra entre las áreas de una forma más eficiente (no lo arregla todo...).

Sin embargo, siempre se demanda más de lo que se ha especificado en la etapa del diseño. La modelización que hemos introducido en esta encuesta nos permite obtener estimaciones de calidad para múltiples subdominios, aunque como todo, no para cualquiera. Existe criterios mínimos para dar una estimación a partir de los datos modelizados para los subdominios.

Metodología Empleada

Solo hemos modelizado 4 variables de la encuesta ESIE (Encuesta de la sociedad de la información a empresas)

- A16 (tiene internet sí/no)
- C1 (tiene página web sí/no)
- C21 (tiene página web en euskera sí/no)
- B4 (hace trámites online sí/no)

La metodología que tiene en cuenta la estructura de los datos, teniendo en cuenta el diseño y los estratos (se tienen en cuenta los pesos de la encuesta). Se ha definido un modelo para cada variable con las variables auxiliares disponibles.

Para cada variable respuesta los análisis se han realizado período a período, se han modelizados 7 años de datos de la encuesta. Para cada uno de los modelos se ha analizado la bondad de ajuste y capacidad predictiva y se ha optado por una estabilidad temporal de los modelos analizados.

Metodología de la modelización

- Estimación
- Ajuste del modelo
- Capacidad Predictiva
- Validación del ajuste
- Extensión de la probabilidad al conjunto de la población

Estimación

El diseño muestral es complejo y las unidades tienen diferente probabilidad de selección, y por lo tanto diferente peso/elevador. Esto se ha tenido en cuenta en la fase de ajuste. En concreto en la función de máxima verosimilitud, y para ello se ha definido una función de pseudo-verosimilitud:

$$l_D(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i{}^{y_i w_i} \, (1 - p_i)^{(1-y_i)w_i}$$

donde wi es el peso de cada unidad.

Los pesos se tienen en cuenta a la hora de ajustar los modelos y obtener los parámetros, en este caso un modelo de regresión logística. Para ello se ha utilizado el paquete "`survey`" de R, que tiene en cuenta la pseudo-verosimilitud. ("Lumley T (2011). Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R. John Wiley and Sons")

Ajuste del modelo

Para cada año/trimestre y variable se ha ajustado un modelo de regresión logística. Se han validado todas las variables explicativas disponibles y se ha optado por una estabilidad temporal. Se han seleccionado los mejores modelos en función del AIC eta MSE más bajos y la capacidad predictiva (AUC más alto).

Podemos definir MSE como: $MSE = \frac{\sum_{r=1}^{k}\left[\left(\sum_{i=1_r}^{n_r} y_i\right) - n_r \hat{p}_r\right]^2}{\sum_{r=1}^{k} n_r}$

Con, $\hat{p}_r,\ r = 1, \dots, k$ las probabilidades estimadas por el modelo, e $y_i,\ i = 1_r, \dots, n_r$ los valores observados.

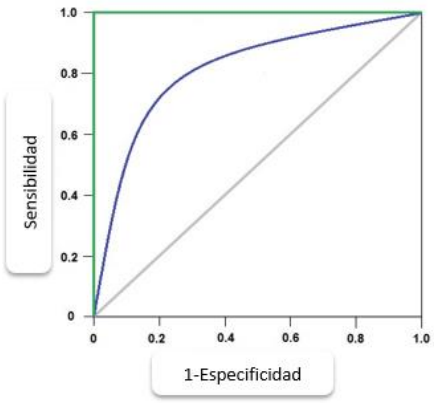
EL parámetro AIC, que tiene en cuenta el muestreo y el diseño (Lumley & Scott, 2015):

$$\mathrm{AIC} = -2\mathrm{nl}_{\mathrm{D}}(\beta) + 2\mathrm{q}\hat{\bar{\delta}}$$

donde $\hat{\bar{\boldsymbol{\delta}}}$ es la traza de la matriz del diseño y q el número de variables explicativas del modelo.

Capacidad predictiva del modelo

Para ello hemos utilizado el parámetro AUC, que tiene la capacidad de medir y diferenciar los individuos que han sido bien clasificados y no.

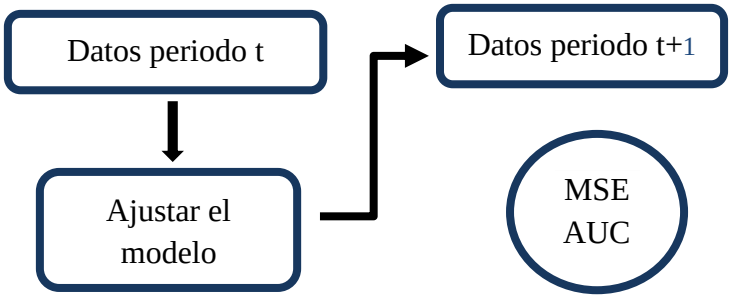


	Clasificación	
Respuesta	$\hat{p}(\boldsymbol{x_i}) < c$	$\hat{p}(\boldsymbol{x_i}) \geq c$
$\boldsymbol{Y = 0}$	Negativo Verdadero (NV)	Falso Positivo (FP)
$\boldsymbol{Y = 1}$	Falso Negativo (FN)	Positivo Verdadero (PV)

- $\mathrm{Especificidad} = \mathrm{P(p(X) < c|Y = 0) = NV/(NV + FP)}$, detecta realmente los casos negativos
- $\mathrm{Sensibilidad} = \mathrm{P(p(X) \geq c|Y = 1) = PV/(PV + FN)}$, detecta realmente los casos positivos

Validación del modelo

Aplicar el modelo a los datos del periodo siguiente y volver a calcular el MSE y el AUC, para validarlo.



Resultados

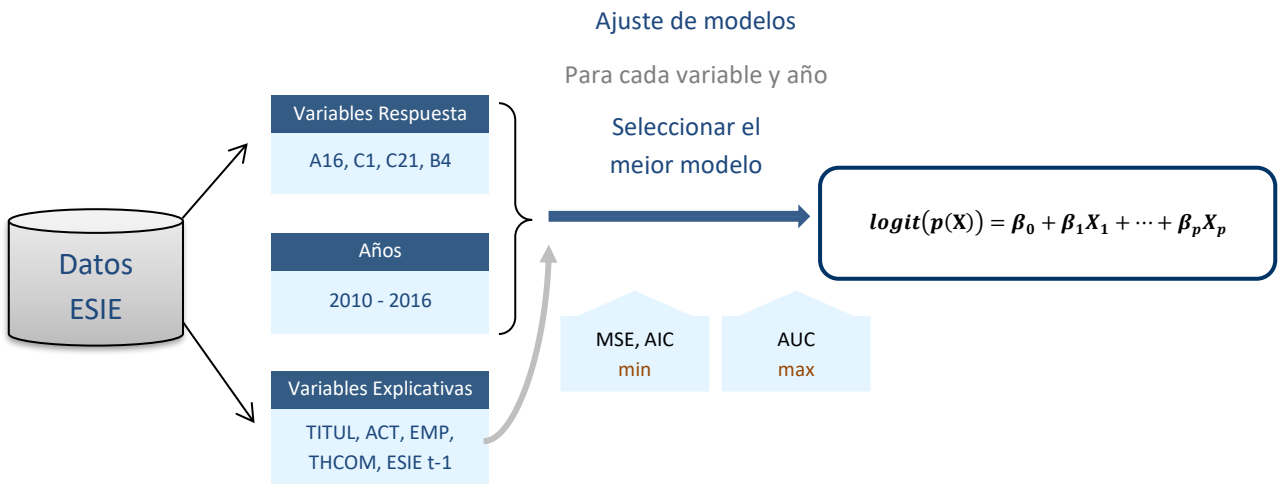
Descripción de la encuesta

- Objetivo de la encuesta: Proporcionar a los responsables políticos, agentes económicos y sociales, universidad, investigadores privados y ciudadanía en general, información periódica sobre la penetración de las nuevas tecnologías de la información y de la comunicación -TIC- en las empresas del País Vasco.
- Una encuesta anual. Panel con renovación anual
- Estratificación: Territorio Histórico x Actividad (2dig de la CNAE) x estrato de empleo (6 categorías)
- Elevación: Se reducen los estratos de empleo a 3 y se hace una corrección de no respuesta.
- Esta encuesta está preparada para obtener resultados de calidad para los TH-s y a nivel de actividad para la CAE
- Información auxiliar: DIRAE, Directorio de Actividades Económicas de Eustat

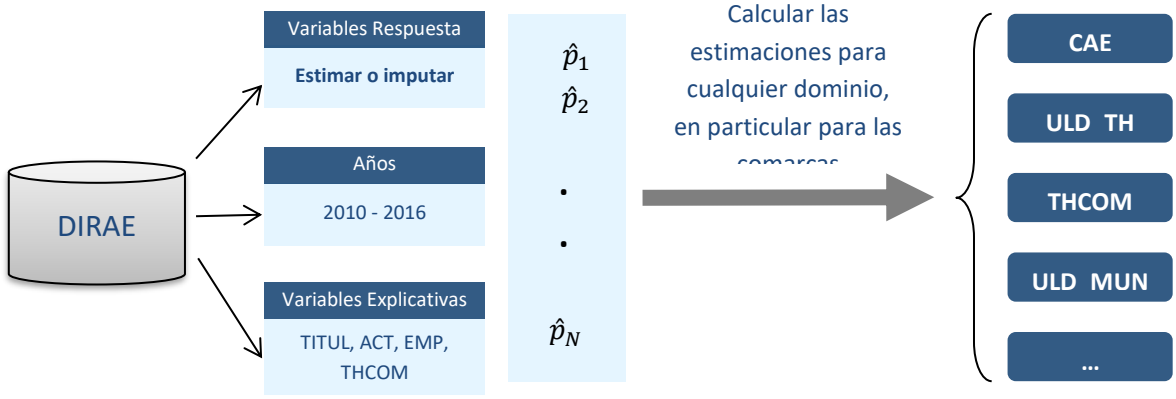
Variables Explicativas para los modelos

- Agregaciones de la CNAE
- Empleo estratificado: 1) 0-2, 2) 3-9, 3) 10-19, 4) 20-49, 5) 50-99, 6) 100 y mas
- Titularidad
- Presencia en el estado
- Estimación sintética de la variable en el año anterior.

Proceso de Modelización



Obtención probabilidades para todas las unidades de la población



Parámetros y estadísticos de la modelización de una variable en concreto. Significatividad de los para metros del modelo

	2010	2011	2012	2013	2014	2015	2016
THCOM23	0.002	0.072	0.775	0.266	0.112	0.029	0.640
A21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.003
EMP	<0.001	<0.001	<0.001	0.009	<0.001	0.023	<0.001
TITUL	<0.001	0.001	0.005	0.061∗	0.002	0.058∗	0.009
A16.2		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
AIC	7152	4064	4139	3844	3483	3420	3316
MSE	0.117	0.050	0.057	0.049	0.043	0.038	0.039
t+1	(0.118)	(0.067)	(0.053)	(0.048)	(0.047)	(0.049)	
AUC	0.864	0.956	0.946	0.948	0.947	0.942	0.925
t+1	(0.866)	(0.940)	(0.943)	(0.944)	(0.930)	(0.904)	

Resultados por Territorio Histórico, diferencias entre la probabilidad estimada y la estimación directa. (Consistencia de las estimaciones)

- Alava

	2010	2011	2012	2013	2014	2015	2016
A16	0.015	0.006	0.005	-0.004	-0.003	0.000	-0.008
C1	-0.062	-0.039	-0.023	0.000	0.017	0.010	0.013
C21	-0.005	-0.003	0.008	0.008	-0.007	-0.011	-0.026
B4	-0.030	-0.033	0.002	0.021	0.012	-0.008	0.011

- Bizkaia

	2010	2011	2012	2013	2014	2015	2016
A16	-0.009	-0.015	-0.014	-0.007	-0.009	-0.008	-0.012
C1	-0.021	-0.038	-0.034	-0.039	-0.026	-0.035	-0.034

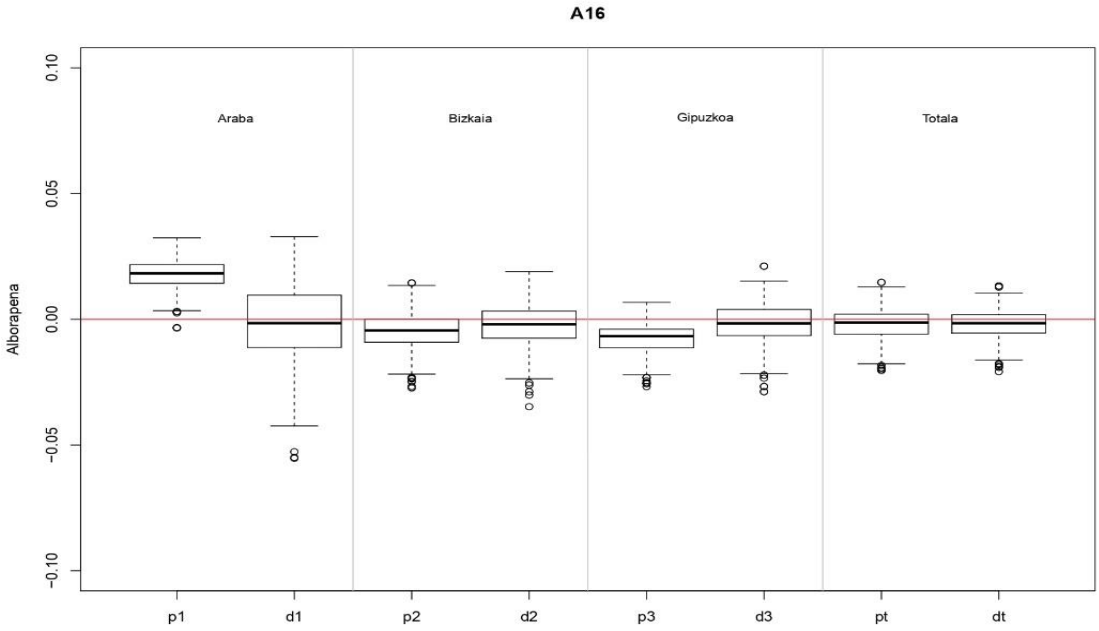
C21	-0.017	0.000	0.002	-0.012	-0.014	-0.027	-0.033
B4	-0.025	-0.001	-0.030	-0.026	-0.019	-0.028	0.004

- Gipuzkoa

	2010	2011	2012	2013	2014	2015	2016
A16	-0.011	-0.015	-0.019	-0.012	-0.007	-0.010	-0.008
C1	-0.035	-0.015	-0.031	-0.003	-0.030	-0.021	-0.017
C21	-0.030	-0.013	-0.002	0.018	0.021	0.024	0.017
B4	-0.031	-0.034	-0.043	-0.028	-0.011	-0.001	-0.016

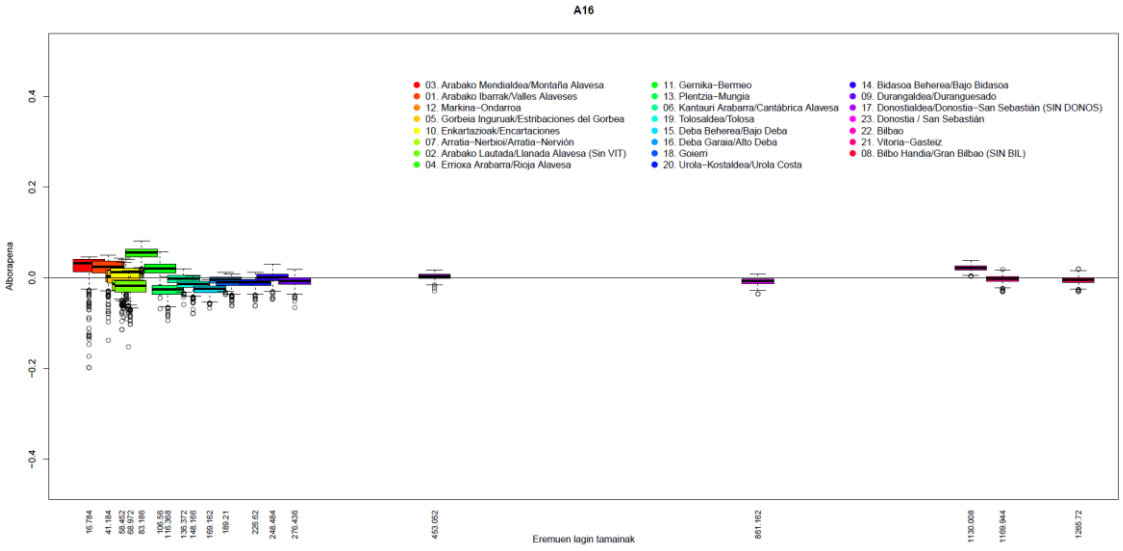
Calidad de los resultados

Mediante un estudio de simulación se ha estudiado la calidad de las estimaciones obtenidas a partir de los modelos desarrollados. No existe una población de referencia por lo que se ha simulado una población para cada una de las variables y se han realizado 500 muestreos siguiendo el diseño de la encuesta. Se ha creado la Pseudo-población estratificado y aplicando aleatoriamente los resultados de la encuesta. No se utilizado el mismo modelo que para estimar, para evitar sobre-ajustes de los datos Para cada una de estas muestras se ha ajustado el modelo y se han obtenido las estimaciones.



p : estimación del modelo
d : estimación directa

Resultados comarcales



- Para la variable A16: el sesgo absoluto medio más alto obtenido es de 0,036 con una desviación estándar de 0,04.
- Para las variables B4, C1 y C21: el sesgo absoluto medio crece, estando en todos los casos próximo a 0,13. En cualquier caso, este valor se da en el dominio más pequeño que se ha estudiado, Arabako Mendialdea/ Montaña Alavesa (de media tiene un tamaño muestral de 15 individuos). Si dejamos a un lado el dominio Arabako Mendialdea/ Montaña Alavesa, el sesgo absoluto medio más alto se reduce a la mitad.
- Para las variables B4 y C21: se han utilizado dos métodos diferentes para calcular el sesgo. En base a las simulaciones realizadas en los dominios analizados, no se han encontrado diferencias entre ambos métodos. Sin embargo, en el caso de C21, los resultados obtenidos con el método B son ligeramente mejores.

Límites de desagregación de las estimaciones según el estudio de la calidad de los resultados.

Variable	Muestra en el Dominio	Desviación media absoluta
A16	$50 \leq n \leq 100$	<0.025
	$100 < n \leq 500$	<0.02
	$n > 500$	<0.008
B4	$50 \leq n \leq 100$	<0.08
	$100 < n \leq 500$	<0.05
	$n > 500$	<0.02
C1	$50 \leq n \leq 100$	<0.08
	$100 < n \leq 500$	<0.05
	$n > 500$	<0.03
C21	$50 \leq n \leq 100$	<0.05

	$100 < n \leq 500$	<0.026
	$n > 500$	<0.02

Conclusiones

Puntos Fuertes

- Con esta metodología se pueden obtener resultados para cualquier dominio de la población. (teniendo en cuenta los límites establecidos en el estudio de calidad)
- Se ha tenido en cuenta el diseño del muestro en la estimación del modelo.
- La capacidad predictiva del modelo es muy alta.

Puntos Débiles

- La información relativa a las variables predictoras que se quieran introducir en el modelo tiene que estar disponible en DIRAE para poder obtener estimaciones en cualquier dominio de la población.