MODELIZACIÓN DE LA ENCUESTA DE POBLACIÓN EN RELACIÓN CON LA ACTIVIDAD (PRA) PARA REALIZAR ESTIMACIONES EN ÁREAS PEQUEÑAS

INTRODUCCIÓN

Las estimaciones a nivel municipal de la PRA están basadas en el ajuste de modelos de regresión logística a los datos de la encuesta. Utilizando esta metodología, y teniendo en cuenta el diseño de muestreo, se han desarrollado modelos que permiten estimar las probabilidades de estar activo, parado y ocupado para cada individuo de la población de interés. Teniendo las probabilidades a nivel individual para toda la población, tendremos la capacidad de estimar las probabilidades para cualquier área de interés.

La idoneidad de los modelos y su capacidad predictiva se han analizado mediante parámetros de calidad que han permitido seleccionar modelos equilibrados en cuanto a complejidad y bondad del ajuste. Además, se han realizado estudios de la calidad de las estimaciones para ciertos dominios que garantizan la robustez de los modelos seleccionados.

Este trabajo se ha realizado de forma conjunta con el Departamento de Matemática Aplicada, Estadística e Investigación Operativa de la Facultad de Ciencia y Tecnología de la UPV/EHU. Eustat colabora desde hace años con este Departamento en diversos proyectos de estimación en áreas pequeñas.

METODOLOGÍA

Los métodos de regresión explican la relación entre una variable de respuesta y una o más variables explicativas. En muchas ocasiones, la distribución de la variable respuesta es binomial, siendo en este caso la **regresión logística** el método más habitual para llevar a cabo la modelización. Este es también el caso que nos ocupa ya que se van a modelizar variables dicotómicas (el individuo es activo o no, o el individuo está ocupado o no).

Sea $X=(X_1,X_2,\ldots,X_q)$ el vector de q variables explicativas e Y la variable de respuesta dicotómica. Entonces, el modelo de regresión logística es el siguiente:

$$p(X) = P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_q X_q}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_q X_q}} \in (0,1)$$

Sea $\beta=(\beta_0,\ \beta_1,\ldots,\beta_q)$ el vector de coeficientes de regresión. Supongamos que tenemos una muestra de tamaño n de la población (X,Y) con valores $\{x_j,y_j\}$ con $j=1\ldots n$. Sea,

$$p_j = p(\boldsymbol{x}_j) = P(Y=1|\boldsymbol{X}=\boldsymbol{x}_j) = \frac{e^{\beta_0+\beta_1 x_{j1}+\cdots+\beta_q x_{jq}}}{1+e^{\beta_0+\beta_1 x_{j1}+\cdots+\beta_q x_{jq}}},$$

la probabilidad condicional de que el individuo j experimente el evento. Entonces, la función de verosimilitud es la siguiente:

$$l(\beta) = \prod_{j=1}^{n} p_j^{y_j} (1 - p_j)^{(1-y_j)}$$

donde:

$$p_j = P(Y = 1 | X = x_j)$$

Sea $L(\beta)$ el logaritmo de la función de verosimilitud, es decir:

$$L(\beta) = \sum_{j=1}^{n} [y_j \ln p_j + (1 - y_j) \ln(1 - p_j)]$$

El objetivo del método de máxima verosimilitud es estimar los parámetros β que maximizan $L(\beta)$.

Sin embargo, cuando se realiza un muestreo complejo y los individuos de la muestra tienen diferentes pesos, estos pesos también deben tenerse en cuenta en la función de verosimilitud (Binder, 1983). En el caso de la PRA, este muestreo se ha realizado en una etapa, es decir, se han definido diferentes estratos en la población y el muestreo se ha realizado aleatoriamente dentro de estos estratos. En este caso, la función de pseudo-verosimilitud se define de la siguiente manera:

$$l_D(\beta) = \prod_{j=1}^{n} p_j^{y_j w_j} (1 - p_j)^{(1-y_j) w_j}$$

Donde ωj *es* el peso del individuo j en la muestra $\{x_j, y_j\}$ con $j=1 \ldots n$.

Por lo tanto, el diseño de muestreo se tiene en cuenta a la hora de ajustar los modelos de regresión logística y estimar los coeficientes de regresión.

DESCRIPCIÓN DE VARIABLES Y MODELOS AJUSTADOS

Las variables respuesta de la PRA que se van a modelizar han sido:

- El individuo es ACTIVO (SI/NO).
- De entre los activos, si el individuo está OCUPADO (SI/NO).

Todo ello para la población de 16 y más años de la C.A. de Euskadi.

Las variables explicativas que se van a utilizar en los modelos deben tener relación con las variables respuesta y, además, deben estar recogidas para todos los elementos de la población. Por lo tanto, se han testeado diversas variables socio-demográficas recogidas en estadísticas censales y otras incluidas en fuentes administrativas relacionadas con la actividad y el empleo de la población, tales como la Seguridad Social, Lanbide y/o Muface.

Las variables más significativas y que finalmente han sido incluidas en los modelos son:

- **TERRITORIO HISTÓRICO**
  - Araba/Álava
  - Guipuzkoa
  - Bizkaia
- **SEXO**
  - Hombre
  - Mujer
- **NACIONALIDAD**
  - Española
  - Otra
- **SITUACIÓN RESPECTO A LA OCUPACIÓN**
  - Sí
  - No
- **EDAD CUMPLIDA** dividida en 7 categorías:
  - 16-24 años
  - 25-29 años
  - 30-34 años
  - 35-44 años
  - 45-54 años
  - 55-64 años
  - 65 años o más
- **NIVEL DE EUSKERA**
  - Euskaldunes
  - Casi euskaldunes
  - No euskaldunes
  - Desconocido
- **TITULACIÓN**
  - Sin estudios
  - Primarios
  - Profesionales
  - Secundarios
  - Medios-superiores
  - Superiores
  - Desconocida

Con estas variables explicativas, y para cada trimestre, se han ajustado dos modelos diferentes para las dos variables respuesta consideradas:

- El modelo para la variable respuesta ACTIVO (SI/NO) se ha ajustado utilizando la información de todos los individuos de la encuesta y como variables explicativas el Territorio Histórico, el sexo, la edad cumplida, la nacionalidad, la titulación y la relación respecto a la ocupación.
- El modelo para la variable respuesta OCUPADO (SI/NO) se ha ajustado considerando solo los individuos activos e incluye como variables explicativas la edad cumplida, la nacionalidad, la titulación, la relación respecto a la ocupación y el nivel de Euskera.

Del modelo ajustado para la variable respuesta ACTIVO (SI/NO) se estiman las probabilidades p_i de estar activo/a para cada individuo i en la población. La probabilidad complementaria (1-p_i) será, por lo tanto, la probabilidad de ser inactivo/a en la población.

Del mismo modo, del modelo ajustado para la variable respuesta OCUPADO (SI/NO) se estiman las probabilidades p_i de estar empleado/a para cada individuo en la población. La probabilidad complementaria (1-p_j) será, por lo tanto, la probabilidad del individuo de estar en paro.

BONDAD DEL AJUSTE Y CAPACIDAD PREDICTIVA DE LOS MODELOS

Para evaluar la bondad del ajuste de los modelos, se han calculado el error cuadrático medio (MSE) y el parámetro AIC (Criterio de información de Akaike). La expresión utilizada para calcular el MSE es la siguiente:

$$MSE = \frac{\sum_{r=1}^{k} \left(\sum_{j=1}^{l_r} y_j / n_r - l_r \hat{p}_r \right)^2}{\sum_{r=1}^{k} l_r}$$

siendo l_r el número de individuos con la probabilidad p_r, y k el número de probabilidades diferentes en la muestra.

En cuanto al parámetro AIC, se ha utilizado una corrección del mismo que tiene en cuenta el diseño y el muestreo de la encuesta (Lumley y Scott, 2015).

Específicamente,

$$AIC = -2 \cdot \frac{1}{N} \ln l_D(\hat{\beta}) + 2q\delta$$

donde δ es la traza de la matriz de diseño y q es el número de parámetros del modelo.

Un valor del AIC más bajo indica un modelo que se ajusta mejor a los datos, teniendo en cuenta la complejidad del modelo. Es decir, penaliza los modelos con más parámetros para evitar el sobreajuste.

Para evaluar la capacidad predictiva se ha calculado el parámetro AUC (área bajo la curva ROC). La AUC mide la capacidad del modelo para distinguir entre individuos que han experimentado el evento y aquellos que no lo han hecho, siguiendo una distribución de Bernoulli. Toma valores entre 0,5 y 1. Un AUC de 0,5 corresponde a un modelo que no proporciona información (equivalente a lanzar una moneda al aire). Un modelo con capacidad predictiva perfecta tiene un AUC de 1. Según la literatura, un modelo con una capacidad predictiva superior a 0,8 se puede considerar excelente (Steyerberg, 2008).

A continuación, se muestran los valores de estos parámetros para los dos modelos ajustados y para diferentes periodos de la encuesta:

MODELO “ACTIVOS”

	2016-1	2016-2	2016-3	2016-4	2017-1	2017-2	2017-3	2017-4
AIC	2989	3125	2499	2229	2110	2104	2179	2087
MSE	0.0714 (0.1378)	0.102 (0.145)	0.104 (0.0626)	0.052 (0.0601)	0.052 (0.113)	0.064 (0.092)	0.101 (0.069)	0.057
AUC	0.8963 (0.8821)	0.8854 (0.9084)	0.9227 (0.9287)	0.9308 (0.9244)	0.9268 (0.9313)	0.9331 (0.9325)	0.935 (0.9268)	0.927

(*) Las cifras entre paréntesis representan el valor del parámetro al aplicar el modelo de un trimestre a los datos del trimestre siguiente

MODELO “OCUPADOS”

	2016-1	2016-2	2016-3	2016-4	2017-1	2017-2	2017-3	2017-4
AIC	4864	4973	4684	4395	4497	4456	4532	4161
MSE	0.0967 (0.0979)	0.102 (0.129)	0.0877 (0.0995)	0.091 (0.099)	0.091 (0.073)	0.065 (0.0797)	0.071 (0.087)	0.074
AUC	0.9708 (0.9685)	0.9688 (0.9712)	0.9719 (0.9739)	0.9739 (0.9724)	0.9728 (0.9746)	0.9748 (0.9752)	0.9750 (0.9775)	0.9780

(*) Las cifras entre paréntesis representan el valor del parámetro al aplicar el modelo de un trimestre a los datos del trimestre siguiente

Se puede observar que con ambos modelos se han obtenido muy buenos resultados, tanto en la adecuación del ajuste (MSE < 0,11 y AIC bajos) como en la capacidad predictiva (AUC altos > 0,89).

CÁLCULO DE LAS TASAS POBLACIONALES

Para el cálculo de las tasas de actividad, ocupación y paro para el dominio D se han utilizado las probabilidades individuales estimadas por los modelos para cada individuo de la población de interés. Sean:

$$p.activo_i = \text{probabilidad estimada por el modelo de que el individuo } i \text{ sea activo}$$
$$p.ocupado_i = \text{probabilidad estimada por el modelo de que el individuo } i \text{ sea ocupado}$$
$$p.parado_i = 1 - p.ocupado_i = \text{probabilidad estimada por el modelo de que el individuo } i \text{ sea parado}$$

Entonces las tasas para el dominio D definen como:

$$Tasa\ actividad_D = \frac{\sum_{i=1}^{N_D} p.activo_i}{N_D}$$

$$Tasa\ de\ ocupación_D = \frac{\sum_{i=1}^{N_D} p.ocupado_i}{N_D}$$

$$Tasa\ de\ paro_D = \frac{\sum_{i=1}^{N_D} p.paro_i}{\sum_{i=1}^{N_D} p.activo_i}$$

con N_D =población de 16 y más años en el domino D[1]

[1] Se trata de la población estimada a partir de las últimas proyecciones de población disponibles en el momento de la elevación de la encuesta.

Las tasas anuales que se publican a nivel municipal se calculan a partir de los promedios trimestrales estimados por los modelos para el total de parados, ocupados y activos respectivamente.

BIBLIOGRAFÍA

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique,* 279-292.

Lumley, T. (2011). *Complex surveys: a guide to analysis using R.* John Wiley & Sons.

Lumley, T., & Scott, A. (2014). Tests for regression models fitted to survey data. *Australian & New Zealand Journal of Statistics* , 1-14.

Lumley, T., & Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology* , 1-18.

Steyerberg, E. W. (2008). *Clinical prediction models: a practical approach to development, validation, and updating.* Springer Science & Business Media.

SOFTWARE

Para el ajuste de los modelos y el análisis de la calidad se ha utilizado la versión 3.4.3 del software R. Concretamente, se han utilizado los paquetes `survey` y `simPopulation`.