MODELLING OF THE SURVEY ON POPULATION IN RELATION TO ACTIVITY (PRA) FOR SMALL AREA ESTIMATION

INTRODUCTION

Municipal-level estimates for the PRA are based on fitting logistic regression models to the survey data. Using this methodology, and taking into account the sampling design, models have been developed that allow us to estimate the probabilities of being active, unemployed and employed for each individual in the population of interest. By having individual-level probabilities for the entire population, we will be able to estimate the probabilities for any area of interest.

The suitability of the models and their predictive capacity have been analysed using quality parameters that have enabled us to select models that are balanced in terms of complexity and goodness of fit. In addition, studies have been conducted on the quality of the estimates for certain domains to ensure the robustness of the models selected.

This work has been carried out in collaboration with the Department of Applied Mathematics, Statistics and Operations Research of the Faculty of Science and Technology at the University of the Basque Country. Eustat has been working with this Department on various small area estimation projects for many years.

METHODOLOGY

Regression methods explain the relationship between a response variable and one or more explanatory variables. Often, the distribution of the response variable is binomial, in which case **logistic regression** is the most common method for modelling. This is also the case here, as dichotomous variables (the individual is active or not, or the individual is employed or not) will be modelled.

Let $X=(X_1,X_2,\ldots,X_q)$ be the vector of q explanatory variables and Y be the dichotomous response variable. Then, the logistic regression model is as follows:

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

Let $\beta=(\beta_0, \beta_1,\ldots,\beta_q)$ be the vector of regression coefficients. Let us suppose that we have a sample size n of the population (X,Y) with values $\{x_j,y_j\}$ with $j=1\ldots n$. Let

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

be the conditional probability that the individual j experiences the event. Then, the likelihood function is as follows:

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

where:

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

Let $L(\beta)$ be the logarithm of the likelihood function, i.e.:

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

The goal of the maximum likelihood method is to estimate the parameters β that maximise $L(\beta)$.

However, when complex sampling is carried out and the individuals in the sample have different weights, these weights must also be taken into account in the likelihood function (Binder, 1983). In the case of the PRA, this sampling has been carried out in one stage, i.e. different population strata have been defined and sampling has been carried out randomly within these strata. In this case, the pseudolikelihood function is defined as follows:

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

Where ωj is the weight of the individual j in the sample $\{x_j, y_j\}$ with $j = 1 \ldots n$.

Therefore, the sampling design is taken into account when fitting logistic regression models and estimating regression coefficients.

DESCRIPTION OF VARIABLES AND FITTED MODELS

The response variables for the PRA to be modelled were:

- The individual is ACTIVE (YES/NO).
- Among those who are active, whether the individual is EMPLOYED (YES/NO).

All the above for the population aged 16 and over in the Basque Country.

The explanatory variables to be used in the models must be related to the response variables and must also be collected for all the elements of the population. Therefore, various socio-demographic variables collected in census statistics and others included in administrative sources related to the activity and employment of the population, such as Social Security, Lanbide [Basque Employment Service] and/or Muface [General Mutual Benefit Society for Civil Servants of the State], have been tested.

The most significant variables that were ultimately included in the models are:

- **PROVINCE**
  - Araba/Álava
  - Gipuzkoa
  - Bizkaia
- **SEX**
  - Male
  - Female
- **NATIONALITY**
  - Spanish
  - Other
- **SITUATION REGARDING EMPLOYMENT**
  - Yes
  - No
- **AGE REACHED** divided into 7 categories:
  - 16-24 years old
  - 25-29 years old
  - 30-34 years old
  - 35-44 years old
  - 45-54 years old
  - 55-64 years old
  - 65 years old or over
- **LEVEL OF BASQUE**
  - Basque speakers
  - Quasi-Basque speakers
  - Non-Basque speakers
  - Unknown
- **QUALIFICATIONS**
  - No studies
  - Primary
  - Professional
  - Secondary
  - Intermediate-higher
  - Higher
  - Unknown

Using these explanatory variables, and for each quarter, two different models have been fitted for the two response variables considered:

- The model for the response variable ACTIVE (YES/NO) has been fitted using information on all the individuals in the survey and includes province, sex, age reached, nationality, qualifications and situation regarding employment as explanatory variables.
- The model for the response variable EMPLOYED (YES/NO) has been fitted considering only active individuals and includes age reached, nationality, qualifications, situation regarding employment and level of Basque as explanatory variables.

Using the fitted model for the response variable ACTIVE (YES/NO), the probabilities p_i of being active are estimated for each individual i in the population. The complementary probability (1-p_i) will therefore be the probability of being inactive in the population.

Similarly, using the fitted model for the response variable EMPLOYED (YES/NO), the probabilities p_i of being employed are estimated for each individual in the population. The complementary probability (1-p_j) will therefore be the probability of the individual being unemployed.

GOODNESS OF FIT AND PREDICTIVE CAPACITY OF THE MODELS

In order to assess the goodness of fit of the models, the mean squared error (MSE) and AIC (Akaike information criterion) parameter have been calculated. The expression used to calculate the MSE is as follows:

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

l_r being the number of individuals with probability p_r, and k the number of different probabilities in the sample.

As for the AIC parameter, an AIC correction has been used that takes into account the survey design and sampling (Lumley and Scott, 2015).

Specifically,

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

where δ is the trace of the design matrix and q is the number of parameters in the model.

A lower AIC value indicates a model that fits the data better, taking into account the complexity of the model. In other words, it penalises models with more parameters in order to avoid overfitting.

In order to assess predictive capacity, the AUC (area under the ROC curve) parameter has been calculated. The AUC measures the model's capacity to distinguish between individuals who have experienced the event and those who have not, following a Bernoulli distribution. It takes values between 0.5 and 1. An AUC of 0.5 corresponds to a model that provides no information (equivalent to flipping a coin). A model with perfect predictive capacity has an AUC of 1. According to literature, a model with a predictive capacity greater than 0.8 can be considered excellent (Steyerberg, 2008).

Below are the values of these parameters for the two fitted models and for different survey periods:

"ACTIVE INDIVIDUALS" MODEL

	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

(*) The figures in brackets represent the value of the parameter when applying the model for one quarter to the data for the following quarter

"EMPLOYED INDIVIDUALS" MODEL

	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

(*) The figures in brackets represent the value of the parameter when applying the model for one quarter to the data for the following quarter

We can see that very good results have been obtained with both models, both in terms of goodness of fit (MSE < 0.11 and low AICs) and predictive capacity (high AUCs > 0.89).

CALCULATION OF POPULATION RATES

In order to calculate the activity, employment and unemployment rates for domain D, the individual probabilities estimated by the models for each individual in the population of interest have been used. Let:

$active\ person_i =$ probability estimated by the model that individual i is active
$employed\ person_i =$ probability estimated by the model that individual i is employed
$unemployed\ person_i = 1 - unemployed\ person_i =$
probability estimated by the model that individual i is unemployed

Then the rates for domain D are defined as:

$$Activity\ rate_D = \frac{\sum_{i=1}^{N_D} active\ person_i}{N_D}$$

$$Employment\ rate_D = \frac{\sum_{i=1}^{N_D} employed\ person_i}{N_D}$$

$$Unemployment\ rate_D = \frac{\sum_{i=1}^{N_D} unemployed\ person_i}{\sum_{i=1}^{N_D} active\ person_i}$$

with N_D=population aged 16 and over in domain D[1]

[1] This is the estimated population based on the latest population projections available at the time of the survey.

The annual rates published at municipal level are calculated based on the quarterly averages estimated by the models for the total number of unemployed, employed and active individuals, respectively.

BIBLIOGRAPHY

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

For model fitting and quality analysis, version 3.4.3 of R software was used. Specifically, the `survey` and `simPopulation` packages were used.