The number of Latvian residents estimation via logistic regression

Jeļena Vaļkovska[1]

[1]Central Statistical Bureau of Latvia, e-mail: jelena.valkovska@csb.gov.lv

Abstract

There were 2 070 371 residents of Latvia on 1 March 2011 according to the results of Population and Housing Census 2011. Further data processing and routine statistics production will be based on the Population and Housing Census 2011 results. Therefore, it is very important to know the number of the residents in some fixed moment of time. It was decided to investigate the logistic regression as the potential instrument. The main aim of this research is to get such estimates of beta coefficients, that can be used to estimate the number of Latvian residents and number of emigrants now and in the future.

Keywords: census, logistic regression, residents, estimators, balanced sampling

1 Introduction

The Population and Housing Census was organised in March 2011 in compliance with the Law on Population and Housing Census and European Parliament and Council Regulation (EC) No 763/2008 of July 9, 2008. The main aim of it was to obtain detailed enough view on structure and characteristics of the population.

Resident population comprised all persons who:

1. have lived in Latvia for at least 12 months before the Census moment (01.03.2011);
2. have arrived in Latvia within 12 months before the Census moment with an intention to spend at least one year in the country.

Persons not counted in the Census:

1. registered at the Population Register, but residing outside of the territory of Latvia for more than 12 months;
2. have entered the country less than 12 months before the Census moment, are residing in the country but are not planning to stay in Latvia for more than 12 months;
3. individuals born after 01.03.2011;
4. individuals who have died before 01.03.2011;
5. foreign armed forces, sea force and consular personnel and their family members residing the country;
6. tourists. (CSB home page: Population Census)

2 Logistic regression

The aim of an analysis using logistic regression is the same as that of any model-building technique used in statistic: to find the best fitting and most parsimonious model to describe the relationship between dependent variable and a set of independent variables. What distinguishes logistic regression from the linear regression model is that the outcome variable in logistic regression is dichotomous.

Let the vector of t independent variables is $x^{'} = (x_1, x_2, ..., x_t)$. The quantity (1) is used in multiply regression to represent the conditional mean Y given x and quantity (2) is the logits transformation.

$$\pi(x) = E(Y|x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_t x_t}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_t x_t}} \tag{1}$$

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_t x_t} \tag{2}$$

The logit $g(x)$ has many of the desirable properties of a linear regression model: it is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$, depending on the range of x.

The model fitting means that it will be obtained the estimates of vector $\beta^{'} = (\beta_1, \beta_2, ..., \beta_t)$. The method of estimation used is the maximum likelihood. The main principle of this method is that is used as the estimate of β the value which maximizes the expression:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)^{(1-y_i)}\right]. \tag{3}$$

The comparison of observed to predicted values using the likelihood function is based on the expression:

$$D = -2ln\left[\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model}\right]. \tag{4}$$

Probabilities are always less than one, so the value of *log likelihood* (LL) is always negative, that is why the expression is multiplied by (-2). Such test is called likelihood ratio test:

$$D = -2ln\sum_{i=1}^{n}\left\{y_i ln\left[\frac{\hat{\pi}(x_i)}{y_i}\right] + (1-y_i)ln\left[\frac{1-\hat{\pi}(x_i)}{1-y_i}\right]\right\}. \tag{5}$$

The statistic D often is called the deviance (it is also known as -2 *log likelihood* ($-2LL$)) and plays for the logistic regression the same role that the residual sum of squares plays in linear regression, it is a measure of how well the estimated model fits the data. The smaller the deviance is, the better the model fits the data.

To ascertain the significance of an independent variable it is necessary to compare the value of D with and without the independent variable in the equation, i.e.:

$$G = D(model\ without\ the\ variable) - D(model\ with\ the\ variable). \tag{6}$$

The statistics G plays the same role in logistic regression as the numerator of the partial F-test plays in linear regression and can be expressed by:

$$G = -2ln\left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right]. \tag{7}$$

Under the null hypothesis that the t "slope" coefficients for the covariates in the model are equal to zero, the distribution of G will be chi-square with t degrees-of-freedom.

The other similar, statistically equivalent test is the univariate Wald test. It is obtained by comparing the maximum likelihood estimate of slope parameter, $\hat{\beta}_i$, to an estimate of its standard error:

$$W_i = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}. \tag{8}$$

Under the hypothesis, that individual $\hat{\beta}_i = 0$, the distribution of univariate Wald test statistics will be the Standard normal distribution. (Hosmer & Lemeshov, 2000)

2.1 Auxiliary variables

First of all the data were taken from the Population Register:

- Sex;
- Age;
- Citizenship;
- Nationality;
- Living region;
- Marital status;
- Country of birth.

Characteristics were recoded as dichotomous variables (i.e. is coded as either 0 or 1) for each individual, for example,

$$x_{1i} = \begin{cases} 1, & \text{if person is male} \\ 0, & \text{if person is female.} \end{cases}$$

Through in *SPSS* conducted experiment we could see, that *Nagelkerke R Square* value were lower than 0.5 and -2 *Log likelihood* statistics value is large, so we can conclude, that this information was not sufficient. Additional data was taken from other available sources of information:

- Information of employment;
- Whether the person is employed or self-employed;
- Personal income of the period (continuous variable).

2.2 The experiment and some conclusions

Using the logistic regression were computed the estimates of beta coefficients and probabilities of the case, that person is the resident of Latvia.

Through experiment we could see that the minimum probability of a case that a person is the resident of Latvia is quite large (about thirty percent). The estimated value of residents is much greater than the real value. Probably the main problem is that the number of emigrants in the investigated population part is too negligible (only 3% of the population). From the other side the unknown population part can be different from population part which the whole information is available. It was considered to select a balanced sample (balancing on the independent variables totals of unknown population part).

2.3 Balanced sampling

Definition. (Tillé, 2006) A sampling design $p(s)$ is said to be balanced with respect to the auxiliary variables $x_1, x_2, ..., x_p$, if and only if it satisfies the balancing equation given by

$$\hat{X}_{HT} = X. \tag{9}$$

which can also be written

$$\sum_{k \in U} \frac{x_{kj} S_k}{\pi_k} = \sum_{k \in U} x_{kj}, \tag{10}$$

for all $s \in S$ such that $p(s) > 0$ and for all $j = 1, ..., p$; or in other words $var(\hat{X}_{HT}) = 0$. Where s_k is the indicator whether the unit is in the sample or not, i.e.:

$$s_k = \begin{cases} 1, & \text{if unit } k \text{ is in the sample} \\ 0, & \text{if unit } k \text{ is not in the sample} \end{cases}$$

and $S_n = \left\{ s \in S | \sum_{k \in U} s_k = n \right\}$.

2.4 The experiment

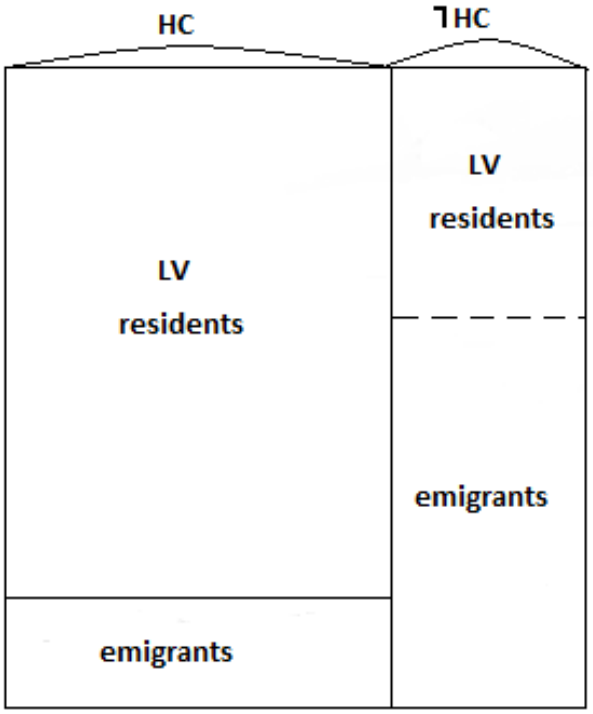
The aim was to select the sample which is balanced on the independent variables totals of unknown population part. The sample size were chosen directly proportional to unknown population part, i.e. $l = \frac{n_z}{N_n}$, and $N_n \leq n_z \leq N_z$, where N_z is the known population part size, N_n is unknown population part size and n_z is sample size. So the sample totals must be proportional to unknown population part totals, i.e.: $t_n = l * t_z$.

The obtained sample include only 5% emigrants and the result is the same as in previous experiments.

Conclusion

Through the experiments we have obtained the conditional probability $g_i = P(i \in LV | i \in HC)$, which means the probability, that person is the resident of Latvia, in condition that this person was participate in Housing Census 2011. To obtained the probability of unknown population part, we have to estimate the probability $p_i = P(i \in LV | i \notin HC)$. It will be the following step in this research.

Figure 1: Emigrants and residents of Latvia in Housing census 2011



References

Hosmer, D. & Lemeshov, S. (2000). *Applied logistic regression.* A Wiley Interscience Publication JOHN WILEY and SONS, ONC.

Tillé, Y. (2006). *Sampling algorithms.* Springer Sciens+Business Media, Inc.

CSB home page (Population Census): http://www.csb.gov.lv/en/statistikas-temas/population-census-30761.html