The number of Latvian residents estimate via logistic regression

Jeļena Vaļkovska
Central Statistical Bureau of Latvia, e-mail: Jelena.Valkovska@csb.gov.lv

Introduction

The main aim of this research is to investigate the logistic regression as the potential instrument for estimating the number of residents and 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.

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.

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

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:

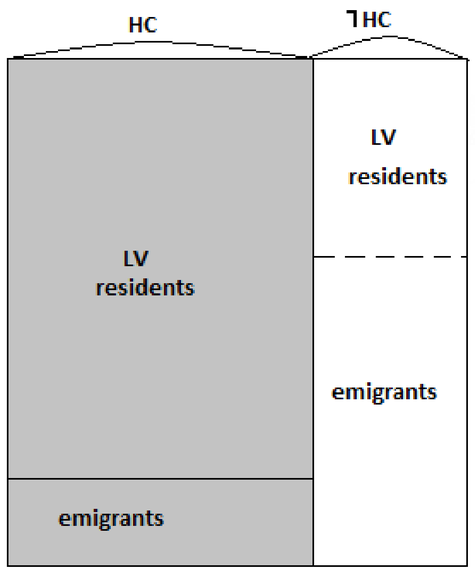
$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)^{(1-y_i)}\right]. \quad (2)$$

Population and Housing Census 2011

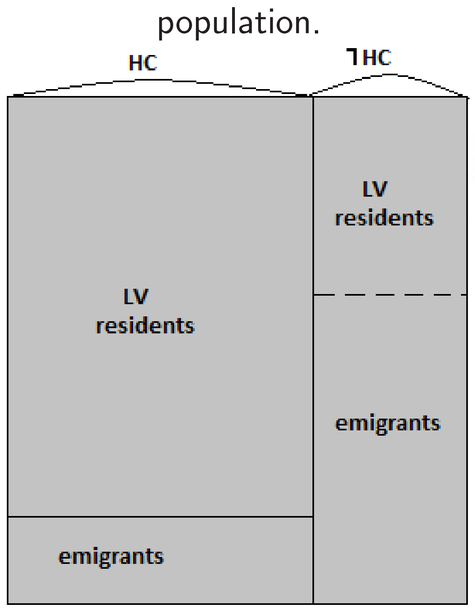
The Population and Housing Census was organised in March 2011. The main aim of it was to obtain detailed enough view on structure and characteristics of the population. There was obtained information about 1,94 mill. persons and counted up about 1,88 mill. residents in the Census. The information (is the person the resident or emigrant) about approximately 300 thousands persons is unknown. The number was estimated and according to the results of Census there were 2 070 371 residents of Latvia on 1 March 2011.

Models

First of all, the logistic regression model was used for known population part.



Beta coefficients were estimated, using logistic regression for the whole population.



Tab. 1: Latvian residents

	01.01.2012.
Real value (CSP)	2 041 763
Logit model	2 140 073

The obtained value is much greater than the real value.

Tab. 2: Latvian residents

	01.01.2012.
Real value (CSP)	2 041 763
Logit model	2 055 350

The obtained result is close to the real value.

Conclusion

The main problem may be that the number of emigrants in the investigated population part is too negligible (only **3%** of the population), and in the unknown population part this relation is different in accordance to estimated value. Using the logistic regression for the known population we have obtained the conditional probability $g_i = P(i \in LV | i \in HC)$. To obtain the probability of unknown population part, we have to estimate the probability $p_i = P(i \in LV | i \notin HC)$.

Model summaries

Tab. 5: Model summaries

Model	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1.	1063065,512	,050	,121
2.	868832,663	,129	,314
3.	866943,933	,130	,316

Auxiliary vectors

Tab. 4: Auxiliary information

	Auxiliary info	Transcription	beta		
			1st model	2nd model	3rd model
Sex	X_1	Sex	-,154	-,104	-,107
Marital statuss	X2_1	Single	-,662	-,488	-,484
	X2_2	Married		-,168	-,171
	X2_3	Divorced	-,659	-,631	-,631
Country of birth	X3	Latvia	,671	,732	,733
Nationality	X4_1	Latvians	1,039	,916	,912
	X4_2	Lithuanians	,768	,765	,767
	X4_3	Estonians	,437	,414	,410
	X4_4	Germans	-,597	-,707	-,705
	X4_5	Belarussians	1,015	,908	,908
	X4_6	Russians	,666	,658	,659
	X4_7	Ukrainians	,675	,638	,639
	X4_8	Poles	,661	,557	,556
	X4_9	Jews	-1,286	-1,447	-1,447
	X4_10	Roma	-,385	,112	,119
	X4_19	Unknown	,626	,588	,587
Citizenship	X5_1	LV citizens	,460	,279	,277
	X5_2	LV non-citizens	,356	,309	,307
	X5_19	no information	,454	,373	,367
Living region	X8_1	Riga	-,143	-,142	-,145
	X8_2	Latgale	,058	,050	,056
	X8_3	Zemgale	,021		
	X8_4	Vidzeme	,058		
	X8_5	Kurzeme	-,046	-,053	-,051
Age groups	X41_00	0- 4		-,154	-,156
	X41_01	4-9	-,348	-,512	-,514
	X41_02	10-14	-,229	-,403	-,406
	X41_2	15-19	-,610	-,839	-,845
	X41_3	20-24	-1,508	-2,245	-2,264
	X41_4	25-29	-1,786	-2,851	-2,873
	X41_5	30-34	-1,659	-2,722	-2,741
	X41_6	35-39	-1,302	-2,369	-2,385
	X41_7	40-44	-1,050	-2,062	-2,075
	X41_8	45-49	-,663	-1,529	-1,538
	X41_9	50-54	,315	,091	,087
	X41_10	55-59	,843	,823	,822
Employment	sum_sum	income		,000	,000
	cik_men	employment		1,173	1,086
	pnz_ien	self- employment		3,841	3,842
Employer's report code	ZK_1	code			1,453
	zaud_st	21.-25.			-,702
	apdr	11.			-,717
	atv	50., 51.			-1,313
Constant			2,168	2,292	2,295

Conclusions

- Using the logistic regression model for a known part of the population, we have obtained high probabilities that a person is Latvian resident. This result could be explained by the fact that the number of emigrants in the population is negligible.
- β coefficients were calculated, using the logistic regression model for the whole population. The estimated value is close to the true value.