Statistical Analysis of International Migration Using R Software

Dobre, Ana Maria

National Institute of Statistics, Romania

2013

Submission for the 2013 IAOS Prize for Young Statisticians

STATISTICAL ANALYSIS OF INTERNATIONAL MIGRATION USING R SOFTWARE

Ana Maria DOBRE
Expert
dobre.anamaria@hotmail.com
anamaria.dobre@insse.ro
National Institute of Statistics
Bucharest, Romania

Abstract

The aim of this paper is to expose the results of my research concerning the migrant's profile built up by means of logit regression model based on social and demographic characteristics. Within these characteristics could be mentioned: the age group, the gender, the education level, the marital status, the activity, the residence area.

The statistical software used is R which represents the most popular and powerful open source programming technology among statisticians during the last years. An application on logistic model and its performance is presented based on 2011 Labour Force Survey (LFS) referring to the international migration in Romania.

Keywords:
R statistical software; logistic regression; statistical analysis; odds ratio; migration

INTRODUCTION

The international migration is a demographic issue often hard to determinate and to estimate. According to Regulation (EC) No 862/2007 on Community Statistics on Migration and International Protection [17], 'emigration' means *the action by which a person, having previously been usually resident in the territory of a Member State, ceases to have his or her usual residence in that Member State for a period that is, or is expected to be, of at least 12 months.*

Every EU citizen has the right to live and work in any country within the Community. The free movement of persons in Europe creates difficulties in the measurement and outline of international migration in general and international emigration in particular.

This paper is an intention to carry out a simulation of outlining the migrant's profile using the logistic regression model.

LITERATURE REVIEW

The literature is wide as regards the international migration, but not regarding the profile of the migrant taking in consideration social and demographic characteristics.

There is an analysis of the migrant workers profile in correlation with labour force factors created in 2011 for Social Care Workforce Research Unit [9] where is compiled a complex analysis of the immigrants workers in England.

As for Romania, in 2007 was conducted a study named Study on Romanian Labour Force Migration in the European Union [18]. The study showed that the most people who migrated from Romania in that period were high-educated and employed, rather from urban area than rural one. Also in that study it is outlined a profile of the migrant by most selected destination country. Even so, the study was not made based on an official statistical survey with national representativeness, the sample having only 884 individuals.

METHOD USED FOR DATA ANALYSIS

The statistical method used for analyzing migrant's profile is the logistic regression model.

The logistic regression is modeling the relationship between some independent variables X_i and one dependent variable (dummy variable) Y. A dummy variable appears, ordinarily, when indicating the absence or presence of some categorical effect that may be expected to shift the outcome – e.g. presence/absence, yes/no.

The logit model is necessary in order to stand out the p likelihood from (0,1) interval to (- ∞, + ∞) interval, being necessary for estimating the parameters of regression equation (βi):

$$LOGIT(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 * X_1 + ... + \beta_k * X_k$$

Where β_0 represents the intercept and $\beta_1, \ldots, \beta_k$ represent the regressors of the model.

The model is directly linked with the concept of odd ratio:

$$OR \equiv \frac{p}{1-p}$$

p= event/success
1-p= non-event/failure

The odd ratio represents the ratio of the probability to ocurre an event to the probability of non-events.

These parameters ($\beta_1, \ldots, \beta_k$) give the variation of the odd ratio's logarithm at X_k's factor increasing with one unit.

CASE STUDY: PROFILE OF MIGRANT POPULATION IN ROMANIA

The R statistical software [16] provides every needed tool for analyzing the logistic regression model.

For the statistical analysis it was used R Commander [7] - a basic statistics GUI for R. It has incorporated the logit regression model within generalized linear model (glm). This function allows for several different types of models. The arguments of the function are: *formula*, *family* (the type of response variable used in the model) and *data* (the dataset used in the analysis). For logistic regression, *family* = binomial.

The sample used is Labour Force Survey 2011 [14]. The size of the survey sample is 112.320 dwellings annually [15]. This sample has a dimension of n=65535 individuals and it has relevance at national level (NUTS 0) and at regional level (NUTS 2). The survey covers all the members of the selected households including the persons absent from home for a longer period (over 12 months). Since LFS is not a statistical survey dedicated to international migration, but to collection of data on labour force, it does not fully compile the migration phenomenon.

Specifically, the LFS file is presented as follows:

Table 1. Labour Force Survey 2011

NO. CRT.	ABSENT	GENDER	AGE GROUP	RESIDENCE AREA	EDUCATION LEVEL	ACTIVITY	MARITAL STATUS
1	1	male	gr_1	urban	low	pupil/student	single
2	0	female	gr_2	urban	medium	unemployed	married
3	1	female	gr_3	rural	high	other situation	single
4	0	male	gr_4	urban	medium	employed	widow/divorced
5	0	male	gr_5	rural	medium	pensioner	married

The variable of interest for the model is *absent* explained as it follows: 1 - if people are left abroad for a period of 12 months and above, 0 – if otherwise. Therefore, in the present case, the categorical effect was represented by the migrant/non-migrant state of person.

The covariates used in the model are the following:

1. The *gender* variable includes two sets of values: male and female
2. The *age group* variable comprises five sets of values: gr_1 (aged 0-14), gr_2 (aged 15-24), gr_3 (aged 25-39), gr_4 (aged 40-64), gr_5 (aged 65 and over)
3. The *residence area* variable comprises two sets of values: urban and rural
4. The *education level* variable consists of three sets of values: low, medium and high
5. The *activity* variable comprises five sets of values: employed, unemployed, pupil/student, pensioner and other situations

6. The *marital status* variable consists of three sets of values: single, married, widow/divorced

Results of the study

The equation of the logistic model is the following:

$$Y_i = \beta_0 + \beta_1 * gender + \beta_2 * age\_group + \beta_3 * residence\_area + \beta_4 * education\_level + \beta_5 * activity + \beta_6 * marital\_status + \varepsilon_i$$

ε_i = random error term

Based on the model above, below it is presented the analysis of the correlation between the dependent variable and every covariate in order to build up the profile of the Romanian migrant.

Table 2. Results of the logistic regression model for migrants

Covariates of the model	*Odds Ratio*	*Confidence Interval*		*p-value*
		2,50%	*97,50%*	
Age (ref – gr_1)				
gr_2	8.31	5.35	13.74	< 2e-16 ***
gr_3	10.74	6.99	17.60	< 2e-16 ***
gr_4	3.88	2.52	6.38	8.45e-09 ***
gr_5	0.14	0.06	0.32	5.65e-06 ***
Gender (ref – female)				
male	1.71	1.37	2.14	1.97e-06 ***
Residence area (ref – rural)				
urban	0.70	0.61	0.79	1.02e-08***
Education level (ref – medium)				
low level	0.53	0.32	0.81	0.00635 **
high level	0.30	0.01	1.36	0.23895
Activity (ref - other situations)				
pupil/student	1	4.91e-08	5.38e+07	1.00
employed	3.81e+04	1.32e-02	NA	0.96
pensioner	1	1.06e-01	2.93e+08	1.00
unemployed	1	5.59e-02	7.11e+23	1.00
Marital status (ref – married)				
single	2.62e-05	3.05e-25	3.49e-02	0.944
widow/divorced	2.62e-05	NA	3.57e+73	0.990

The analysis resulted from R's output has 3 components presented below:

1. *The odds ratio* represents the exposure associated with the event of being a migrant.
2. *The 95% confidence interval* for the odds ratio; confidence intervals that do not contain one are significant
3. *The p-value* represents the chance that the obtained result would be achieved if the null hypothesis were true.

The reference group is the one generated by the model which has null regressors.

The results in the table above highlight the profile of the Romanian migrant at 1st January of 2012 using LFS 2011[1]. The model confirms that migrants are significantly 25-39 aged, more likely to be men and considerably from rural residence area. The analysis shows up that the migrants are more likely to be employees with medium education level.

The interpretation of the output is described in the following statement.

The reference age group is the 1st one (0-14 years old). The odds of a 15-24 aged person being a migrant are 8.31 times the odds of a the 1st age group person to migrate. The data available from the output show that the most people migrating are forming an integral part of 2nd and 3rd age group.

One of the strongest associations of the model was observed in relation to the gender variable. The people leaving from Romania are more likely to be men.

The reference group for the residence area variable is the rural one. The people from the rural area have more chances to leave abroad than the ones from urban residence area.

The output shows up that there are 53% odds for the people with low education to migrate towards the ones with medium education. The ones with high education migrate the least of all.

According to the analysis, the employees have the widest chance to be migrants. On the other hand, the pensioners, the unemployed and the students are considerably less likely to be migrants. This variable does not have statistical significance.

The probability of being a migrant among single, widow and divorced persons is insignificant towards to the married ones. This variable does not have statistical significance.

CONCLUSION

The data have a number of limitations because of the sample size and it consists in a lack of other dimensions like motivation to emigrate, push and pull factors, skills or future plans. However, the available data and the quantitative analysis presented in the paper address an existing knowledge gap related to the migrant's profile. A number of important findings from this paper are relevant to current phenomenon of emigration.

The exposed model is very important because it can be used to create not only the profile of migrant, but also the profile of other social or demographic category like: profile of unemployed, profile of employed, profile of pensioner, profile of immigrant etc.

The model can be extended also for long-time data series as well as for creating profile for long-term migrants and for short-term migrants or for analysing the brain drain phenomenon.

ACKNOWLEDGEMENT

The present paper is part of a research project of Romanian R-userRs Team[2]. I would like to express special thanks to Nicoleta Caragea[3] and Ciprian Antoniade Alexandru[4]. They provided me their support and guideline in this project. Also, I would

[1] At the moment (January 2013) is the most recent available LFS database for Romania
[2] http://www.r-project.ro/
[3] Senior Expert, National Institute of Statistics, Romania/Lecturer at Ecological University of Bucharest
[4] Dean, Faculty of Economics, Senior Lecturer at Ecological University of Bucharest

like to give special gratitude to Silvia Pisica[5], who expressed her sincere guidance for this project and provided the data sources for the analysis.

References

[1] Agresti, A. (2007), *An Introduction to Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley.

[2] Caragea, N., (2010), *Economic Statistics*, Ed. Mustang, ISBN 978-606-8058-37-5

[3] Caragea-Hrehorciuc, N., (2008), The work in Romania: An Analysis of the Macroeconomic Perspective, *Revista Economica/ RePEc, DOAJ/EconPapers,* Available at: http://econpapers.repec.org/article/blgreveco/v_3a42-43_3ay_3a2008_3ai_3a5-6_3ap_3a117-126.htm, ISSN 1582-6260

[4] Caragea, N., Alexandru, A.C., Dobre, A.M. (2012), "Bringing New Opportunities to Develop Statistical Software and Data Analysis Tools in Romania", *The Proceedings of the VIth International Conference on Globalization and Higher Education in Economics and Business Administration*, ISBN: 978-973-703-766-4, pp.450-456.

[5] Castles, S., Miller, J. M. (2003), *The Age of Migration: International population movements in the modern world* (3rd edition). Palgrave: Macmillan

[6] Christensen, Ronald (1997). *Log-linear models and logistic regression*. Springer Texts in Statistics (Second ed.). New York: Springer-Verlag. pp. xvi+483. ISBN 0-387-98247-7. MR 1633357.

[7] Fox, J. (2005). The R Commander: A Basic Statistics Graphical User Interface to R. *Journal of Statistical Software*, 14(9): 1--42

[8] Fox, J. (2007) Extending the R Commander by "plug in" packages. *R News*, 7(3): 46–52.

[9] Hussein, S. (2011) Migrant Workers in Long Term Care: Evidence from England on Trends, Pay and Profile, in *Social care Workforce Periodical*, Issue 12

[10] Long, J. Scott (1997). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage Publications.

[11] Vasile, V (2011), "Youths on Labour Market. Features. Particularities. Pro-mobility Factors for Graduates. Elements of a Balanced Policy for Labour Migration", *Romanian Journal of Economics*, 32, (1(41)), 97-123

[5] General Director, General Department of Social Statistics and Demography, National Institute of Statistics, Romania

[12] Vasile, V. (2009) “International economic integration and migration: The case of Romania” Chapter 11, in *Migration and Human Capital*, Ed. Jacques Poot, Brigitte Waldorf, Leo van Wissen, Edward Elgar Publishing House**,** 2008, published in 2009 pp.225-247; UK, ISBN 978-1-84720-084-6.

[13] EUROSTAT (2011) - „The Eurostat Database on International Migration”

[14] National Institute of Statistics, Romania (2011) – “Labour Force Survey”

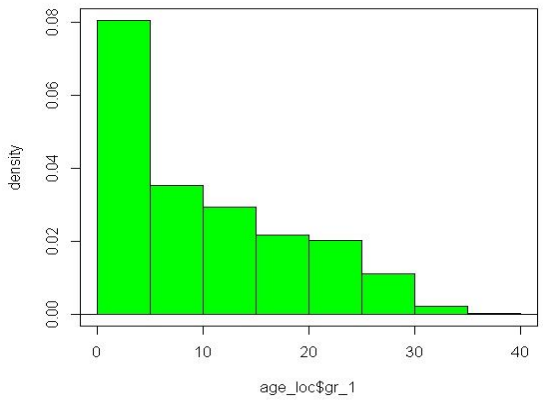
[15] *Quality Report On 2011 Household Labour Force Survey,* Available at: http://www.insse.ro/cms/files/Rapoarte%20de%20calitate/Amigo/RO_LFS%20Quality%20Report_2011.pdf

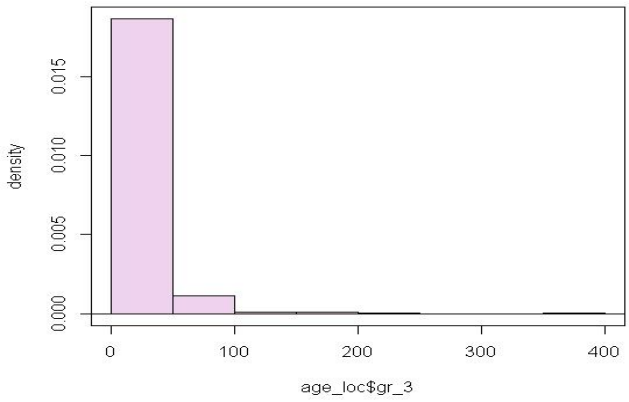
[16] R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

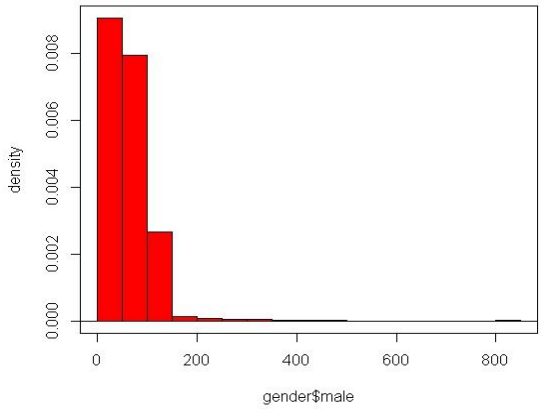
[17] *Regulation (EC) No 862/2007 of the European Parliament and of the Council*, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:199:0023:0029:EN:PDF

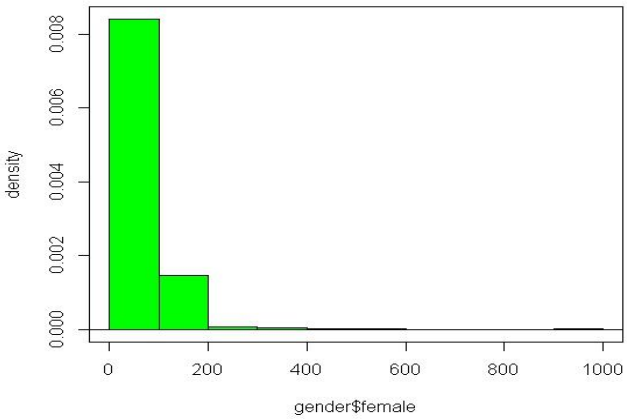
[18] *Study on Romanian Labour Force Migration in the European Union*, available at: http://www.robcc.ro/ro/_Cu_bine_din_Europa_Studiu_despre_migratia_fortei_de_munca_romanesti_in_Uniunea_Europeana-146

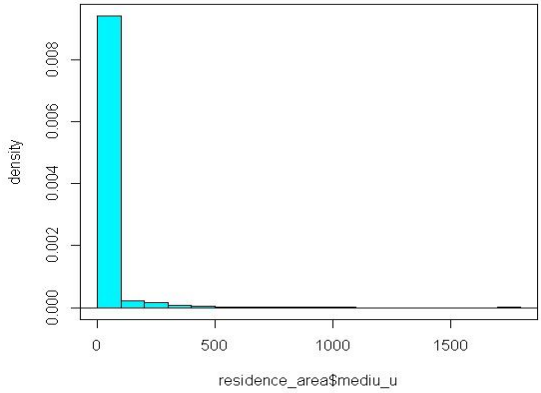
APPENDIX
Some examples of the distributions of the analyzed variables

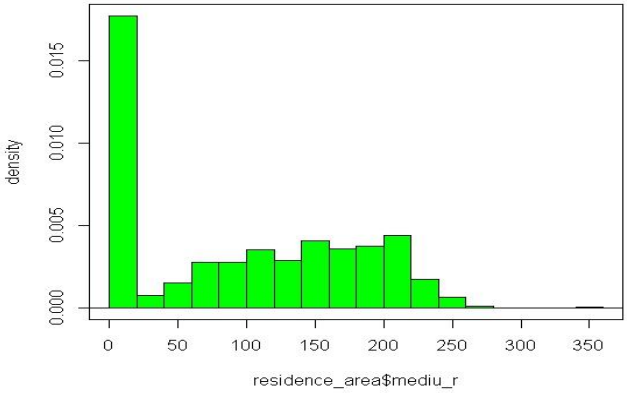


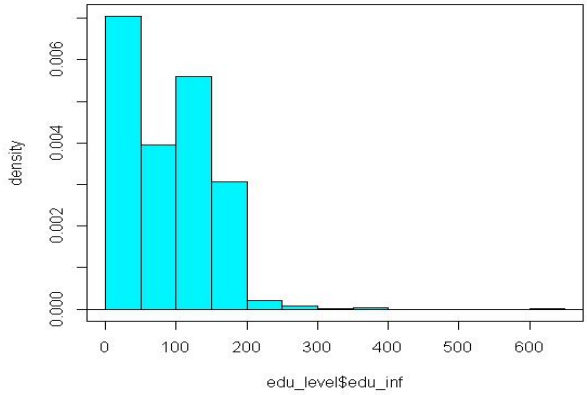


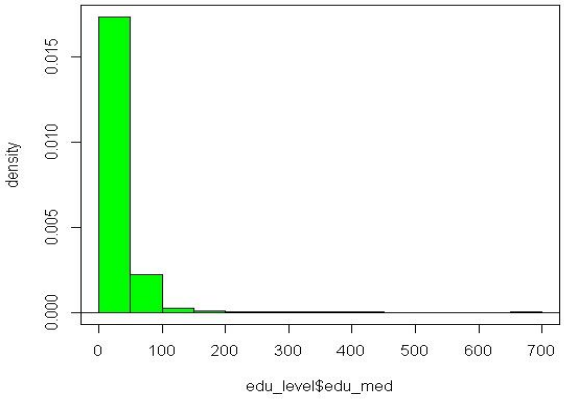

density
0.000
0.002
0.004
0.006
0.008
0
200
400
600
800
gender$male

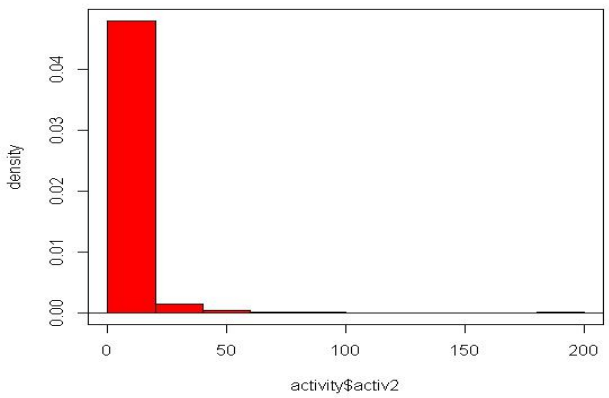

density
0.000
0.002
0.004
0.006
0.008
0
200
400
600
800
1000
gender$female

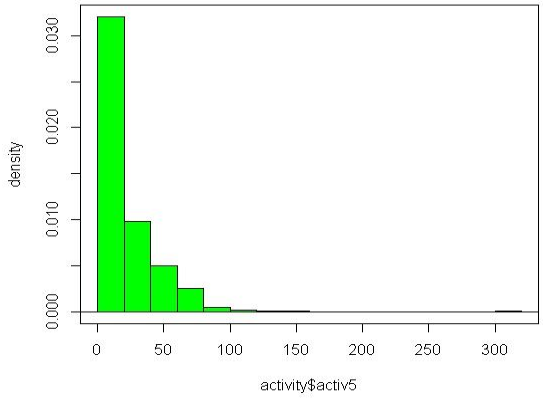

density
0.000
0.002
0.004
0.006
0.008
0
500
1000
1500
residence_area$mediu_u


density
0.000
0.005
0.010
0.015
0
50
100
150
200
250
300
350
residence_area$mediu_r


density
0.000
0.002
0.004
0.006
0
100
200
300
400
500
600
edu_level$edu_inf


density
0.000
0.005
0.010
0.015
0
100
200
300
400
500
600
700
edu_level$edu_med


density
0.00
0.01
0.02
0.03
0.04
0
50
100
150
200
activity$activ2


density
0.000
0.010
0.020
0.030
0
50
100
150
200
250
300
activity$activ5