Statistical Analysis of International Migration Using R Software

Ana Maria Dobre

National Institute of Statistics, Romania

2013

Online at http://mpra.ub.uni-muenchen.de/48804/
MPRA Paper No. 48804, posted 2. August 2013 12:44 UTC
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.


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 (- $\infty$, + $\infty$) interval, being necessary for estimating the parameters of regression equation ($\beta$):

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

Where $\beta_0$ represents the intercept and $\beta_1, ..., \beta_k$ represent the regressors of the model.

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

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

$p$= event/success

$1-p$= non-event/failure
The odd ratio represents the ratio of the probability to occur 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**

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

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:

\[ Y_i = \beta_0 + \beta_1 \text{gender} + \beta_2 \text{age_group} + \beta_3 \text{residence_area} + \beta_4 \text{education_level} + \beta_5 \text{activity} + \beta_6 \text{marital_status} + \epsilon_i \]

\( \epsilon_i \) = 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**

<table>
<thead>
<tr>
<th>Covariates of the model</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>p-value</th>
<th>2.50%</th>
<th>97.50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (ref – gr_1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gr_2</td>
<td>8.31</td>
<td>5.35</td>
<td>13.74</td>
<td>&lt; 2e-16 ***</td>
<td></td>
</tr>
<tr>
<td>gr_3</td>
<td>10.74</td>
<td>6.99</td>
<td>17.60</td>
<td>&lt; 2e-16 ***</td>
<td></td>
</tr>
<tr>
<td>gr_4</td>
<td>3.88</td>
<td>2.52</td>
<td>6.38</td>
<td>8.45e-09 ***</td>
<td></td>
</tr>
<tr>
<td>gr_5</td>
<td>0.14</td>
<td>0.06</td>
<td>0.32</td>
<td>5.65e-06 ***</td>
<td></td>
</tr>
<tr>
<td>Gender (ref – female)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>1.71</td>
<td>1.37</td>
<td>2.14</td>
<td>1.97e-06 ***</td>
<td></td>
</tr>
<tr>
<td>Residence area (ref – rural)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>urban</td>
<td>0.70</td>
<td>0.61</td>
<td>0.79</td>
<td>1.02e-08 ***</td>
<td></td>
</tr>
<tr>
<td>Education level (ref – medium)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low level</td>
<td>0.53</td>
<td>0.32</td>
<td>0.81</td>
<td>0.00635 **</td>
<td></td>
</tr>
<tr>
<td>high level</td>
<td>0.30</td>
<td>0.01</td>
<td>1.36</td>
<td>0.23895</td>
<td></td>
</tr>
<tr>
<td>Activity (ref - other situations)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pupil/student</td>
<td>1</td>
<td>4.91e-08</td>
<td>5.38e+07</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>employed</td>
<td>3.81e+04</td>
<td>1.32e-02</td>
<td>NA</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>pensioner</td>
<td>1</td>
<td>1.06e-01</td>
<td>2.93e+08</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>unemployed</td>
<td>1</td>
<td>5.59e-02</td>
<td>7.11e+23</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Marital status (ref – married)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>single</td>
<td>2.62e-05</td>
<td>3.05e-25</td>
<td>3.49e-02</td>
<td>0.944</td>
<td></td>
</tr>
<tr>
<td>widow/divorced</td>
<td>2.62e-05</td>
<td>NA</td>
<td>3.57e+73</td>
<td>0.990</td>
<td></td>
</tr>
</tbody>
</table>

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 2\(^{nd}\) and 3\(^{rd}\) 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\textsuperscript{5}, who expressed her sincere guidance for this project and provided the data sources for the analysis.

References


\textsuperscript{5} General Director, General Department of Social Statistics and Demography, National Institute of Statistics, Romania
APPENDIX
Some examples of the distributions of the analyzed variables