Modelling the potential human capital on the labor market using logistic regression in R

Ana-Maria Ciuhu (dobre.anamaria@hotmail.com)
Institute of National Economy, Romanian Academy; National Institute of Statistics

Nicoleta Caragea (nicoleta.caragea@insse.ro)
National Institute of Statistics; Ecological University of Bucharest

Ciprian Alexandru (alexcipro@yahoo.com)
National Institute of Statistics; Ecological University of Bucharest

Motto:

"If you wanted to do research in statistics in the mid-twentieth century,
you had to be bit of a mathematician, whether you wanted to or not . . .
If you want to do statistical research at the turn of the twenty-first century,
you have to be a computer programmer."
Andrew Gelman, Department of Statistics,Columbia University

ABSTRACT

This paper exposes the methodology of creating the profile of two categories of potential human capital using logistic regression in R. The profiles were created based on some social and economic characteristics provided by the 2015 Labour Force Survey, assuring the representativeness of results at national and regional level. In this sense, the logistic regression was used to model the relationship between economically inactive persons who are seeking for a job, but are not immediately available to start working, respectively economically inactive persons who are not seeking for a job, but are immediately available to start working, and some socio-economic predictors. The aim is to identify the impediments which determine inactive people not to become active on the labour market.

Keywords: *R statistical software, labor force, logistic regression, odds ratio*
JEL Classification: *J21, C50, C87*

1. INTRODUCTION

Taking part of the labor force is very important because not only the individual's life depends upon it but also it participates in the economic development of the country. The key issue to be discussed in this study is to analyze, through statistical tools, potential employment in Romania based on socio-economic characteristics of the population. The economic problem

could be regarded as a risk analysis while an individual is economically inactive people, being a chance to take part of the labor force. As a result of the binomial logit model, the most significant factor to consider here is that each one tells the effect of the predictors of risk on the probability of success in that category, in comparison to the reference category. These kind of econometric models have a different approach comparing with the parametric models, being part of the class of generalized linear model - GLM. These models have been formulated for the first time by John Nelder and Robert Wedderburn (1972).

Logistic regression is a probabilistic model of statistical analysis between two or more processes, based on certain characteristics, the result being a categorical variable. The main issue of a logit model is to predict the likelihood of dependent variable to register one of the possible response categories. The estimation of the parameters of regression equation is based on MLE (maximum likelihood estimation). This method involves:

- finding the coefficients (βk) that makes the log of the likelihood function ($LL < 0$) as large as possible (maximize the probability that event to occur);
- or, finds the coefficients that make -2 times the log of the likelihood function (-2LL) as small as possible.

There are situations where the dependent variable can record two or more categories of response; if there are two categories, the variable is binary or dichotomous (for example, the sex variable can record two values: male and female) - in this case the binomial logistic regression is applied; if there are multiple response categories of the resulting variable, the multinomial logistic regression applies (for example, the education level variable can record multiple categories: low, medium, or high).

The aim of the present analysis is to identify the impediments which determine inactive people not to become active on the labour market. Certain kind of factors are being considered, such as gender, age, education level, marital status, residence area, household's structure, economic sector of the last employer and reason to decline a job offer.

Similar logit models have been realized in the literature, of which for modelling the long-term unemployment (Obben J. et al, 2002), the probability of becoming employed (Luckanicova M. et al, 2012) and the profile of international migrants (Caragea N. et al, 2013).

2. METHOD DESCRIPTION

2.1. Fitting a binary logistic regression model

Binomial logistic regression - model the relation between a set of independent variables x_i (categorical, continuous) and a dichotomous (nominal, binary) dependent variable y. The multiple logistic regression model is given by the following equations:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \quad (1)$$

or

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \quad (2)$$

The multiple regression model could also be represented as:

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k} \quad (3)$$

Or, as odds ratios:

$$\Omega = \frac{p}{1-p} \quad (4)$$

The model could be expresed as:

$$\Omega = e^{\beta_0+\sum_k \beta_k x_k} \quad (5)$$

Where:

p = the probability of y to be equal to 1 (success);

1-p = the probability of y to be equal to 0 (non-success);

$\beta_0, \beta_{1, ..., } \beta_k$ = parameters of regression equation;

k = number of observations.

Transformation of logit into probabilities is represented below:

$$p = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}} \quad (6)$$

The odds ratio compares the chances of two population groups characterized by different values recorded by the independent variable (x_j), while all other predictors remain constant ($x_i = \text{const}, i \neq j$). It could be expressed by the following formula:

$$OR = \frac{\Omega_{(x_j+1)}}{\Omega_{(x_j)}} = \frac{e^{\beta_0+\beta_j(x_j+1)}}{e^{\beta_0+\beta_j x_j}} = \frac{e^{\beta_0} \times e^{\beta_j x_j} \times e^{\beta_j}}{e^{\beta_0} \times e^{\beta_j x_j}} = e^{\beta_j} \quad (7)$$

Therefore, e^{β_j} represents the odds ratio that shows what happens when x_j changes with one unit, and the other predictors do not have any influence on the change of the dependent variable.

2.2. Modeling the potential human capital on the labor market using a binary logistic regression model

The regression function is modelling the potential category of human capital that could be on labour force, but is not yet. Suppose we are interested in estimating the proportion of inactive persons in a population, having potential of being employed. Naturally, we know that entire population does not have equal probability of 'success' (i.e. being employed). Lower educated people are more likely to be inactive, even they are included in working-age population. Consider the predictor variable X (education level) to be any of the success/risk factor that might contribute to the economically inactive status of a person. Probability of success (to be employed) will depend on the levels of the success/risk factors (level of education).

In the presented study, there were used two sub-population consisting in two types of economically inactive persons: economically inactive persons who are seeking for a job, but are not immediately available to start working, respectively economically inactive persons who are not seeking for a job, but are immediately available to start working (within 2 weeks). According to LFS methodology, the potential additional labour force represents the sum of the two categories mentioned above.

Data sources and software used

Data used are based on the Romanian Labour Force Survey sample 2015, conducted by the National Institute of Statistics. Data were collected quarterly, in order to capture the effects of seasonal variations. Conceived as important source of intercensus information on labour force, the survey provides, in a coherent manner, essential data about all the population segments, with several possibilities of correlation and structuring by various demographic, social and economic characteristics, under the conditions of international comparability (Pisica S, 2015).

For computing the logistic regression models was used the glm function from the stats package in R. The model summary output includes the coefficients, standard errors, z values, p-values, the null and residual deviances, the Akaike Criterion and the number of Fisher Scoring iterations.

Description of the variables

Dependent variable is the potential of labour force, in terms of a categorical variable with 2 groups as follows:

Y1 = economically inactive persons who are seeking for a job, but are not immediately available to start working (within 2 weeks). These are persons aged 15-74 years, neither employed nor in unemployment, who looked for a job, during the 4 weeks previous to the interview, but are not available to start work in the next 2 weeks.

Y2 = economically inactive persons who are not seeking for a job, but are immediately available to start working (within 2 weeks). These are persons aged 15-74 years, neither employed nor in unemployment (economically inactive persons), who wish to work, and are available to start working in the next 2 weeks, but did not look for a job during the 4 weeks previous to the interview.

Independent variables (predictors) are the following:

- *Gender* – is a dummy variable for gender [Gender (Male = 1, Female = 2)];
- *Age (Age Groups)* - Age variable was available in LFS 2015 as a continuous variable that was further converted into a categorical variable with different groups showing five different stages of life [Age Group (1=15 to 24, 2=25 to 34, 3=35 to 44, 4=45 to 54, 5=55 years or more)];
- *Residence area* – is a dummy variable for residence area [Residence area (Urban = 1, Rural = 3)];
- *Education* - a categorical variable of education with 3 categories [Education (1=low education, 2=medium education, 3=superior education)];
- *Marital status* – a categorical variable for marital status with 4 categories [Marital status (1=single, 2=married, 3=widowed, 4=divorced)];
- *Number of persons in the household* – a continuous variable with values from 1 to 9;
- *Economic sector* – a categorical variable for economic sector of the last employer [Activity (B=industry, C=services). In the database there are only registrations for industry and services, excluding agriculture sector;
- *Reason* – a categorical variable for the reason to decline a job offer with 3 categories [Reason (1=distance e.g. changing the usual residence, long distance to home and shuttling, 2=qualification e.g. underqualification and requalification, 3=lower earnings)].

The models could be represented in the equation below:

$$
\begin{aligned}
&ln\left(\frac{P(Y=1)}{P(Y=0)}\right)=\beta_{inactive0}+\\
&+\beta_{inactive,female}\times(gender=2)+\\
&+\beta_{inactive,urban}\times(resid=1)+\\
&+\beta_{inactive,25-34}\times(age=2)+\beta_{inactive,35-44}\times(age=3)+\beta_{inactive,45-54}\times(age=4)+\beta_{inactive,55+}\times(age=5)+\\
&+\beta_{inactive,low}\times(edu=1)+\beta_{inactive,lhigh}\times(edu=2)+\\
&+\beta_{inactive,married}\times(marital=2)+\beta_{inactive,widowed}\times(marital=3)+\beta_{inactive,divorced}\times(marital=4)+\\
&+\beta_{inactive,2\,pers}\times(pers=2)+\beta_{inactive,3\,pers}\times(pers=3)+\beta_{inactive,4\,pers}\times(pers=4)+\\
&+\beta_{inactive,5\,pers}\times(pers=5)+\beta_{inactive,6\,pers}\times(pers=6)+\beta_{inactive,7\,pers}\times(pers=7)+\\
&+\beta_{inactive,8\,pers}\times(pers=8)+\beta_{inactive,9\,pers}\times(pers=9)+\\
&+\beta_{inactive,services}\times(sec\,tor=C)+\\
&+\beta_{inactive,qualification}\times(reason=2)+\beta_{inactive,lowerearnings}\times(reason=3)
\end{aligned}
\tag{8}
$$

2.3. Empirical Results

Model 1

The first model (for dependent variable y1= economically inactive persons who are seeking for a job, but are not immediately available to start working) was computed, but it did not accomplish the expected empirical results. The output of the logit in R showed up a perfect convergence between the dependent variables and the predictors (very low odds ratios and p-values very close to 1, meaning that the coefficients are not statistically significant). Therefore, the model explaining the probability of economically inactive persons who are seeking for a job, but are not immediately available to start working to enter on the labour market depending on socio-economic characteristics is not statistically valid.

Results of the Logistic Regression Model for Inactives (y1), reference year 2015

Table 1

Covariates of the model	*Odds Ratio*	*Confidence Interval*		
		Lower 95%	*Upper 95%*	*p-value*
Age (ref – gr 1)				
gr_2 (25-34 years old)	9.42e-09	NA	1.91e+199	0.9959
gr_3 (35-44 years old)	9.42e-09	NA	2.65e+171	0.9952
gr_4 (45-54 years old)	9.42e-09	NA	9.02e+172	0.9952
gr_5 (55 years or more)	9.79e-02	4.84e-03	1.41	0.0442 *
Gender (ref – male)				
female	3.34e+07	1.80e-100	NA	0.992
Residence area (ref – rural)				
urban	3.06e-01	1.51e-02	2.39e+00	0.305
Education level (ref – medium)				
low level	4.10e-07	NA	9.62e+79	0.992
high level	1.16e+00	1.39e-01	9.66e+00	0.883
Marital status (ref - single)				
married	1.35e-01	6.68e-03	1.05e+00	0.0829 .
widowed	4.95e-08	NA	2.19e+118	0.9937
divorced	4.95e-08	NA	7.41e+197	0.9961
No. of persons in the household (ref - 1)				
2 persons	1.47e+07	0.00e+00	NA	0.998
3 persons	7.93e+06	0.00e+00	NA	0.998
4 persons	5.30e+06	0.000e+00	NA	0.998
5 persons	9.99e-01	1.35e-194	7.36e+193	1.000
6 persons	9.99e-01	1.178e-200	8.48e+199	1.000
7 persons	4.28e+07	0.00e+00	NA	0.998
8 persons	9.998e-01	1.43e-256	6.96e+255	1.000
9 persons	9.99e-01	2.39e-254	4.18e+253	1.000
Economic sector (ref - industry)				
services	1.000000e+00	NA	NA	1.000
Reason (ref - distance)				
qualification	9.09	8.47	9.73	0.992
lower earnings	12.10	10.29	14.15	0.997

Source: R output on logistic regression

The unavailability to start work within 2 weeks is not influenced by factors included in the analysis. Hence, regardless of age, gender, education level of the persons, the dependent variable does not change.

Model 2

The results for the second model (for dependent variable y2= economically inactive persons who are not seeking for a job, but are immediately available to start working), consisting in computed odds ratios, confidence intervals and p-values are exposed in the Table 2. The reference group is the group with null regressors generated by the model.

Results of the Logistic Regression Model for Inactives (y2), reference year 2015

Table 2

Covariates of the model	*Odds Ratio*	*Confidence Interval*		
		Lower 95%	*Upper 95%*	*p-value*
Age (ref – gr_1)				
gr_2 (25-34 years old)	1.03	0.92	1.16	0.567
gr_3 (35-44 years old)	0.96	0.86	1.07	0.421
gr_4 (45-54 years old)	0.78	0.69	0.87	1.30e-05 ***
gr_5 (55 years or more)	1.28	1.17	1.41	5.32e-08 ***
Gender (ref – male)				
female	2.06	0.66	0.78	< 2e-16 ***
Residence area (ref – rural)				
urban	0.27	0.26	0.29	< 2e-16 ***
Education level (ref – medium)				
low level	3.06	2.87	3.26	< 2e-16 ***
high level	0.47	0.44	0.51	<2e-16 ***
Marital status (ref - single)				
married	1.50	1.07	1.24	0.000126 ***
widowed	1.96	1.80	2.14	< 2e-16 ***
divorced	0.68	0.57	0.81	1.95e-05 ***
No. of persons in the household (ref – 1 person)				
2 persons	0.77	0.63	0 94	0.01010
3 persons	0.63	0.52	0.76	2.81e-06 ***
4 persons	0.59	0.49	0.72	1.11e-07 ***
5 persons	0.63	0.52	0.77	3.76e-06 ***
6 persons	0.66	0.54	0.81	7.17e-05 ***
7 persons	0.79	0.64	1.00	0.04987 *
8 persons	0.84	0.64	1.00	0.19709
9 persons	0.69	0.53	0.91	0.00841 **
Economic sector (ref - industry)				
services	0.55	0.47	0.65	8.78e-13 ***
Reason (ref - distance)				
qualification	9.09	8.47	9.73	<2e-16 ***
lower earnings	12.10	10.29	14.15	<2e-16 ***

Source: R output on logistic regression

In 2015 the economically inactive persons who are not seeking for a job, but are immediately available to start working in Romania are mostly persons aged 55 years or more. The most important reason for their status is they are before or in the age of retiring. There are two situations. For those inactive people in the 55-65 age groups, it is hard to find a job, being discouraged because of age. On the other side, the people aged 65+, the situation is different, they are not seeking for a job, because are pensioners. The odds ratio confirms that the probability to be an economically inactive person who are not seeking for a job, but is immediately available to start working is 1.28 times higher than for the reference age group (15-24 years). Moreover, also the young people aged 25-34 are likely to be in the same situation, with an odds ratio of 1.03.

In terms of gender, female population has higher risk (2.06 times more) to be inactive seeking a job and immediately available to start working, rather than men.

The majority of inactive people who are not seeking for a job, but are immediately available to start working are from the rural area.

Education level is another factor which determines inactive people not to participate on the labour market. The probability to be inactive person with low education is 3.06 times higher than to be inactive with medium education. More educated people (high level education) have lower risk to be inactive.

Regarding the marital status, most of inactive people are widowed or married. The second status (married) should be a warning for the economic status of the households. Married people who do not seek for a job but are immediately available to start working could sustain a lack of livelihood of the families, which have a direct impact on poverty.

The majority of inactive people come from households with one member. Starting with 3 persons, one-unit increase of family size reduces the risk of a person to be out of the labour market.

People which have worked before in industry sector have higher chance to be inactive than the people who have worked in the services sector.

Economically inactive persons who are not seeking for a job, but are immediately available to start working could have some reasons to decline job offers. The empirical results show that:

- the probability of a person to decline a job offer because of lower earnings is 12.1 times higher than the reason of distances (changing the usual residence, long distance to home and shuttling).

- the probability of a person to decline a job offer because of changing qualification (under-qualification or requalification) is 9.09 times higher than the distance from home to job location.

2.4. Fit of model

A logistic regression is said to provide a better fit to the data used in the analysis if it demonstrates an improvement over a model with fewer predictors. This is performed using the likelihood ratio test, which compares the likelihood of the data under the full model against the likelihood of the data under a model with fewer predictors. Removing predictor variables from a model will almost always make the model fit less well (i.e. a model will have a lower log likelihood), but it is necessary to test whether the observed difference in model fit is statistically significant. Given that H0 holds that the reduced model is true, a p-value for the overall model fit statistic that is less than 0.05 would compel us to reject the null hypothesis. It would provide

evidence against the reduced model in favour of the current model. The likelihood ratio test can be performed in R using the anova() function in base installation:

```
> anova(mylogit,test="Chisq")
```

The values of the chi squared tests were exposed in the Table 3.

Results of ANOVA (chi squared) for Model 2

Table 3

Covariates of the model	*ANOVA(chi squared)*
Age	< 2.2e-16 ***
Gender	< 2.2e-16 ***
Residence area	< 2.2e-16 ***
Education	< 2.2e-16 ***
Marital status	< 2.2e-16 ***
No. of persons in the household	6.897e-11 ***
Economic sector	4.494e-13 ***
Reason	< 2.2e-16 ***

Source: R output on logistic regression

These values indicate that every predictor improves the model.

3. CONCLUSIONS

In this paper a logit model of inactive people in Romania in the year 2015 was estimated. Age, gender, residence area, education level, marital status, number of persons in the household, economic sector of the last employer and the reason to decline a job offer were proven to have significant impact on the employability of inactive people in the model 2. The concept of the paper started from the idea that both type of economically inactive people (economically inactive persons who are seeking for a job, but are not immediately available to start working – model 1, respectively economically inactive persons who are not seeking for a job, but are immediately available to start working – model 2) are expected to be influenced by these predictors. Nevertheless, the practice has demonstrated that only the model 2 is statistically significant. Hence, regardless of age, gender, education level of the persons, the dependent variable does not change for model 1. The unavailability to start work within 2 weeks is not influenced by factors included in the analysis.

In model 2, all the predictors included in analysis represent more or less impediments which determine inactive people who are not seeking for a job, but are immediately available to start working not to become active on the labour market.

The study revealed some facts on employability of inactive people, as follows:

- Persons are more willing to change their residence than to work on lower wages or to be re-qualified or under-qualified;
- The inactive persons who are not seeking for a job, but are immediately available to start working are mostly older persons (aged 55 years or more) or people aged 25-34;
- The gender represents also an impediment, which may be a result of the discrimination on the labour market. The females have double chance more than men to be inactive seeking a job and immediately available to start working;
- The residence area is, as it has been expected, an important drawback for employability of inactive people. The majority of inactive people who are not seeking for a job, but are immediately available to start working live in the rural area;
- Rather people with low education and widowed or married could be inactive people who are not seeking for a job, but are immediately available to start working;
- Regarding the household structure, one-unit increase of family size does not affect much the risk to be out of the labour market;
- People which have worked before in industry sector have higher chance to be inactive than the people who have worked in the services sector. The capability of services sector to absorb labour force is obvious.

Taking into account the results of the analysis, could be noticed that the national social policies on employment should be revised. In order to attract on the labour market the inactives available to start work, employment measures should be reformulated, especially for those aged close to retirement and those in rural areas. The results of the paper show the general characteristics of an emerging economy, as is the case with Romania.

References

1. **Caragea N., Dobre A.M, Alexandru C.**, 2013, Profile of Migrants in Romania – A Statistical Analysis Using "R", Working papers No. 4 from Ecological University of Bucharest, Department of Economics
2. **Hosmer D., Lemeshow S., Sturdivant R.**, 2013, Applied Logistic Regression, Third Edition, Wiley, ISBN 978-0-470-58247-3
3. **Luckanicova M., Ondrusekova I., Resovsky M.**, 2012, Employment modelling in Slovakia: Comparing Logit models in 2005 and 2009, Economic Annals, Volume LVII, No. 192 / January – March 2012, ISSN: 0013-3264
4. **Nelder J. A., Wedderburn R. W. M.**, 1972, Generalized Linear Models, Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3 (1972), pp.370-384, available at: https://docs.ufpr.br/~taconeli/CE225/Artigo.pdf

5. **Obben J., Hans-Jürgen Engelbrecht H.-J., Thompson V.W.**, 2002, A logit model of the incidence of long-term unemployment, Applied Economics Letters, Vol. 9, No. 1, January 2002, pp. 43-46
6. **Pisică S.** (coord.), "Labour Force in Romania – Employment and unemployment" (annual publication), National Institute of Statistics, 2005 – 2015, ISSN 1842-3671
7. **R Core Team**, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/