Usage of the Internet for looking for a job in European Countries: Multiple regression and cluster analysis (Utilisation d'Internet pour la recherche d'un emploi dans les pays Européens: De régression multiple et l'analyse de cluster)

Ksenija Dumicic
Faculty of Economics and Business, University of Zagreb, CROATIA
Trg J. F. Kennedy 6, Zagreb (HR-10000), CROATIA
E-mail: kdumicic@efzg.hr

Introduction

The purpose of this paper is to investigate whether and how six selected variables influence the dependent variable under study defined as Y- Percentage of individuals aged 16 to 74 using the Internet for looking for a job or sending a job application. The aim of the study is to contribute the understanding of how two groups of independent variables impact the Y variable based on linear regression models: the *first group* being comprised of two ICT related variables (X1-Level of Internet access in households as Percentage of households who have Internet access at home; and X5- Individuals' level of computer skills as Percentage of the total number of individuals aged 16 to 74, who have carried out 1 or 2 of 6 computer related items), and the *second group* of independent variables being comprised of four economic development variables (X2- GDP per capita in PPS (EU-27 = 100); X3- Total unemployment rate; X4- Public expenditure on education as percent of GDP); and X6 - A dummy variable for a country being in transition or not.

The research hypothesis is that at least one variable from each of two groups of variables is statistically significant for explanation of variability in variable Y. Data used are cross sectional EUROSTAT data for EU-27 and Turkey, Iceland and Croatia, based on average for period 2001-2008.

For data exploration, both descriptive and cluster analysis using *Minitab 14* and *Megastat*, and afterwards methods of multiple linear regression using *EViews 7* were applied. After examining several models, the linear regression model with two regressors, X1 and X2, is found to be the most appropriate. This model is tested for diagnostics, and no model assumptions are violated.

A few papers treating variable Internet search for a job by regression and multivariate methods are already published. These papers consider mostly USA and Asian countries, but in a slightly different way than study presented in this paper, compare to Kuhn and Skuterud (2000), Suvankulov (2010), Fountain (2003), Brenčič and Norris (2010) Hogler, Henle and Bemus (1998), and Tso, Yau and Cheung (2010). But, Jackson (1998), and Kinder (2000) investigated the phenomenon for European countries, especially UK and Scottland.

I this paper after data definition and exploration, multiple regression analysis results are given and discussed.

Data

Based on data exploration, different shapes of distributions for each of variables were recognised. Data for variable Y, defined as Percentage of individuals aged 16 to 74 using the Internet for looking for a job or sending a job application, vary over 30 countries a lot with a mean of 10,3%, standard deviation of 5,88%, and coefficient of variation of 57,12%, see *Table 1*. The distribution of Y is more flat than the normal distribution (kurtosis<0) and it is slightly positively skewed (skewness=0,72), which shows that, on average for the period 2004 to 2008, there were a few (Scandinavian) countries with higher percentage of people who were using the Internet for looking for a job or sending a job application by Internet (Finland with 24,8, Denmark 20,8% and Sweden 20,6). Turkey with the value of 2,3%, and Romania with 2,5% are the opposite extremes. All the extreme values are within 3 standard deviations around the mean, being only suspicious outlying values, and not serious ones.

Table 1 Summary descriptive statistics for variables Y, X1 and X2

	Y_INTJOB0408	X1_INTER0408	X2_GDPAV0408
Mean	10,30	47,40	95,42
Median	9,05	45,20	92,40
Maximum	24,80	84,00	266,40
Minimum	2,30	17,75	38,80
Std. Dev.	5,89	18,87	44,72
Coeff. of. Var.	57,12%	39,81%	46,87%
Skewness	0,68	0,2611	1,7291
Kurtosis	2,70	2,19	8,09
Sum	308,95	1421,85	2862,60
Observations	30	30	30
Jarque-Bera statistic	2,42	1,17	47,27
Probability	0,2988	0,5579	0,0000

Source: *EViews* and *Megastat*, Author's calculation

Extreme small data for variables X1=X1_INTER0408 and X2=X2_GDPAV0408 belong to Bulgaria. Data for variable X1 is the highest for Iceland. For variable X2 the Luxemburg's data is the highest, being the only serious outlying value in the whole dataset with standardized value higher than 3 (z>3), and it influences the shape of distribution to be significantly apart from the normal (see *Table 1*: the Jarque-Bera test on normality has got the p-value=0,0000<0,01). Data for variable X3_UNEMP0408 is the smallest for Denmark, and the highest for Slovakia. Variable X4_EDGDP0408 has got the smallest value for Turkey, and the largest for Denmark. Data for variable X5_CSKILL0408 is the smallest for Italy, and the greatest for Sweden.

Using multivariate analysis of standardized data based on cluster analysis with n=30 countries and 7 numerical variables, Ward linkage and Euclidean distance, four clusters were created, with the countries counted in *Table 2* (numbers are given in the dendrogram in ***Figure 1***, respectively).

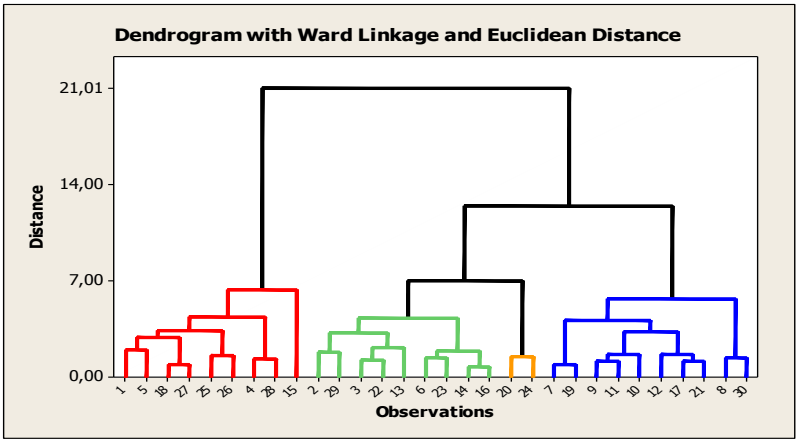
Table 2 Cluster analysis: n=30 European countries in four clusters according to seven variables

Cluster	Countries grouped into the clusters
Cluster 1	Belgium 1, Germany 5, The Netherlands 18, United Kingdom 27, Finland 25, Sweden 26, Denmark 4, Iceland 28, Luxembourg 15
Cluster 2	Bulgaria 2, Croatia 29, Czech Republic 3, Romania 22, Latvia 13, Estonia 6, Slovenia 23, Lithuania 14, Hungary 16
Cluster 3	Ireland 7, Austria 19, Spain 9, Italy 11, France 10, Cyprus 12, Malta 17, Portugal 21, Greece 8, Turkey 30
Cluster 4	Poland 20, and Slovakia 24

Source: *Mintab*, Author's calculation

Transition countries (Bulgaria, Czech Republic, Estonia, Latvia, Lithuania, Hungary, Romania,

Slovenia and Croatia) are grouped in one cluster, with exception of Poland and Slovakia, who created a cluster of their own. The most developed, mostly Scandinavian and northern countries as Finland, Luxembourg, and highly developed Sweden, Iceland, Belgium, Denmark, The Netherlands, Germany and United Kingdom are grouped in a separate cluster, too. And, finally, a special cluster was comprised of countries Spain, Portugal, France, Italy, Cyprus, Greece, Turkey, Malta, Austria, and Ireland.



Source: *Mintab*, Author's creation

Figure 1 The dendrogram: n=30 European countries in four clusters according to seven variables

The highest variability among countries is noticed within Cluster 1 (SS=37,575). Also, for the same Cluster the value of average distance from the centroid (1,8707) is the highest, and the highest is maximum distance from the centrolid (3,8871), see *Table 3*. Concerning variables under study, Belgium, Germany, The Netherlands, UK, Finland, Sweden, Denmark, Iceland, and Luxembourg seem to be over-average.

Table 3 Final Partition: n=30 European countries in four clusters according to seven variables

Cluster	Number of observations	Within cluster sum of squares (SS)	Average distance from centroid	Maximum distance from centroid
Cluster 1	9	37,575	1,8707	3,8871
Cluster 2	9	18,565	1,4143	1,8104
Cluster 3	10	27,236	1,5680	2,7475
Cluster 4	2	1,087	0,7372	0,7372

Source: *Mintab*, Author's calculation

The largest distance between cluster centroids appears between Cluster 1 (highly developed countries Belgium, Denmark, Germany, The Netherlands, Finland, Sweden, United Kingdom, Iceland, Luxembourg) and Cluster 4 (less developed, transition countries Poland, Slovakia) with value 4,906, *Table 4*.

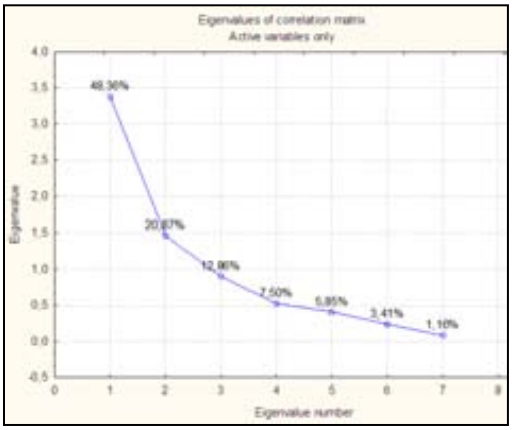
Table 4 Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	0,000			
Cluster2	3,949	0,000		
Cluster3	3,106	2,310	0,000	
Cluster4	4,906	3,066	4,137	0,000

Source: *Mintab*, Author's calculation

In Principal Components analysis a Scree Plot limits the number of PC by their contribution or numerical relevance, towards representing fractions of the total variance of the data. Only PCs associated with respective eigenvalues greater than or equal to 1E-8 are included in the calculation result set. Even though in practice, PCs close to 1 with respective eigenvalues (i.e., fractions of data total variance) could

be of interpretive use. After a Scree Plot in *Figure 2*, only independent variables X1 and X2 are suggested as regressors in the following regression alanysis.



Source: Statistica 9, Author's creation

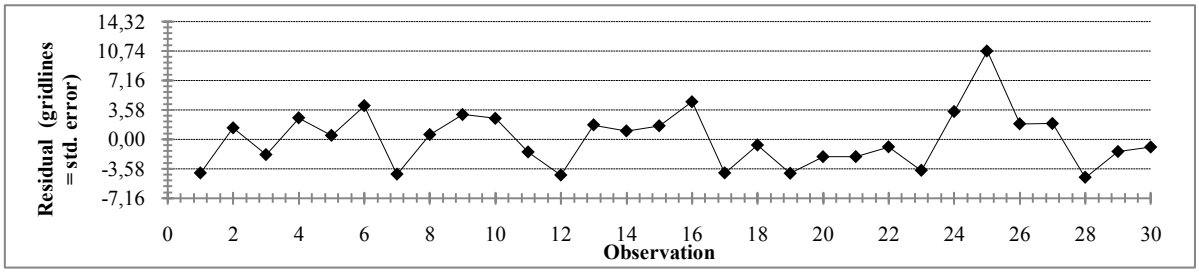
Figure 2 Scree Plot

Regression analysis

Since the multiple linear regression equation with all six proposed regressors and estimates of parameters calculated using OLS method showed that only variable X1 is statistically significant, but with multicollinearity problem (Variance Inflation Factor VIF_{X1}=5,25>5), some other models were evaluated. In further analysis multiple linear regression models based on different number of regressors *were* us*ed-T*he regression coefficients with *regressors* X2 and X4 (parameters β_2 and β_4) arise to be statistically significant at Type I Error of 5% barely X1 is excluded from the model. At the other hand, the parameter β_1 appeared to be statistically significant in all models *examined* with K=5 variables, when each time one of regressors X_j, for j=2, 3, 4, 5 or 6, was excluded. Finally, *model* with only K=2 regressors, X1 and X2, was accepted as the most appropriate. The model with estimated parameters looks as follows:

$\hat{Y} = -0{,}5670 + 0{,}3101\, X_1 - 0{,}0402\, X_2$ $(1{,}826)\quad (0{,}050)\quad (0{,}021)$	$\hat{\sigma} = 3{,}580$ $R = 0{,}809$ $R^2 = 0{,}655$ $\bar{R}^2 = 0{,}630$ $DW = 1{,}73$ n=30	(1)

The F-test for overall regression and the p-value=5,75E-07 shows that at least one regressor is statistically significant for the final model (given above) at 1% significance level. Coefficient of determination R^2 shows that 65, 51% of the total sum of squares is explained. Further, the regression diagnostics was conducted using the t –test for testing the significance of each of independent variables. The variable X1- Level of Internet access in households (Percentage of households who have Internet access at home) is statistically significant, with t-statistic=6,249 and p-value=1,10E-06, with Type I Error of 1%. The variable X2- GDP per capita in PPS (EU-27 = 100) is statistically significant, with t-statistic=-1,918 and p-value=0,0657, with Type I Error of 7%. *Figure 3* shows the regression residuals for the chosen model. The highest residual value appears to be for Finland ($e_{Studentized,\ 25}$=3,083).



Source: *Megastat*, Author's creation

Figure 3 Residuals for adopted regression model: n=30, K=2 (X1 and X)

Using *EViews 7* the regression diagnostics was conducted:

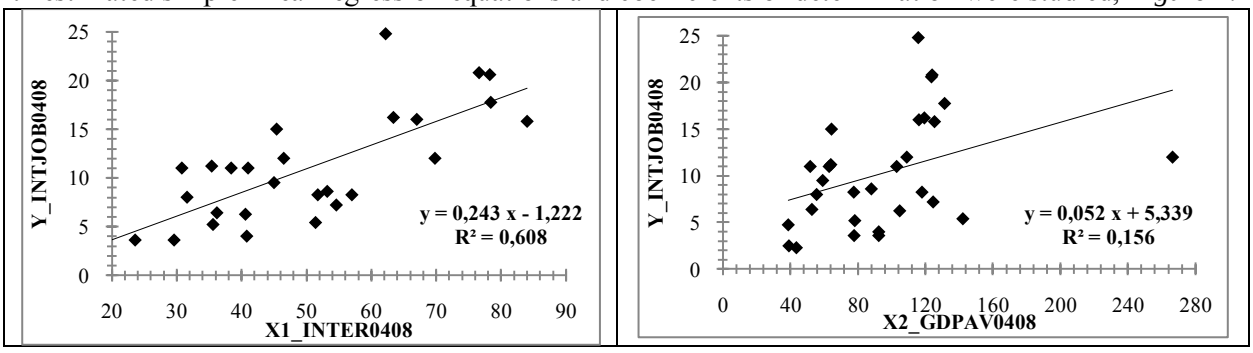
1. *Normality of residuals*. Based on the Jarque–Bera test of normality for residuals, with the Jarque-Bera statistics=4,9996 and the p-value=0,082, the null-hypotheses that residuals are from the normal distribution, may not be rejected at significance level of 5%.

2. *Autocorrelation of residuals.* Since, with n=30 and K=2, the values d_l=1,070 and d_U=1,331, the Durbin- Watson test value is greater than the critical value, i.e. $DW = 1,73 > d_U$, at 5% significance level there is no positive autocorrelation of residuals. Further, the Breusch-Godfrey test value is LM=0,4515 with p-value=0,7979. So, the null-hypotheses that there is neither the first nor the second order autocorrelation of the residuals may not be rejected at any reasonable significance level.

3. *Homoscedasticity of residuals.* The White test based on auxilliary regression gives the test statistic LM=3,8296 with the p-value=0,5742. So, at any reasonable significance level the null-hypotheses that there is no heteroscedasticity may not be rejected.

4. *Multicollinearity of regressors.* Possible multicollinearlty was diagnosed using the coefficient of determination and the Variance Inflation Factor *(VIF)* criterion. In Model II, the coefficient of determination R^2_{12}=R^2_{21}=0,4959 which does not reach 0,8, and the *VIF1=VIF2*=1,984<5 indicates there is no multicollinearity of regressors.

Going into more details, the scatter diagrams for the pairs of Y and each of the regressors X1 and X2, with estimated simple linear regression equations and coefficients of determination were studied, *Figure 4.*



Source: *Megastat*, Author's creation

Figure 4 Scattergrams with estimated simple linear regression equations and coefficients of determination for the pairs of Y and each of the regressors X1 and X2

In each of two simple linear regression models given in *Figure 3* the regressors are statistically significant and no regression model assumptions are violated.

Conclusion

Data exploration of all seven variables shows that Northern Europe, mostly Scandinavian, countries

are the leading ones concerning both economic development and ICT related variables, and they form cluster of their own being apart from the rest of European countries.

When studying impacts on dependent variable Percentage of individuals aged 16 to 74 using the Internet for looking for a job or sending a job application, research methods applied gave the final conclusion that only the model with two regressors is relevant, and that among six only the following two regressors: X1, Level of Internet access in households as Percentage of households who have Internet access at home, and X2, GDP per capita in PPS (EU-27=100), are statistically significant with no violations of the multiple linear regression model assumptions. Coefficient of determination for the adopted model shows that 65,51% of the total sum of squares is explained. The regression coefficient $\hat{\beta}_1$ shows that if X1, Percentage of households who had Internet access at home, would increase by one, with unchanged value of variable X2, GDPpc in PPS (EU-27=100), the regression value of a percentage of individuals aged 16 to 74 using the Internet for looking for a job or sending a job application, would increase by 0,3102. The regression coefficient $\hat{\beta}_2$ shows that if X2, GDPpc in PPS (EU-27=100), would increase by one, with unchanged value of variable X1, Percentage of households who had Internet access at home, the regression value of a percentage of individuals aged 16 to 74 using the Internet for looking for a job or sending a job application would decrease by 0,0402 (all data treated are based on average for 2004-2008). Estimated standard error of the regression model is 3,580. The t-test with p-value=0,000 says that variable X1-percentage of households who have Internet access at home, is statistically significant at 1% significance level. The t-test with p-value=0,066 shows that variable X2, GDPpc in PPS (EU-27=100) is significant at 7% significance level. Based on all diagnostics tests, no model assumptions are violated for the regression model adopted.

Restriction of this research is that some additional independent variables might be included into analysis and recommendation for a future research is to include them. Also, untypical extreme values for here investigated numerical variables recorded for highly developed countries, mostly Scandinavian, and especially discovered outlier for the Luxembourg' s data for GDP per capita in PPS, should be excluded from the analysis.

REFERENCES (RÉFERENCES)

Brenčič V., Norris J. B. Do employers change job offers in their online job ads to facilitate search? Econ. Letters; 2010(108): p.46–48.

Fountain C. Finding a Job in the Information Age: Job Searching, Labour Market Outcomes, and the Internet, 2003. http://www.allacademic.com/meta/p106817_index.html. [01/27/2011]

Gujaraty D. N., Porter D.C. Essentials of Econometrics. 4th Edt., McGraw-Hill IRWIN; 2010.

Hogler R. L., Henle C., Bemus C. Internet recruiting and employment discrimination: a legal perspective, Human Resource Management Review; 1998. 8(2): p. 149-194.

Hair F. H., Black W. C., Babin B. J., Anderson R.E. Multivariate Data Analysis, 7th Edt., 2008: Prentice Hall

Jackson A. A careers service on the Internet, Computers&Education; 1998. 30(1/2): p. 57-60.

Kinder T. The use of the Internet in recruitment—case studies from West Lothian, Scotland, Technovation; 2000 (20): p. 461–475.

Kuhn P., Skuterud M. Job search methods: Internet v. traditional, Monthly labour review; 2000, 123(10): p.3-11.

Suvankulov F. Job Search on the Internet, E-Recruitment, and Labor Market Outcomes, 2010-01-02; 2010. http://www.rand.org/pubs/rgs_dissertations/RGSD271.html [01/22/2011]

Tso G., Yau K., Cheung, M. Latent constructs determining Internet job search behaviours: Motivation, opportunity and job change intention, Computers in Human Behavior; 2010- 26: p. 122–131.b

RÉSUMÉ (ABSTRACT) -