Abstract of Doctoral Thesis "*Methodological and Applicative Problems of Correlation in the Analysis of Socio-Economic Phenomena*"

Author: Daniela-Emanuela Dănăcică

The main objective of this doctoral thesis is to examine methodological and applicative problems of correlation in the analysis of socio-economic phenomena. To achieve this objective, I focused on two main lines of research: methodological problems arising from the use of correlation and regression in the study of socio-economic phenomena, discussed in Chapter I of the thesis, and applicative problems of correlation in the analysis of these phenomena, presented in Chapters II, III and IV of the thesis.

In Chapter I of the thesis, "*Methodological Problems of Using Correlation in the Factorial Analysis of Socio-Economic Phenomena"* first I have concentrated my attention on the methodological problems arising from the interpretation of Pearson linear correlation coefficient. Pearson coefficient is widely used in analyzing the relationships between socio-economic phenomena, however the subtle nuances of this statistics are not sufficiently known in the literature and occasionally there are confusions about the interpretation thereof. Different interpretations attached to particular values of Pearson correlation coefficient must be presented very carefully, because the similarities between certain interpretations are subject to specific constraints. The size and interpretation of Pearson correlation coefficient is influenced by the shape of distribution, the sample size, restriction of empirical data amplitude, non-linearity and influence of other variables not included in the model. The influence of each factor is described in detail in paragraph (1.4) of the thesis. The paragraph (1.5) of the thesis is devoted to presenting special cases of Pearson correlation coefficient, the scope and methodology for calculating them (*point biserial* coefficient, *phi* coefficient, *biserial* coefficient, *tetrachoric* coefficient and *eta* coefficient). Since regression is an inferential method, operating on a sample of observations that offer the possibility to deduce and generalize the conclusions on the entire population, it is absolutely necessary to test the significance of regression parameters and to determine the corresponding confidence intervals for a given level of λ significance. The paragraph (1.6) of the thesis sets out the statistical inference in the simple linear model.

The stochastic linear dependence between an endogenous variable Y and a set of independent variables X_i is described using multiple regression. The paragraphs (1.7) and (1.8) deal with the methodological issues arising from the estimation of regression parameters, the inference of multiple regression model and ensuring of estimator stability. Multicollinearity, the primary generator of instability in the multiple regression model, raises significant methodological problems when it is sufficiently severe to affect the estimation of regression coefficients. Multicollinearity detection is performed using the correlation matrix of exogenous variables X_i, calculating and interpreting the determinant of this matrix, or using the *tolerance* statistics and *variance inflation factor* statistics. In the statistics literature we can find a number of solutions to mitigate or even eliminate the multicollinearity. But the solutions offered are not sufficient for the complexity of the socio-economic analysis. Thus, estimators lose their usefulness, as some of the factors whose influence was sought to be analyzed are no longer found in the calculated indicator or the estimators can not be assigned clear socio-economic interpretations.

If distributions of correlated variables are not normal, or if the individual values of variables are not expressed numerically, in order to assess the intensity of their relationship we shall use nonparametric methods. In this case we do not use the real values of variables but their ranks. Rank correlation coefficients are only measurement indicators of the relations between two statistical variables. Chen and Popovich (2002) point out that they may be affected by sample selection fluctuations.

The paragraph (1.11) deals with methodological problems specific to survival analysis. Although at the beginning the survival analysis was used to study death as an event specific to medical studies, as from the '70s these statistical techniques have been increasingly used in economics and social sciences. Besides the fact that survival data are not normally distributed, they often contain incomplete information, censored subjects. Censoring of subjects may be on the right or left. It is vital to include censored subjects in the statistical analysis. But, according to Greene (2003), a very large number of censored subjects may affect the accuracy of statistical tests. I have presented in this section methodological aspects of Kaplan-Meier analysis, statistical significance testing for the resulted survival curves and the peculiarities of various statistical tests used. I have also concentrated my attention on the Cox regression, and I have set out the concept of *hazard, baseline hazard, hazard rate, hazard rate interpretation*. I have also pointed out that the proportional hazards assumption is crucial for the Cox regression model. The proportional hazards assumption can be checked using the *log-minus-log* curve or with the help of *partial (Schönfeld) residuals.* In the first case, if baseline hazards are proportional, then the lines corresponding to individual layers must be parallel; in the second case proportional hazards assumption requires that in the *Schönfeld* graph there should be no *pattern*. If proportional hazard assumption is violated, then we shall build a Cox model with non-proportional hazard by entering an interaction between the specific covariate and time. Paragraph (3.4) shows the applicative problems of using the Cox regression model with a time-dependent covariate.

In Chapters II, III and IV, I wanted to present the applied side of the methods and models presented in Chapter I of the thesis. I focused on studying the factors influencing the labour resources, unemployment duration and the likelihood of employment. I chose employment and unemployment as my research area because the identification of factors influencing the labour resources, unemployment duration and the likelihood of (re)employment remains one of the main issues faced by developed economies.

Chapter II "*Estimating the Effect of Factors Influencing the Evolution of Labour Resources in Romania during 1990-2006 Using Correlation and Regression Analysis*" opens the series of applied studies of my thesis. The chosen sequence is not random, as a starting point I wanted to analyze how the labour resources are trained, investigating further issues related to the duration of unemployment and the likelihood of employment. Starting from the studies of Earle and Pauna (1998, 1999, 2006), Bugundi (2006), Kollo and Vincze (1999), Fields (2004) I have used three models specific to the labour market, *the total population model*, *the employed population model* and *the unemployment rate model* and using correlation and regression I have estimated the influence of exogenous variables of these models on specific outcome variables. During the analyzed period, 1990-2006, the total population, one of the analyzed endogenous variables had a downward trend; the evolution of this variable was positively influenced by the evolution of natural growth of population and by the gross domestic product index, and adversely by the emigration trend. The value of correlation coefficient R equal to 0,906 suggests a strong relationship between the total population endogenous variable and the exogenous variables of natural growth, emigration and the index of gross domestic product. A second endogenous variable studied, the employed population, had a downward trend throughout the analyzed period. It has been positively influenced by the total population evolution and by the activity rate of population; the saving rate of the population has adversely influenced the endogenous variable, but it did not enjoy a statistical significance. The third endogenous variable analyzed, ILO unemployment rate had an upward trend for the period 1990-1995 and 1996-1999, then decreasing until 2006, the final year of my study. It has been positively influenced by the evolution of GDP index per capita and by the rate of social pressure, and adversely by the consumer price index and real wage index. The calculated values of Fisher-Snedecor statistics, higher than the critical values, proved the validity of the chosen models.

In Chapter III of the thesis, "*Estimating the Effect of Factors Influencing the Duration of Unemployment – a Duration Model Approach"* I have examined the influence of exogenous variables *gender*, *age* and *educational level* for the duration of unemployment in Gorj County, using the survival analysis. I used the data made available by the National Agency for Employment (NAE) Bucharest; the sample was made of 80.961 records, with information on the start date of unemployment, end date of unemployment, gender, age, educational level and the reason of unemployment leaving for each registered person. The minimum duration of unemployment was 0 months, the maximum duration of 57 months, the average 8.8 months and median of 6 months. The corresponding unemployment distribution is asymmetric, with a kurtosis of 5.583. Kaplan-Meier analysis showed that the probability to remain unemployed is higher for unemployed men compared with women; the median of unemployment duration calculated for female unemployed is 10 months and for men is 13 months. As for the age variable, along with the growing age of registered persons, their probability to remain unemployed increases as well. For the educational level variable, the probability to remain unemployed is higher for people without education, followed by apprentice school graduates and post high school graduates. For university graduates the probability to remain unemployed until the time *t* or after this time decreases much more rapidly compared with other groups, indicating that people with higher educational levels have better employment opportunities on the labour market of Gorj County. Logrank test results suggested the existence of statistical significance for all three variables analyzed. In the next step I used the Cox proportional hazard model to determine the hazard rates for each variable analyzed. Cox regression results showed that the hazard of leaving unemployment is 14% lower for women compared to unemployed men. The increase of age variable causes a reduction in the hazard of leaving unemployment by 0.2% annually. Regarding the educational level variable, all its 5 levels present hazard rates lower than 1; the lowest hazard rate (0.277) was recorded by a group of unemployed persons without education, and the highest value of hazard was recorded, as I expected, by the unemployed graduates of higher education. It was interesting the observation that the hazard of (re)employment shows a slightly higher value for the educational level 1 compared to the educational level 2, although the subjects that make up this group are less educated than those who make up group 2. This result shows that the theoretical high school graduates have fewer opportunities for (re)employment compared with graduates from vocational apprentice schools because, on the one hand, general theoretical training does not provide an advantage on the labour market, on the other hand, the expectations of a person with a higher educational level are higher. Hazard proportionality testing was made using the *log-minus-log* curve and with the *partial (Schönfeld) residuals.* Both graphical representations have suggested the violation of proportional hazard assumption. In the next step, beside the age variable we introduced into the Cox proportional hazard model the term of interaction between time and age (*time*age*). The results showed that the hazard rates for the variables estimated in the Cox model with *time*age* interaction are slightly lower compared with the hazard rates calculated using the Cox proportional hazard model. The study performed is the first of its kind in Romania, using survival analysis for modelling the duration of unemployment. As a future research project, I would like to extend this type of study for the entire country, analyzing the influence of other exogenous variables such as region, marital status, health, specialization, etc. for the duration of unemployment and the potential relationship between duration of unemployment and migration registered within the territory of our country.

Chapter IV, "*Employed or Unemployed? An Logistic Regression Approach"* completes the research area of applicative problems of correlation in the analysis of socio-economic phenomena. In this chapter I have analyzed the influence of the exogenous variables *gender*, *age* and *educational level* on the probability of employment of persons registered as unemployed at NAE Bucharest in the period 1 January 2002-31 August 2006. As a methodology I have used Kaplan-Meier analysis and binary logistic regression, estimating with these statistical techniques the probability of individuals to be employed or unemployed at the end of the analyzed period, depending on gender, age and educational level. Of 80961 persons registered in the database of Gorj County as unemployed, during 1 January 2002 - 31 August 2006, 19369 persons became employed until August 31, 2006; the reason for their unemployment leaving was filled with *"employed"*. The average duration of unemployment until finding a job is of 6 months, the median of 2 months, the maximum value - 57 months and the minimum value - 0 months. For the gender variable the survival curves have shown a probability of employment higher for male subjects compared with women during the first months after the entry into unemployment, then the situation is reversed, women record a higher probability compared with men. After 30 months the curves coincide, the gender factor loses its influence on the binary outcome variable *status* (*1 – employed, 0 unemployed*). For the age factor the survival curves have showed that the probability of event occurrence – *employment* – decreases with the growing age of subjects in the database. In the first 10 months the probability of the pre-established event occurrence is higher for the 55-64 year-old group compared with the 25-34, 35-44 and 45-54 year-old groups. The 15-24 year-old group has the highest probability of employment. After 40 months from the date of registration as unemployed in the database, the curves coincide and the influence of age factor became insignificant for the binary variable *status*. Regarding the educational level variable, survival curves revealed that the subjects with university education had the highest probability of achieving the event, followed by subjects graduating from vocational and foremen school; the most disadvantaged category is the group of people without education or with primary education / secondary education. Logrank test results revealed the existence of statistical significance for all investigated variables. The results of logistic regression showed that the probability of employment increased by 1.6 for men registered as unemployed compared to women with the same status. In the period analyzed, in the database there were registered 33270 female unemployed and 47691 male unemployed. Among them, at the end of the spell 6390 women (19.21%) and 12979 men (27.21%) became employed. Although there much more men registered as unemployed in the analysed database, the number of those who manage to find a job is higher compared to women, which indicates that although there are more unemployed men, however they are preferred by employers. For the age variable, the probability of employment increases by 1.06 at one unit (year) change. By age groups, the probability of employment increases by 1.3 for the 15-24 year group, compared to the 55-64 year group, by 2.107 for the 25-34 year group, compared to the specified group, by 2.125 for the 35-44 year group, compared to the last age group, and by 1.664 for the 45-54 year group, compared to the 55-64 year group. As for the educational level variable, the probability of employment increases by 1.25 at one unit change. Indeed, the higher the educational level, the higher the subject's probability of employment. The most disadvantaged educational groups proved to be the persons without education, the persons with unfinished secondary school, vocational school, apprenticeship complementary education and special education and the theoretical high school graduates. The results of Omnibus, Hosmer & Lemenshow and Wald test have shown the validity of the model and the statistical significance for the estimates of model variables.