BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LX No. 1/2008	57 - 62	Seria Matematică - Informatică - Fizică

An Application in Social Domain using SAS

Mădălina Cărbureanu

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti, Catedra de Informatică
e-mail: mcarbureanu@upg-ploiesti.ro

Abstract

For accomplish a good analysis of the imense volumes of data, it is neccesary the usage of some powerful instruments. In this paper, are presented some features of SAS software and an example of utilizing this software for solving a certain problem from the social domain. This software is a comprehensive package with very powerful data management tools, a wide variety of statistical analysis and a batch of graphical procedures.

Key words: *procedure, data analysis, correlation, regression*

Introduction

In this paper we present some basic statistical procedures of an instrument for data analysis, named SAS (Statistical Analysis System). The SAS version we used is 9.1.3 Service Pack 4 with license, developed in 2005, by the SAS Institute. The most important aspect is that it allows the user to achieve descriptive analysis of the data, correlation, regression and variance analysis.

This software can be applied to solve different types of problems from various domains. In our case is approached a problem from the social domain, regarding the factors witch influence the salary income of a person. For solving this problem we used the values of the gross domestic product (*GDP*), the salary income and the wage earners values for each development region of our country, offered by the National Institute of Statistics, for 2006, as we can see in Table1 [1].

Table 1. The data offered by INSSE

Region	Wage_earner_no	Salary_income	GDP (RON)
N-E	567000	8530000	2941820
S-E	551000	9060000	2941310
S-Muntenia	571000	9240000	3170950
S-V Oltenia	393000	9580000	2196250
V	484000	9260000	2529690
N-V	580000	8710000	3026980
C	576000	8570000	3009640
Bucharest-Ilfov	837000	1291000	4807760

A Few Words about SAS

SAS was conceived by Anthony J. Barr in 1966. This software was improved lengthways the years, knowing different versions, as follows: SAS 71, 72, 76, 79.3-82.4, version 4-9 series.

This software has different components and licenses, which must be installed separately. Some of these components are:

- SAS Add-In for Microsoft Office (provide acces to the data, analysis and many others facilitations, using menus integrated into Office applications);
- Base SAS (is used to manage data and represents the *core* of SAS);
- SAS Enterprise Business Intelligence Server (contains various instruments and a platform who provide acces to the data);
- SAS/ACCESS (offers SAS the posibility of sharing data with different types of applications);
- SAS/ASSIST (represents the SAS interface);
- SAS/CONNECT (offers SAS the posibility to comunicate with other platforms);
- SAS/GRAPH (contains instruments for building different types of graphs);
- SAS/STAT (supplies different tools for making analysis of variance, regression, multivariate analysis, and categorical data analysis).

In this software, for achieving the proposed goals, the user must write SAS programmes. An SAS programme contains two sections: the *DATA* section and the *PROC* section.

The *DATA* section contains the analised data and a data dictionary. This dictionary contains informations about the variables used and their properties.

The *PROC* section is the section in witch the user may achieve different types of data analysis [3].

SAS is very useful because the user may achieve diferent types of operations upon the data, for instance [2]:

- The generating of frequency histograms (*PROC FREQ*);
- The t-test performing (*PROC TTEST*);
- Analysis of variance and correlational analysis (*PROC GLM* and *PROC CORR*);
- The generating of the graphical images (*PROC UNIVARIATE*);
- Graphical presentation of the data (*PROC CHART, PROC PLOT* and *PROC BOXPLOT*);
- Transformations of the data (*PROC SORT, etc.*);
- The prediction of a variable using another variable (*PROC REG*).

Application Example

Upon the data offered by the National Institute of Statistics, presented in figure 1, we chose to apply the corelation and the regression, supplied by the SAS software.

A useful statistic is gived by the *Pearson correlation coeficient*, knowed as *Pearson's*. In SAS, simple or multiple correlations can be made using the *PROC CORR* procedure. This procedure supplies descriptive information and indicates the *p* value at witch the correlation can be considerate semnificative. The Pearson correlation indicates the association level between variables.

The data base implemented in SAS has three variables, as follows: *region*, *wage_earner_no*, *salary_income* (milions) and *GDP* (Gross domestic produc, RON) and is presented in figure 2:

```
DATA TEST;
INPUT region $ wage_earner_no salary_income GDP;
CARDS;
N-E 567000 8530000 2941820
S-E 551000 9060000 2941310
S-Munt 571000 9240000 3170950
S-V-Olt 393000 9580000 2196250
V 484000 9260000 2529690
N-V 580000 8710000 3026980
C 576000 8570000 3009640
Buc-Ilf 837000 12910000 4807760
;
PROC PRINT DATA=TEST;
RUN;
PROC CORR DATA=TEST;
var wage_earner_no salary_income GDP;
RUN;
PROC REG DATA=TEST;
MODEL salary_income=GDP wage_earner_no;
RUN;
PLOT r.*p.;
run;
```

Fig. 2. The SAS data base

Appling the corelation using *PROC CORR* procedure, as we can see in figure 2, we obtained the following statistical data, presented in figure 3:

The SAS System 13:53 Wednesday, Jur

The CORR Procedure

3 Variables: wage_earner_no salary_income GDP

Simple Statistics

	N	Mean	Std Dev	Sum	Minimum
wage_earner_no	8	569875	125653	4559000	393000
salary_income	8	9482500	1433315	75860000	8530000
GDP	8	3078050	767495	24624400	2196250

Pearson Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	wage_earner_no	salary_income	GDP
wage_earner_no	1.00000	0.72536 0.0417	0.98842 <.0001
salary_income	0.72536 0.0417	1.00000	0.80924 0.0150
GDP	0.98842 <.0001	0.80924 0.0150	1.00000

Fig. 3. The *PROC CORR* results

Simple Statistics supplies informations regarding the used variables, as follows: the number of instances, the mean, the standard deviation and the maximum and minimum values of the variables.

The second part consists in presenting a *correlation matrix* for *wage_earner_no*, *salary_income* and *GDP* variables. The correlation of a variable with themselve is perfect positive (r =1) and doesn't represent any interest for the user. Each cell of the correlation matrix contains the value of the Pearson Correlation Coefficient (r) and the level of signifiance p.

As we can see from the figure 3, we have a semnificative correlation between *wage_earner_no* variable and *GDP* variable (r =0.98, p<0.0001). Also, good correlations we have bettween *wage_earner_no* and *salary_income* (r =0.72, p<0.0417), also between *salary_income* and *GDP* (r =0.80, p<0.0150).

The SAS procedure *PROC REG* is used for achieve simple and multiple regressions. For using this procedure all the variables from the model must be of continuous type. Otherwise it is used the *PROC GLM* procedure.

We can use *PROC REG* for predicting *salary_income* values using *GDP* and *wage_earner_no* variables. The application of the *PROC REG* is presented in figure 2. The result of this procedure is presented in figure 4.

The SAS System 13:53 Wednesday, June 4, 2008 3

The REG Procedure
Model: MODEL1
Dependent Variable: salary_income

Number of Observations Read	8
Number of Observations Used	8

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.288417E13	6.442086E12	21.52	0.0035
Error	5	1.496579E12	2.993158E11		
Corrected Total	7	1.438075E13			

Root MSE	547098	R-Square	0.8959
Dependent Mean	9482500	Adj R-Sq	0.8543
Coeff Var	5.76955		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7479771	1153714	6.48	0.0013
GDP	1	7.48351	1.77543	4.22	0.0084
wage_earner_no	1	-36.90615	10.84443	-3.40	0.0192

Fig. 4. The *PROC REG* results

PROC REG procedure, supplies a table in witch is presented among others the standard errors, the t and p values. The results presented in figure 4, shows that *GDP* and *wage_earner_no* variables are predictind the *salary_income* value.

We have also, several simple statistics. The *Root MSE* (Root Mean Squared Error) is an estimation of the standard deviation of the error term. The coefficient of variation, or *Coeff Var*, is a unitless expression of the variation in the data. The *R-Square* (0.8959) and *Adjusted R-*

Square (0.8543) are two statistics used in assessing the fit of the model. Values of these two statistics close to 1 indicate a better fit. In our case, the value of *R-Square* statistic of 0.89 indicates that *GDP* and *wage_earner_no* variables accounts for 89% of the variation in *salary_income* variable.

From the *Parameter Estimates* section, the fitted model is:

$$salary\_income = 7480000 + 74835 \times GDP - 36906 \times wage\_earner\_no .$$

The *REG* procedure can be used also interactively. After we specify a model with the *MODEL* statement and submit the *PROC REG* statements, we can submit further statements without reinvoking the procedure. The *PLOT* command can now be issued to request a plot of the residual versus the predicted values, as shown in figure 5.

In figure 5, we can observe a random fluctuation about zero, so we don't have a pattern for the analised values.

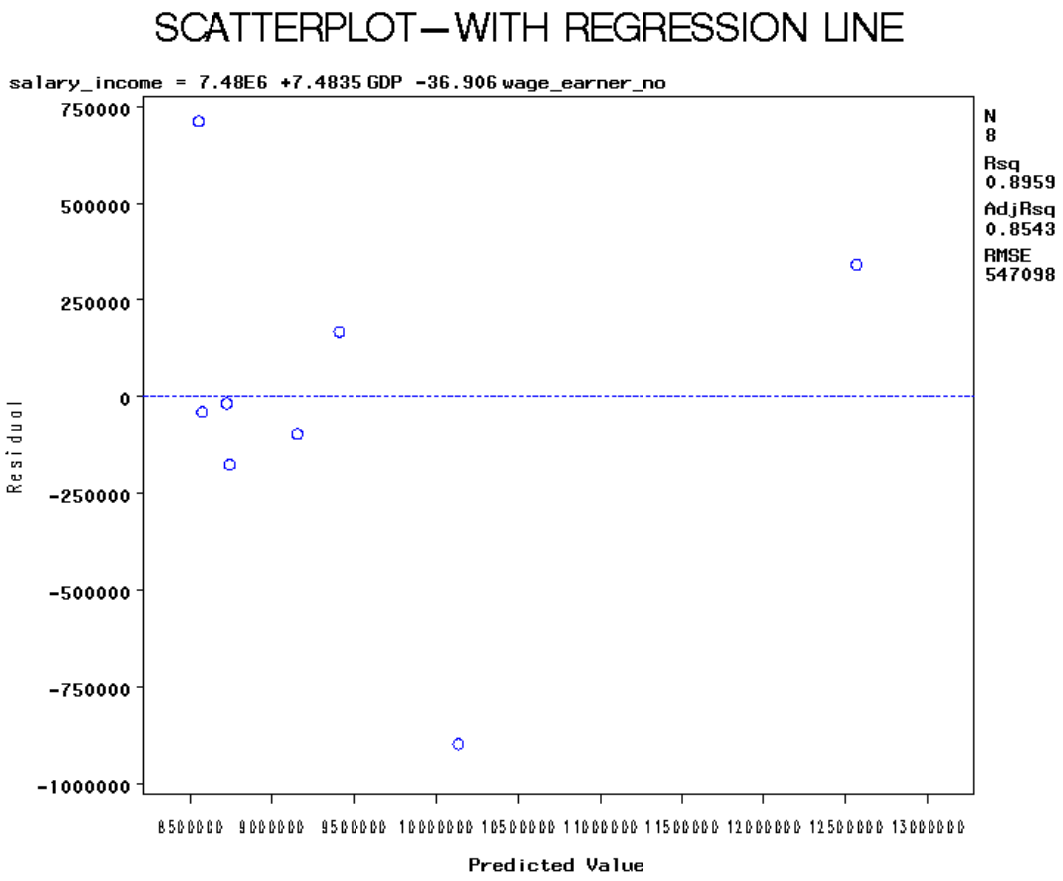

Fig. 5. The *PROC REG* results

Conclusion

Having a batch of components which offers the user the possibility of making statistical analysis for solving problems from any domain, SAS became a very powerful data analysis tool. Although, the earlier versions of SAS software had a poor interface, SAS 9.1.3 Service Pack 4

version, delivers comprehensive data integration, business intelligence, and analytical capabilities beyond just the statistical functions of others analogous free tools.

References

1. * * * - Institutul Național de Statistică, http://www.insse.ro, accessed 10 May 2008
2. * * * - *SAS Tutorials*, http://web.fccj.org/~jtrifile/SAS2.html, accessed 20 April 2008
3. * * * - *SAS Tutorials*, http://instruct.uwo.ca/sociology/300a/SASintro.htm, accessed 15 April 2008

O aplicație în domeniul social utilizând SAS

Rezumat

Pentru a realiza o bună analiză a unor volume imense de date, este necesară utlizarea unor instrumente puternice. În cadrul acestei lucrări, vom prezenta câteva caracteristici ale software-ului SAS şi un exemplu de utilizare a acestuia pentru a rezolva o anumită problemă din domeniul social. Acest software este un pachet cuprinzător conținând instrumente pentru managementul datelor, foarte puternice, o gamă largă de proceduri de analiză statistică şi o serie de proceduri grafice.