Statistical Data Analysis via R and PHP: A Case Study Of the Relationship Between GDP and Foreign Direct Investments for The Republic Of Moldova

PhD Candidate Ştefan Cristian CIUCU (stefanciucu@yahoo.com)
The Bucharest University of Economics Studies, Romania

ABSTRACT

This paper provides an overview over a way of integrating R with PHP scripting language in order to analyze statistical data (time series).

We analyze the relationship between the foreign direct investments and GDP of the Republic of Moldova over 1992-2012 time period.

***Keywords**: R software; PHP; programming; GDP; FDI.*
***J.E.L. Classification:** C88; L8.*

INTRODUCTION

R is a language and environment for statistical computing and graphics. It is a GNU project (free software) which was designed by Ross Ihaka and Robert Gentleman.

R software is gaining ground in Romania, as research papers appear. The study of (Caragea, 2012) can be mentioned, mostly for the fact that it underlines the importance and usability of R for statistical computations, data analysis, visualization and applications in various fields.

Also, R is a powerful software that can work with databases, for example in (Dobre, 2013) the manipulation of large databases is treated.

At the moment, PHP is a popular general-purpose scripting language that is especially suited to web development[1]. PHP is behind most of the websites, from blogs to presentation websites or web-based applications.

1. http://php.net/

It has appeared in 1995 and it is influenced by Perl, C, C++, Java and Tcl programming languages.

The idea of integrating R with PHP has been around for a few years. A software called R-PHP, is developed within the Department of Statistical and Mathematical Sciences of the University of Palermo (Italy) and can be used as it is an open-source project with the code released by the authors with a General Public License (GPL) (so it can be freely installed and used). A lot of information can be found in (Mineo, Pontillo, 2006).

Also, there are many studies of the impact of FDI on GDP. For example, the study of (Agrawal, 2011). The foreign direct investments are very important for countries in transition. They usually bring a great plus to the economic development of a country.

THE USE OF PHP IN STATISTICS

PHP language has some mathematical extensions that include numerous functions. Three of the extensions are worth to be mentioned: "Math" (with basic Mathematical Functions), "Statistics" - a statistics extension that contains functions for statistical computations and "Trader" - Technical Analysis for Traders which contains some functions for linear regressions.

All of these three extensions contain a few dozens of functions useful for statistical computations, but a large part of them are not documented at the moment (see [3] and [7]).

From the statistical functions of PHP we can mention:

- stats_absolute_deviation - returns the absolute deviation of an array of values;
- stats_cdf_f - calculates any one parameter of the F distribution given values for the others;
- stats_cdf_t - calculates any one parameter of the T distribution given values for the others;
- stats_standard_deviation - returns the standard deviation;
- stats_variance - returns the population variance;
- trader_linearreg - linear regression;
- trader_linearreg_slope - linear regression slope;
- trader_linearreg_intercept - linear regression intercept;
- trader_linearreg_angle - linear regression angle**.**

For example, the description provided on php.net website for the linear regression function:

```
array trader_linearreg (array $real [,integer $timePeriod])
```

where

real - Array of real values.

timePeriod - Number of period. Valid range from 2 to 100000.

and the return values:

Returns an array with calculated data or false on failure.

An example of PHP for statistics implementation:

- stats_stat_correlation - it calculate the Pearson's Correlation Coefficient of two arrays.

```
<?php
function correlation($x, $y)
{
$PPMCC = stats_stat_correlation($x, $y);
echo "Pearson product-moment correlation coefficient is " .$PPMCC;
}
$array_x = array(5,3,6,7,4,2,9,5);
$array_y = array(4,3,4,8,3,2,10,5);
correlation($array_x, $array_y);
?>
```

The main observation is that plain PHP is not powerful enough for statistical analysis and doesn't have enough already made functions for statistics or econometrics (even if we custom create the possibility of reading data from files or validate data). Also, writing PHP code for statistics and econometrics can be pretty laborious and requires massive programming knowledge (mostly because a lot of the functions are not documented).

R VIA PHP

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering and so on) and graphical techniques, and is highly extensible[1].

Accessing R via PHP can be done multiple ways and when implemented, it provides great flexibility for the user. Basically a web application using R, accessible from anywhere (over an Internet connection) can be done. The final user only needs the installation of a Web Browser and a stable Internet connection.

1. http://www.r-project.org/

In this article we will take a look at a way to access R via PHP already implemented, with a software called R-PHP.

R-PHP implements **two modules**:

1. The first module allows the simple insertion of the R code and it prints its output (analyses and plots) in another page.
2. The second module makes some statistical analysis by using a GUI.

It is available for download with proper documentation and demonstrations.

In our case, in order for R-PHP software to work, it was installed on a computer with an Intel Core2Duo E8500 3.16 GHz processor, 4GB RAM, Linux version Fedora 13 and with PHP version 5.3.8.

Also, the R-PHP software can be configured to work on a virtual machine, for example Oracle VM VirtualBox.

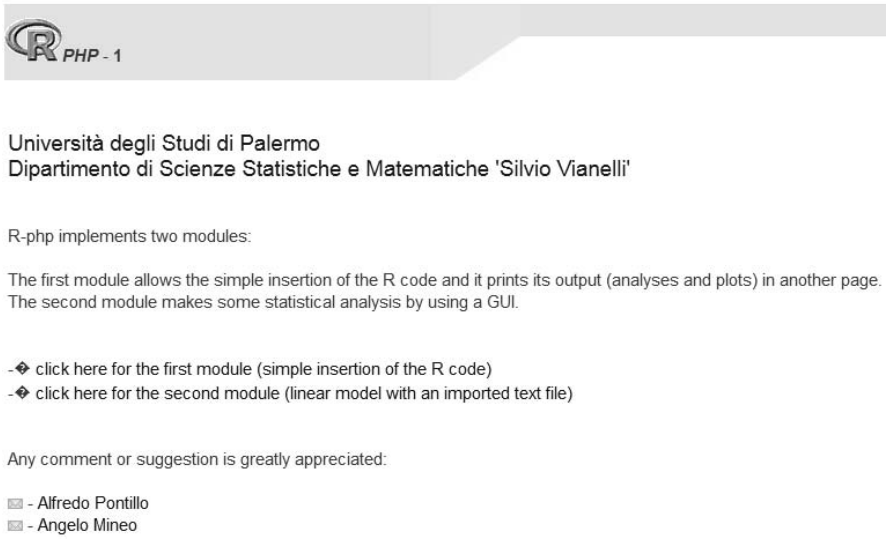
Besides the standard procedure of installation, the following steps and commands are necessary:

- chmod 0777 R/pages/tmp
- chmod 0777 R-gui/pages/tmp
- in the file include/conn.php at the command "CREATE TABLE 'dangerous' ... - delete "TYPE=MyISAM"

The chmod is a Unix shell command used to change access permissions to files and directories. The rights are given to three groups: OWNER, GROUP and OTHERS, the 7 gives the read, write, and execute permissions.

Home page of R-PHP

Figure 1



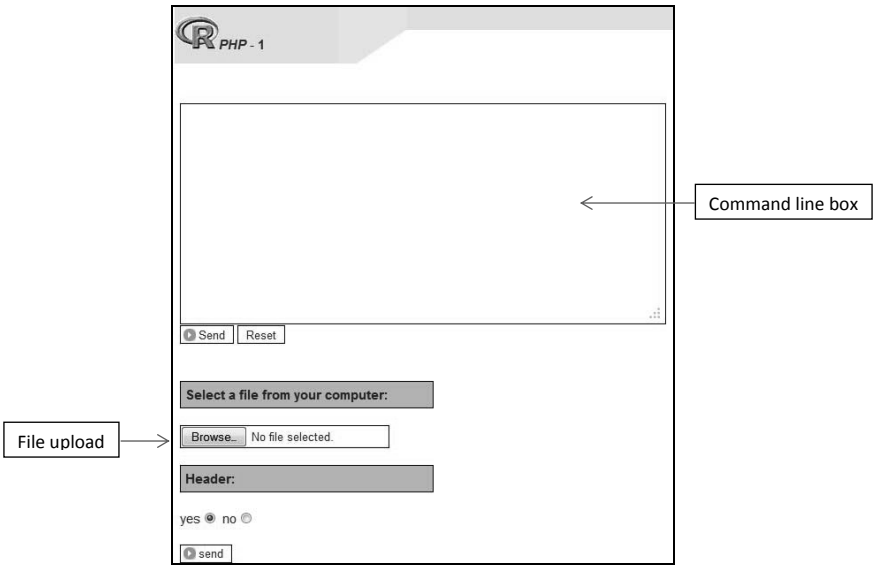
Using **the first module** requires knowledge of R programming language, but it gives total freedom to the user to run any commands supported by R software on the data.

The second module is a user friendly implementation and requires no programming skills. As a final user, you upload the data on the server and then you can run some statistical analysis. The menus are pretty suggestive and well thought out.

Using the second module the final user can run a descriptive analysis, a linear regression, an analysis of variance, a cluster analysis, a principal component analysis, a factor analysis or a metric multidimensional scaling.

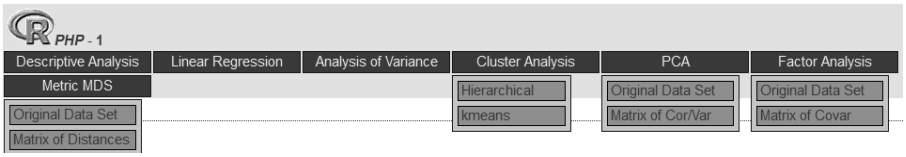
R-PHP module no. 1

Figure 2



R-PHP module no. 2 menu

Figure 3



DATA ANALYSIS USING R-PHP

In this section of the article, R-PHP will be used in order to analyze the relationship between GDP and foreign direct investments in the Republic of Moldova.

This section has **two major goals**:

1. to show basic way of using R-PHP for data analysis;

2. to interpret data and draw some conclusions about the economic situation in the Republic of Moldova.

This paper adopts a country-specific time series data from 1992 to

2012 (data for a longer period of time is unavailable). The data source is *The World Bank* - http://data.worldbank.org[1].

The data used:

GDP and FDI of the Republic of Moldova over 1992-2012 time period

Table 1

Year	GDP (US$)	FDI (US$)	Year	GDP (US$)	FDI (US$)
1992	2.319.243.407	17.000.000	2003	1.980.901.554	73.750.000
1993	2.371.812.924	14.000.000	2004	2.598.231.467	87.690.000
1994	1.702.314.353	11.568.000	2005	2.988.172.424	190.700.000
1995	1.752.995.314	25.910.000	2006	3.408.454.198	258.680.000
1996	1.695.130.484	23.740.000	2007	4.402.495.921	536.020.000
1997	1.930.071.445	78.740.000	2008	6.054.806.101	726.610.000
1998	1.639.497.207	75.500.000	2009	5.439.422.031	135.150.000
1999	1.170.785.048	37.890.000	2010	5.811.622.394	201.500.000
2000	1.288.420.223	127.540.000	2011	7.015.201.446	276.420.000
2001	1.480.656.884	54.540.000	2012	7.252.769.934	184.940.000
2002	1.661.818.168	84.050.000			

1. Consulted on 15 February 2014.

In the above table we have:

- GDP (current US$)[1] - GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars. Dollar figures for GDP are converted from domestic currencies using single year official exchange rates. For a few countries where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used.

- Foreign direct investment, net inflows (BoP, current US$)[2] - Foreign direct investment are the net inflows of investment to acquire a lasting management interest (10 percent or more of voting stock) in an enterprise operating in an economy other than that of the investor. It is the sum of equity capital, reinvestment of earnings, other long-term capital, and short-term capital as shown in the balance of payments. This series shows net inflows (new investment inflows less disinvestment) in the reporting economy from foreign investors. Data are in current U.S. dollars.

The data must be prepared for the R software, so a file with the data will be created, a *.txt file, with the data organized on columns with a tab separator. R supports many file types, but in this article a *.txt file will be used.

R-PHP - first module - code insertion

In the browser, after accessing the URL of the installed application, the file from the computer will be selected (in order to be uploaded on the server) and then in the command line of the R-PHP application, the following code can be run.

```
date.analizate <- read.table("a.tab.txt", header=TRUE)
attach(date.analizate)
names(date.analizate)
date.analizate
```

After running the code, a new tab will be created in the browser with the following result (meaning that our data has been read and stored into a data-frame):

1. http://data.worldbank.org/indicator/NY.GDP.MKTP.CD

2 http://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD

R-PHP data read output

Figure 4


```
> date.analizate<-read.table("a.tab.txt", header=TRUE)
> attach(date.analizate)
> names(date.analizate)
[1] "Year" "GDP"  "FDI"
> date.analizate
   Year        GDP       FDI
1  1992 2319243407  17000000
2  1993 2371812924  14000000
3  1994 1702314353  11568000
4  1995 1752995314  25910000
5  1996 1695130484  23740000
6  1997 1930071445  78740000
7  1998 1639497207  75500000
8  1999 1170785048  37890000
9  2000 1288420223 127540000
10 2001 1480656884  54540000
11 2002 1661818168  84050000
12 2003 1980901554  73750000
13 2004 2598231467  87690000
14 2005 2988172424 190700000
15 2006 3408454198 258680000
16 2007 4402495921 536020000
17 2008 6054806101 726610000
18 2009 5439422031 135150000
19 2010 5811622394 201500000
20 2011 7015201446 276420000
21 2012 7252769934 184940000
>
```

Now that our data could be read by the R via PHP, we can run typical R commands for statistical analysis.

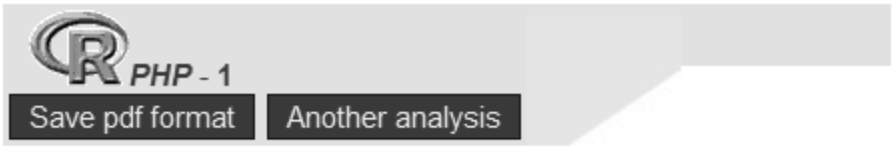
For data summary, the following code can be run:

```
date.analizate <- read.table("a.tab.txt", header=TRUE)
summary(date.analizate)
```

In a new browser tab, the summary will be presented:

Data summary

Figure 5




```
> date.analizate<-read.table("a.tab.txt", header=TRUE)
> summary(date.analizate)
      GDP                 FDI
 Min.   :1.171e+09   Min.   : 11568000
 1st Qu.:1.695e+09   1st Qu.: 37890000
 Median :2.319e+09   Median : 84050000
 Mean   :3.141e+09   Mean   :153425619
 3rd Qu.:4.402e+09   3rd Qu.:190700000
 Max.   :7.253e+09   Max.   :726610000
>
```

It can be seen that R analyzed the three variables successfully. For each variable the min., max., median, mean and 1st and 3rd quantiles are displayed.

Next, a linear regression for the data will be run. The regression model used is:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

where

y y = GDP (current US$);
x_1 x_1 = foreign direct investment, net inflows (BoP, current US$);
ϵ ϵ = random variable.

The code for the linear regression is:

```
date.analizate <- read.table("a.tab.txt", header=TRUE)
summary(lm(date.analizate$GDP~date.analizate$FDI))
```

Where the *lm* command runs the linear regression and the summary gives in the output all the details needed. The *lm* command only returns the coefficients, so the *summary* command is added.

Linear regression – GDP & FDI variables

Figure 6


```
> date.analizate<-read.table("a.tab.txt", header=TRUE)
> summary(lm(date.analizate$GDP~date.analizate$FDI))

Call:
lm(formula = date.analizate$GDP ~ date.analizate$FDI)

Residuals:
       Min         1Q     Median         3Q        Max
-1.672e+09 -9.691e+08 -4.965e+08  1.320e+08  3.891e+09

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.068e+09  4.594e+08   4.502 0.000244 ***
date.analizate$FDI 6.993e+00  1.969e+00   3.551 0.002135 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.586e+09 on 19 degrees of freedom
Multiple R-squared: 0.3989,     Adjusted R-squared: 0.3672
F-statistic: 12.61 on 1 and 19 DF,  p-value: 0.002135
```

The regression equation can be written as:

$$\hat{y} = 2068327443{,}60 + 6{,}992669198 x_1$$

It can be noticed that GDP and FDI are positively correlated.

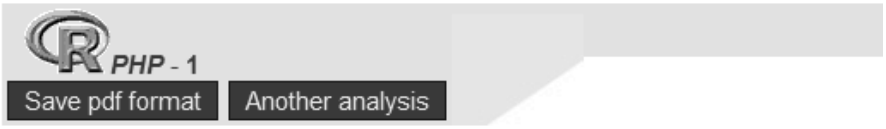
From the output, the coefficient of determination, $R^2 = 0.3989$, meaning that 39,89% of the variability of the GDP is explained by the foreign direct investments.

The code for the ANOVA analysis:

```
date.analizate <- read.table("a.tab.txt", header=TRUE)
anova(lm(date.analizate$GDP~date.analizate$FDI))
```

ANOVA analysis output

Figure 6




```
> date.analizate<-read.table("a.tab.txt", header=TRUE)
> anova(lm(date.analizate$GDP~date.analizate$FDI))
Analysis of Variance Table

Response: date.analizate$GDP
                   Df     Sum Sq    Mean Sq F value   Pr(>F)
date.analizate$FDI  1 3.1702e+19 3.1702e+19  12.607 0.002135 **
Residuals          19 4.7779e+19 2.5147e+18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

From the ANOVA table (table 6), we can establish the overall significance of the model. The p-value (significance F) is equal to 0,002135 0,002135, which is quite small and the null hypothesis can be rejected. The P value tests the null hypothesis that data from all groups are drawn from populations with identical means[1]. In our case the overall p-value is small, so we can reject the hypothesis that all the populations have identical means.

Also, more complex analysis of the data can be done, but the purpose of this research is to show some of the programming commands in R and the fact that R works via PHP.

R-PHP - second module – GUI

The second module is quite interesting because no programming skills in R are required. All of the commands that were ran in section 4.1. can be run

1. http://www.graphpad.com/guides/prism/6/statistics/index.htm?f_ratio_and_anova_table_%28one-way_anova%29.htm

in this module using the appropriate command buttons in the upper part of the Web page.

The only step that might cause some problems is the data preparation step. After uploading the file with data, the application becomes quite self-explanatory.

Depending on what data analysis the user needs, the proper commands from the menu should be chosen.

For a quick preview of this module see figure 3.

CONCLUSIONS

In the transition Republic of Moldova, as seen from the data presented in this article, the GDP over the 1992 and 2012 time period has an increasing trend. But an increasing GDP doesn't always mean a good life for everyone. The Republic of Moldova's biggest problem is that income is distributed unevenly.

The foreign direct investments are very important for a country economy because it can create jobs (reduce unemployment) and also increase productivity. In the Republic of Moldova's case, the FDI are not stable. From our data, it can be observed that there are some years with very high FDI and other years with low FDI.

From the linear regression model briefly presented, it can be observed that, in the Republic of Moldova, between GDP and FDI there is a positive relationship.

Using R via PHP, means that a user can access a Web page and run commands on data, without installing R on their computer. It also means that the user will use the hardware and software components of the server, where the application is stored, so if the application is configured on a powerful server, then the processing speed of R will significantly increase.

The R-PHP software tool developed by the Department of Statistical and Mathematical Sciences Silvio Vianelli of the University of Palermo (Italy) (with contact persons Alfredo Pontillo and Angelo Mineo) is definitely an eye-opener and a great software.

As further development, more commands in the user friendly area are suitable and a mobile friendly version of the application might be of interest.

References

1. Caragea, N., Alexandru, C. A., Dobre, A.-M., Bringing new opportunities to develop statistical software and data analysis tools in Romania, Published in: The Proceedings of the VIth International Conference on Globalization and Higher Education in Economics and Business Administration, 2012.
2. Dobre, A.-M., Gagiu, A., Manipulation of large databases with R, presented at EUB-2013 International Conference - Ecological University of Bucharest, 4-5 April, 2013.
3. http://www.php.net/manual/en/ref.stats.php
4. http://dssm.unipa.it/R-php/
5. Mineo, A., Pontillo, A., Using R via PHP for Teaching Purposes: R-php, Journal of Statistical Software, Volume 17, Issue 4, October 2006.
6. Agrawal, G., Khan, A., Impact of FDI on GDP: A Comparative Study of China and India, International Journal of Business and Management, Vol. 6, No. 10; October 2011.
7. http://www.php.net/manual/en/ref.trader.php