DE GRUYTER OPEN

STUDIA UNIVERSITATIS ECONOMICS SERIES
"Vasile Goldiş" Western University of Arad

DOI: 10.1515/sues-2016-0014

COMPARATIVE STUDY ON DIFFERENT TYPES OF REGRESSION APPLIED TO UNEMPLOYMENT IN MARAMURES COUNTY OF ROMANIA

Magnolia Tilca
"Vasile Goldis" Western University of Arad
e-mail: tilca.magnolia@uvvg.ro
Meda Bojor
"Gheorghe Sincai" National College Baia Mare

(Received May 2016; Accepted July 2016)

Abstract: We are studying the economic phenomenon of the unemployment in Maramures County of Romania. To obtain plausible conclusions regarding this study we apply different types of regression: the linear regression, polynomial regression, spline and B-spline regression. In this paper we focus on the numerical side of the research and we compare the predicted values, the graphic representation of the evolution, the future predictions and the errors generated by the regressions mentioned above. The calculations are performed in R, a programming language for statistical computing. An implementation in R is given.

Keywords: linear regression, splines regression, B-spline regression, R language, the evolution of the unemployment in Maramures
J.E.L. CODES: C68, R15

Introduction

National and County Employment Agencies (ANOFM and AJOFM), together with the National Statistical Institute of Romania (INSR), study the evolution of unemployment in Romania using statistical methods that compare the number of unemployed relative to unit time, relative to the region, educational studies, sex, age or unemployment rate. (For example BIM - International Labour Office - uses the *Holt method*; it uses exponential data series with a linear trend; it is the method of estimating unemployment measured according to the criteria of the International Labour Office (National Institute of Statistics, 2011)). We propose a study of econometric methods used in estimating and predicting the unemployment data in Maramures through the open-source mathematical-statistical programming environment R, namely the regression method.

The linear regression, polynomial regression, spline regression and B-spline regression are used. *The linear regression* is particularly popular due to its ease of use and determines the line which passes through or adjacent to data points, based on the principle of least squares. But the linear regression and *polynomial* *regression* are not smooth enough to model real situations. The s*pline regression* ensures flexibility in the estimation of the model. It is a mathematical-economic model adapted to changes in the economy which “keeps the continuity of the regression function, but splits the studied data in intervals with homogenous characteristics." (Geambasu et al, 2010). The *B-Spline regression* is also a flexible model that offers numerical stability to the algorithm and it is easy to implement.

The comparative study of the four variants of these unemployment progressions are applied over a period of time (years).

1. Literature review

It is known that the regression is an econometric technique which investigates the relationship between a dependent and independent variable, “used for forecasting, time series modeling and finding the causal effect relationship between the variables.” (Sunil, 2015) There are many forms of regressions. The challenge is to identify what regression best fits the problem investigated.

Linear regression was the first type of regression analysis and it is used widely in practical applications. The frequent economic changes highlight the necessity of adapting regression models to variation of the economy. This is the case of the spline regression model.

The spline regression is a relatively recently studied technique. In 1991, Friedman extended the recursive partitioning regression model of Morgan and Sonquist (1963) and Breiman, Friedman, Olshen and Stone (1984) by constructing a cumulative function which reunites the functions adapted to each sub-region, thus offering a better understanding of the evolution of the dependent variable. He applied the model to data sets from signal theory (numbers of shots), analytical chemistry (Portuguese olive oil) or artificial functions. Since then, specific economic problems were better solved by using the spline regression model. Thus Blindell, Chen and Kristensen (2007) studied the relation between the demand of goods and household budgets, using a continuous function on subsets of data resulted also from adding the subsets functions. Engle and Gonzalo (2008) studied the relation between the financial market and macroeconomic evolution using a GARCH-spline regression model for describing the trend with reduced frequency of macroeconomic variable volatility in time. Greiner (2009) used penalized splines to study the relation between the “primary surplus to GDP ratio and the debt ratio” (Greiner, 2009). Liu and Yang (2010) proposed the spline-backfiteed local linear procedure and applied it to a “varying coefficient extension of the Cobb-Douglas model for the US GDP that allows non neutral effects of the R&D on capital and labor as well as on the Total Factor Productivity.” Haupt, Kagerer and Steiner (2014) illustrate spline and B-spline regression in an empirical application using unit sales, retail prices and display activities on the store level from a food chain in Chicago. For the computations they used the open-source software R. Using spline regression and R, Shujie et al (2015) computed their results on GDP growth and OECD Status. The list doesn't stop here.

If the mentioned papers use R for the computation of the created spline regression models, the present paper reunites four of the regression models (linear, polynomial, slpine and B-spline) and implements an R-function which generates results specific to each regression. The function compares these results in order to help the users in applying the proper model. We illustrate the functionality of the implemented R-function using unemployment data from the Maramures County.

2. Methodology

One of the advantages of the R statistical environment is the fact that complicated mathematical definitions and formulae may be avoided. The R code doesn't require these.

The piecewise linear model is "a non-linearity captured by estimating a linear regression through several intervals" (Ruppert et al, 2003) and is given by the equation

$$y_i=a_i+b_ix+\varepsilon_i, \text{ if } x\in[t_{i-1},t_i],\ i=1,\ldots,k,$$

where a_i,b_i are the coefficients of the model.

But the regression function is not smoothed at the knots $t_1,\ldots,t_{k-1}$. (Ruppert et al, 2003) Spline is a smooth piecewise regression model with transition at the knots. The cubic spline regression model is a continuous piecewise function with continuous second order derivatives:

$$y=a+b_1x+b_2x^2+b_3x^3+\Sigma_{j=1,\ldots,k}b_{j+3}(x-t_j)_+^{\,3}+\varepsilon,$$

where the notation $(\cdot)_+$ is the mathematical notation for truncated power functions.

The truncated functions often generate unstable numerical results. (Ruppert et al, 2003) The truncated basis is replaced by the B-spline basis which has compact support (i.e. values are equal to 0 outside two adjacent knots).

B-spline is a spline model smooth at the boundary knots expressed as a combination of basic B-spline functions (cubic B-spline):

$$y=\Sigma_{i=1,\ldots,k}b_iB_{i,3}(x)+\varepsilon,\ x\in[a,b].$$

These B-spline basis functions $B_{i,3}$ can be determined from the recurrent formula of Carl de Boor (Carl de Boor, 1972); this recurrent formula makes the algorithm easy to implement.

We use the pre-defined R-functions for regressions:

Table 1. The R-functions used in regression models

model	R-function	Interpretation
linear	`lm(Y~X)`	the adequacy of the linear model having Y as the response variable and X as predictor
polynomial	`lm(Y~poly(.))`	Y is modeled linearly by $y=a+b_1x+b_2x^2$
spline	`lm(Y~ns(.))`	function to model Y by natural cubic spline function
B-spline	`lm(Y~bs(.))`	function to model Y through cubic B-spline function

Source: adapted from R library and (Paradis, 2013)

R is a free environment language and software for statistical calculation and graphics, developed since 1995 by *R*oss Ihaka and *R*obert Gentleman (hence the name R). It quickly gained the attention of programmers and statisticians becoming a powerful environment for statistical computing. R contains linear and non-linear statistical modeling techniques, statistical tests, techniques for linear time series analysis etc. R is an integrated suite of software facilities for data manipulation, calculation, graphics display that includes a simple and effective programming language, allowing statistical techniques to be implemented. R is a case-sensitive dialect of the S language (*S language* is a statistical programming language, developed since 1975, and updated to its modern version since 1988.). The code may be placed in the Command prompt window or may be used from a source file. There are a variety of data types including vectors (numeric, character, logic), matrices, lists and data windows. R's functionality comes from the basic functions found in the dedicated R packages or from complex functions created by the user.

R provides facilities for implementing regressions through regression functions which are stored in packages distributed with the installation of R. A characteristic feature of R is the regression model `lm(Y~model,data)`, where Y is the analyzed response and `model` is a set of terms for which some parameters will be estimated. For the linear regression the model is considered to be the set of independent variables X, the `poly(.)` model corresponds to the polynomial regression, the `ns(.)` model to the natural spline regression and the `bs(.)` model to the B-spline regression (see Table 1).

The source code function that includes these formulas was written in R, in an editor file (New Script) and saved as `Regression.R`. The input data of the function (which has the same name: `Regression`) is the set of independent variables X, the set of dependent variables Y and the p value for which the prediction is intended. The output data is data specific to each regression, mainly the predicted values and the graphs of the regressions. The interpretation of the results is reported in the last part of the algorithm, *Conclusions*. Main steps of the algorithm are described below.

Table 2. Other R-functions used in the algorithm

R-function	Used to
`cat("...")`	print the information on-screen
`cor.test(X,Y)`	test the degree of correlation between two variables
`predict(.)`	compute the estimated values for new data of the model
`plot(.)` and `abline(.)`	`abline` function plots the regression line y=a+bx
`data.frame(.)`	create a sequence of data saved as a table
`predict(.,interval =”predict”,level=.95)`	predict data with the 95% confidence
`summary(.)$r.squared`	extract the R-squares coefficient; it specifies how much the variation of `Y` is explained by the independent variable
`summary(.)$sigma`	extract the sigma value of the standard error; it specifies how much the average observed values deviate from the theoretical values
`summary(.)$fstatistic` and `pf(.)`	extract the p-value of the F distribution; this value determines whether the two variables are significantly different by checking if the coefficient β is nonzero

Source: adapted from R library

Data input: X, Y, p
Data output:
Step I. For the linear regression

I.1. the coefficient of correlation between `X` and `Y`
I.2. the linear regression coefficients `a`, `b`
I.3. the approximate values generated by the linear regression equation
I.4. errors (residue)
I.5. the predicted value for an independent variable `p`
I.6. the graph of the linear regression equation

Step II. For the polynomial regression

II.1. the quadratic polynomial coefficients
II.2. the predicted value for `p`
II.3. the predicted values generated by the polynomial regression
II.4. the errors (residue)
II.5. Plotting the polynomial curve

Step III. For the natural spline regression

III.1 the matrix of the cubic B-spline values for the natural spline function

`ns(.)` function generates the matrix of order `(card(X),df)` of the B-spline basis function for a natural cubic spline (`df=5`=>3 inner knots)

(`card(X)` is the cardinal of `X`, equivalent to the `length(X)` function in `R`, and `df` represents here the number of the degrees of freedom; for `ns(.)` function, the `df` is given by `df = internal knots no.+1+1`)

III.2. the natural cubic spline regression coefficients

III.3. the predicted value by natural spline regression for known `p`

III.4. the errors (residue)

III.5. the spline regression curve graph

`lines(.)` function takes the horizontal axis as the sequence `seq(.)` of 200 points lying between the extremes of set `X` and the vertical axix the set of the values predicted by the spline regression calculated in the abscissa points

Step IV. For the B-spline regression

IV.1. the matrix of the basic B-spline functions

`bs(.)` function generates the matrix of order `(card(X),df)` of the values of the B-spline basis functions for a polynomial spline function (cubic) (`df=5`=>1 inner knot) (We note that the number of the degrees of freedom `df` for the `bs(.)`function is given by `df = internal knots no.+1+degree of the spline function`)

IV.2. the cubic B-spline coefficients of the regression

IV.3. the predicted value of the B-spline regression for `p`

IV.4. the predicted values generated by B-spline regression for the observed variable X

IV.5. the errors (residue)

IV.6. the B-spline regression curve

Step V. Generating the conclusions

V.1. the values predicted by the 4 regressions for a given value

V.2. the prediction interval (The prediction interval represents all the predicted values between the observed values which are contained with a certain probability)

V.3. the errors generated by the 4 regressions

V.4. the R-Squared coefficient value for the 4 regressions

V.5. the standard error

V.6. the p-value

The applicability of this algorithm is highlighted by the study of regressions in the case of the unemployment data from Maramures County.

3. Main findings

The issue studied is the evolution of unemployed in Maramures County and its prediction in time. Yearly reports from the AJOFM database regarding the number of registered unemployed in 2000-2015 in Maramures County are used as statistical data. The method used here is the regression that uses the years 2000-

2015 as the independent variable and the number of the unemployed from Maramures as the dependent variable. It is known that the number of the unemployed is influenced by a series of factors such as the number of job offers, the GDP and so on. For the sake of simplicity, the comparative study is applied for the observations of the annual unemployed number. Thus the study underlines the unemployment evolution through the linear regression, polynomial, spline and B-spline regression, using the algorithm described above.

The first step is to write the observed data (using a comma separator for data) and to save it in a text document (Notepad, WordPad, MS Word, s.o.) as a simple `.txt` file.

The sequence of instructions (Fig.A.1) reads the data from the file: setting the access path to the directory where the `.txt` file was saved -> reading the data from the file (in this case `UnemloyedEvolution.txt`) in the mandatory variable `givendata` -> viewing the data by calling `givendata` variable -> reading the independent variable Year in the `X` mandatory variable -> reading the `unemployedNo` dependent variable in the `Y` mandatory variable -> displaying `X` and `Y`. The function `source(Regression.R)` calls the source code of the algorithm from the file `Regression.R`. The function `Regression(X,Y,2011)` applies the linear regression model, polynomial, spline and B-spline for the variables `X` and `Y`, and it performs the prediction for the independent variable 2011.

For each model the function returns the model coefficients, predicted/ theoretical values which are generated by the model for both an independent variable and for independent initial values, errors (residue) and the statistical parameters of the method (p-value, the multiple R-squared coefficient, the adjusted R-squared, the residual standard error, the number of degrees of freedom, the significance coefficients). Results can be viewed in steps I-IV (see Fig.A.2 for polynomial regression), but the comparison and interpretation may be made easier by using the graphical representation (Figure 1) and the conclusions contained in Step V. The table 3 summaries the results contained in conclusions.

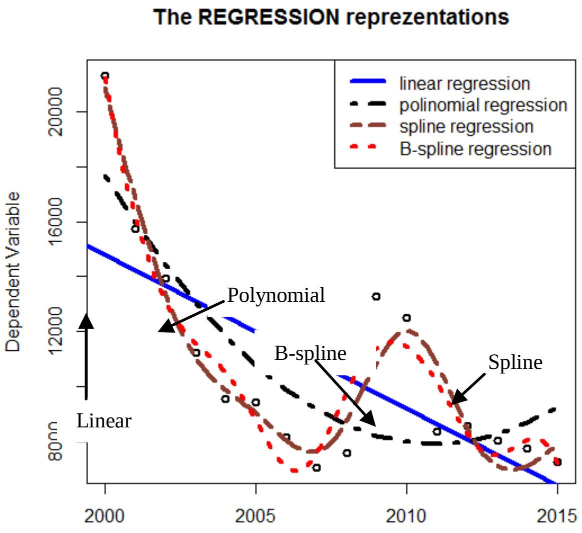
Table 3. The results generated by the algorithm in the case of the four regressions

Summary Regressions	Median Residual	Residual standard error	Multiple R-squared	Adjusted R-squared	Coefficients	Predicted value for 2011
Linear	368.2	2,950	0.4651	0.4269	Intercept: 1,131,383.2 b: --558.3	8,656
Quadratic Polynomial	710.2	2,539	0.6322	0.5756	Intercept: 3.302e+0.8 b1: -3.284e+0.5	7,921

Summary Regressions	Median Residual	Residual standard error	Multiple R-squared	Adjusted R-squared	Coefficients	Predicted value for 2011
					b2: 8.165e+0.1	
Cubic Spline	127.5	1,391	0.9236	0.8727	Intercept: -38,69 b1: 38,774 b2: 48,198 b3: 44,442 b4: 53,886 b5: 21,044 b6: 115,733 b7:lin.comb.	10,820
Cubic B-spline	52.95	1,673	0.9263	0.8158	Intercept: 21,242 b1: -4,467 b2: -9,715 b3: -10,093 b4: -16,700 b5: -8,888 b6: -9,810 b7: -15,565 b8: -12,130 b9: -14,017	10,144

Source: own results adopted from `Regression.R` function

Figure 1. The graphical representation of the regressions applied to the annual unemployed number in Maramures



Source: own results generated by `Regression.R` function

The graphical representation (Figure 1) shows that the B-spline and spline regressions approximate the observed data more accurately than the linear regression and polynomial regression (see also Fig.A.5 for generated results). This is primarily due to the definition of the spline function. In the case of the unemployment in Maramures, the nonlinear manner of distribution of the observed data makes linear and polynomial regression to be discredited in the face of spline and B-spline regressions.

The theoretical values for periods of time within the 2000-2015 range differ as precision from one model to another. Thus, for a theoretical value for the number of unemployed in 2011, the linear regression is the method that sets the minimum error (Table 1 or Fig.A.3 for generated results) and for 2013 the optimum theoretical value is given by the spline regression. This result can be observed also from the analysis of the model that generates the lowest residue, as it is shown in conclusion 3 (Fig.A.5).

With a probability of 95%, the number of unemployed in 2011, given by the linear regression, can oscillate between 2,024 and 15,289, a relatively broad range compared to the spline regression (Table 4 and Fig.A.4 for generated results).

Table 4. The predicted intervals (with a probability of 95%) in which the theoretical value in 2011 is located

Regression	Prediction interval		p-value
	min	max	
Linear	2,024	15,289	0.00361301
Polynomial	2,136	13,707	0.00150077
Spline	7,190	14,449	0.00014791
B-spline	5,071	15,217	0.00879304

Source: own results adopted from `Regression.R` function

Within the studied phenomenon, the comparative analysis of the residue shows that the spline regression model is valid. In order to identify which method can be used to a greater extent forecasters, the R-squared is compared. The B-spline regression can be used in predictions with 92.63% confidence interval (see Table 1 and Fig.A.6 for generated results).

The standard errors of the studied phenomena are relatively high in all four estimates (Table 1 and Fig.A.7 for generated results), which means that the average deviation of the observed values compared to the theoretical ones is high.

The predicted result with the lowest standard deviation of the estimation is given by the splines regression and has an error of ± 1,390 unemployed.

The results investigated so far raise the question: is there a significant relationship between the amount of time (years) and the number of the unemployed? The null hypothesis test β = 0 leads to p-values that are smaller than 0.05, so the null hypothesis is rejected in all four cases, with a high degree of trust in the case of the spline regression (p-value=0.00014791). See Table 4 and Fig.A.8.

The intensity of the relationship between the two variables `X` and `Y` is given by the correlation coefficient of linear regression model, namely -0.6819874, the correlation being a relatively good, negative one, which emerges from the trend of the decreasing number of the unemployed.

For a future prediction, the function `Regression(X,Y,2016)` returns the predictions from Fig.A.9. and Table 5. The polynomial regression and B-Spline regression in the near future indicate an increase, and the linear regression and spline regression indicate a decrease. For immediate future predictors, the spline and B-spline regression predictions will take extremely large values, or even negative ones, which confirms the known outcome that the spline functions are used in interpolation problems rather than extrapolation. The spline regressions are useful for predictions in which the unknown `X` values are within the range of the independent variables.

Table 5. The predicted regressions value for x=2016

Regression	linear	polynomial	spline	B-spline
Predicted value for p=2016	5,865	10,029	8,805	1,679

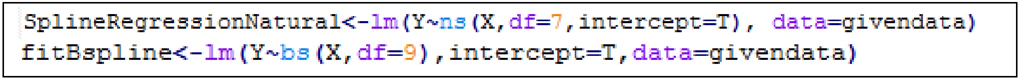
Source: own results adopted from `Regression.R` function

4. Discussions

By applying the linear regression, polynomial regression, spline regression and B-spline regression to the real model of the unemployed number in Maramures, the generated results may be interpreted as follows:

1. The validity of the model: the spline regression and B-spline regression models are valid because they have the regression line closer to the observed values. Their validity alternates depending on the choice of the model's degree of freedom, i.e. the number of internal knots. The function implemented in `R`, `Regression`, uses the same number of intermediate knots (five) for both models. For one internal point (`df = 3`, respectively `df = 5`), the B-spline regression is valid.

Figure 2. The code of the spline regression and B-spline regression with five internal knots (`df=7`, respectively `df=9`)



```
SplineRegressionNatural<-lm(Y~ns(X,df=7,intercept=T), data=givendata)
fitBspline<-lm(Y~bs(X,df=9),intercept=T,data=givendata)
```

Source: author's view

2. The relevance of the equation: The spline and B-spline regressions are the models which have the estimated equations with a high degree of confidence; these models alternating if a different number of degree of freedom is chosen.
3. The level of significance: all four regressions state that the relationship between the number of unemployed and the period of time is significant, with greater significance in the case of the spline regression.
4. The residue: the spline and B-spline regressions ensure a variation of the theoretical dependent errors which is smaller than the observed dependent one.
5. The predictions: the spline and B-spline regressions are more accurate in the case of predictions for independent values `X` within the range of observations, for example if the number of unemployed in a given year located between 2000 and 2015 is not known. To follow the future evolution of the unemployment, the spline and B-spline regressions tend to generate false values, unstable results outside the range of given values.
The spline and B-spline regressions prove a higher tracking power of trends resulted from empirical data, and provide better tracking of the evolution of data over time. Although the spline regression is a useful approach in describing the economic evolution of the studied event, there are pros and cons, depending on the researched topic and the desired results.
The researcher can apply the implemented function `Regression.R` to his own phenomena to generate the regression results.

Conclusions

Correct assessment of the evolution of annual unemployment is a constant concern for economists. Regressions are relatively easy to create and apply models giving very good results in estimations.
In this paper we apply regression techniques in employment analysis and we compare the results in the case of Maramures County.
The dataset comes from AJOFM and consist of the last 16 years (2000-2015) and the number of unemployed corresponding to these years. From this dataset we observe that after an increase of the number of unemployed in the crisis years (2009-2010), the number of unemployed decreases in time (2011-2015).
The graphical representation shows that the data is oscillating.
A linear or polynomial regression will not efficiently generate the evolution of the unemployment. Only in 2002 and 2011 are the estimations of the unemployed number closer to the observed values. The tails are less reliable than the central portion.
A spline and B-spline representation will fit the given data more accurately due to their flexibility in knots. The B-spline regression provides the best information

about the evolution of the unemployed because B-spline knots are not restricted at the ends.

For investigations which include future prediction of the unemployment, we can use linear regression. In 2016 the number of the unemployed is estimated to be 5865. For investigations which include the evolution of the unemployment in past years, we can use spline and B-spline regression because they fit well the observed data. Setting five interior knots automatically selected by `R`: 2002.5, 2005, 2007.5, 2010, 2012.5, the entire interval is split into 6 sub-regions which can offer a better analysis of the unemployment trends in Maramures. Applying `Regression(X,Y,p)` with `p=2012,2012.1,…,2012.5` we obtain the unemployment evolution in the fifth sub-region.

Thus, the implemented function `Regression` unifies the regression results. Its application only requires entering data and using several commands calling the function `R`. The graphic helps with observing the phenomenon. The challenge lies in interpreting the results. Spline and B-spline regressions explain the phenomenon within X very well and prove a higher power of tracking the trends.

References

1. Blundell, R., Chen, X. K. (2007). Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves. *Econometrica, Vol. 75, No. 6* , pp. 1613-1669.
2. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and regression trees. *Wadsworth and Books/Cole, Belmont, CA.*
3. de Boor, C. (1972). On calculating with B-splines. *J. Approx. Theory, 6* , pp 50-62.
4. Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics, Vol. 19, No. 1* , pp. 1-67.
5. Geambasu, L., Jianu, I., Geambasu, C. (2010). Spline linear regression used for evaluating financial assets. *Analele Universitaţii "Constantin Brâncusi" din Târgu Jiu, Seria Economie, Nr. 4/2010* , pp. 310-321.
6. Greiner, A. (2009). Estimating penalized spline regressions: theory and application to economics. *Applied Economics Letters, 16* , pp. 1831-1835.
7. Haupt, H., Kagerer, K., Steiner, W. (2014). Smooth quantile bases modeling of brand sales, price and promotional effects from retail scanner panels. *Journal of Applied Econometrics, DOI 10.1016/j.jeconom.2014.06.003*.
8. Liu, R., Yang, L. (2010). Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory 26* , pp. 29-59.
9. Ma, S., Racine, J., Yang, L. (2015). Spline regression in the presence of categorical predictors. *J. Appl. Econ. 30* , pp. 705-717.
10. Morgan, J. N., Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc. 58* , pp. 415-434.

11. National Institute of Statistics. (2011). *Unemployment BIM methodology on a monthly basis.* Retrieved from www.insse.ro: www.insse.ro/.../Metodologia_rata%20somaj%20 BIM%20lunar_rev.doc. Accessed 6 April 2016.

12. Paradis, E. (2013). *R for beginners*. Retrieved from https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pd. Accessed 6 Aplir 2016.

13. *Cambridge Series in Statistical and Probabilistic Mathematics, volume 12, Cambridge University Press*.

14. Ruppert, D., Wand, M., Carroll, R. (2003). *Semiparametric regression.* New York: Cambridge University Press.

15. Sunil, R. (2015). *7 tipes of regression techniques you should know!* Retrieved from Analytics Vidhya: http://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/. Accesed 12 April 2016.

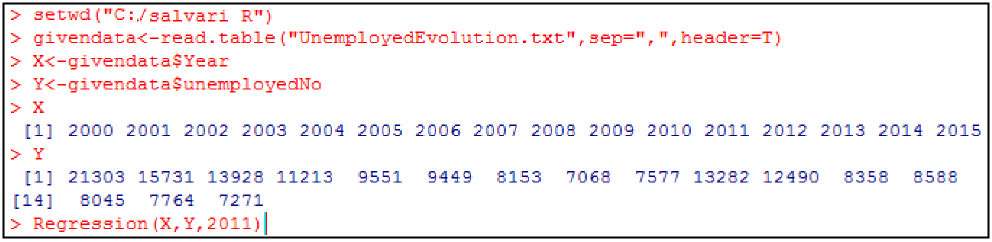
16. Vlad, C., Brezeanu P., Fiscality – a relevant factor influencing regional development in Romania and the European Union, *Studia Universitatis "Vasile Goldis" Arad Economics Series, Vol. 26, Issue 2/2016,* pp 48-62.

17. Wu, W. (2009). *An application of spline regression to dose-response analysis in observational study*. Retrieved from https://medschool.vanderbilt.edu/cqs/files/ cqs/media/2009Jan16William.pdf. Nashville - Vanderbilt University Medical School, USA. Accesed 13 April 2016.

Appendices

Appendix A. The interpretation of the results generated by the implemented R-function `Rgegression.R`

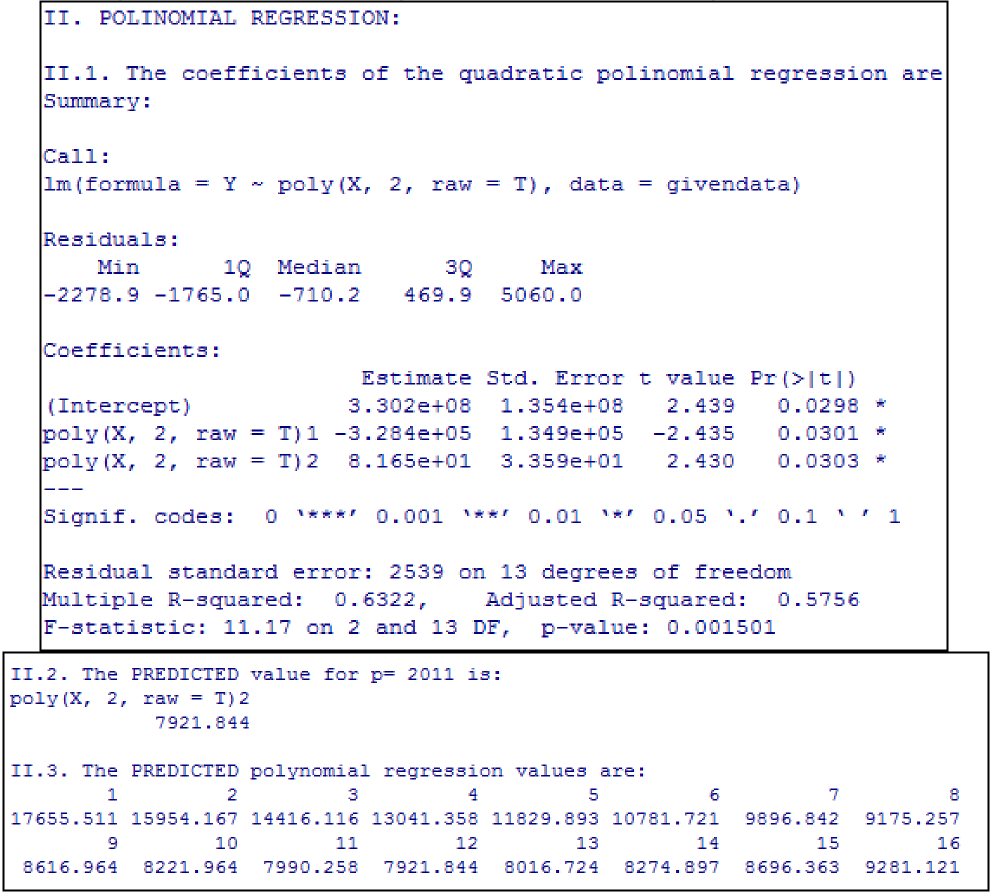
Fig.A.7. The steps of reading the yearly unemployed number from the file `UnemployedEvolution.txt`



```
> setwd("C:/salvari R")
> givendata<-read.table("UnemployedEvolution.txt",sep=",",header=T)
> X<-givendata$Year
> Y<-givendata$unemployedNo
> X
 [1] 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
> Y
 [1] 21303 15731 13928 11213  9551  9449  8153  7068  7577 13282 12490  8358  8588
[14]  8045  7764  7271
> Regression(X,Y,2011)
```

Source: own results generated by `Regression.R` function

Fig.A.8. The polynomial regression results for assessing unemployment

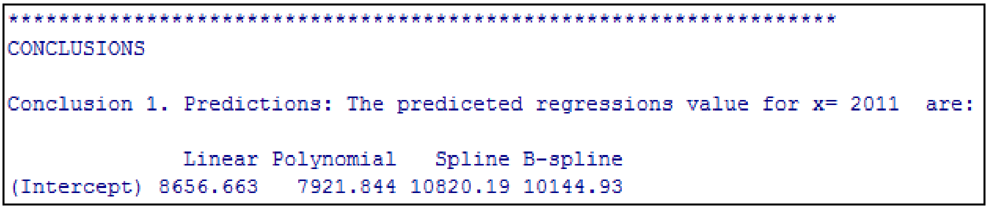


```
II. POLINOMIAL REGRESSION:

II.1. The coefficients of the quadratic polinomial regression are
Summary:

Call:
lm(formula = Y ~ poly(X, 2, raw = T), data = givendata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2278.9 -1765.0  -710.2   469.9  5060.0 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)  
(Intercept)           3.302e+08  1.354e+08   2.439   0.0298 *
poly(X, 2, raw = T)1 -3.284e+05  1.349e+05  -2.435   0.0301 *
poly(X, 2, raw = T)2  8.165e+01  3.359e+01   2.430   0.0303 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2539 on 13 degrees of freedom
Multiple R-squared:  0.6322,    Adjusted R-squared:  0.5756 
F-statistic: 11.17 on 2 and 13 DF,  p-value: 0.001501
```

```
II.2. The PREDICTED value for p= 2011 is:
poly(X, 2, raw = T)2 
            7921.844 

II.3. The PREDICTED polynomial regression values are:
        1         2         3         4         5         6         7         8 
17655.511 15954.167 14416.116 13041.358 11829.893 10781.721  9896.842  9175.257 
        9        10        11        12        13        14        15        16 
 8616.964  8221.964  7990.258  7921.844  8016.724  8274.897  8696.363  9281.121 
```

Source: own results generated by `Regression.R` function

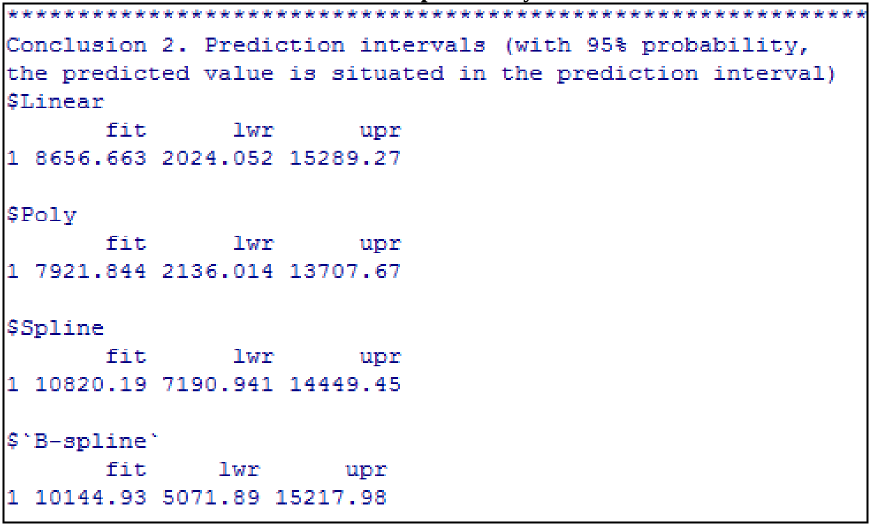
Fig.A.3. The predictions generated by the four regressions for the independent variable x = 2011



```
**********************************************************************
CONCLUSIONS

Conclusion 1. Predictions: The prediceted regressions value for x= 2011  are:

              Linear Polynomial   Spline B-spline
(Intercept) 8656.663   7921.844 10820.19 10144.93
```

Source: own results generated by `Regression.R` function

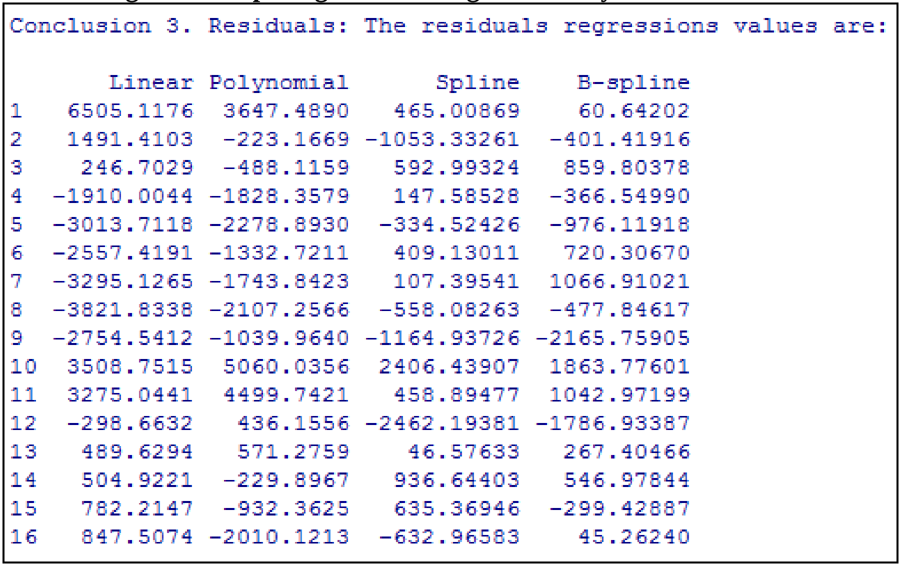
Fig.A.4. The predicted intervals in which theoretical value in 2011 is located, with a probability of 95%



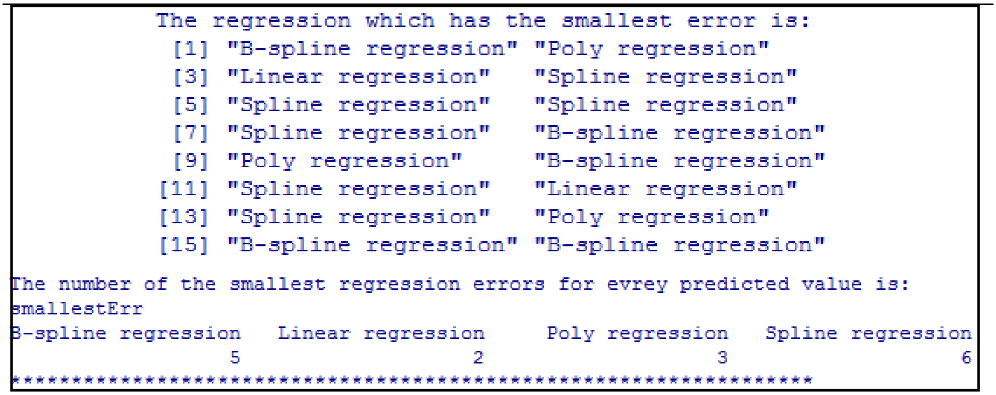
```
*************************************************************
Conclusion 2. Prediction intervals (with 95% probability,
the predicted value is situated in the prediction interval)
$Linear
       fit      lwr      upr
1 8656.663 2024.052 15289.27

$Poly
       fit      lwr      upr
1 7921.844 2136.014 13707.67

$Spline
       fit      lwr      upr
1 10820.19 7190.941 14449.45

$`B-spline`
       fit     lwr      upr
1 10144.93 5071.89 15217.98
```

Source: own results generated by `Regression.R` function

Fig.A.5. Comparing the residue generated by the four models



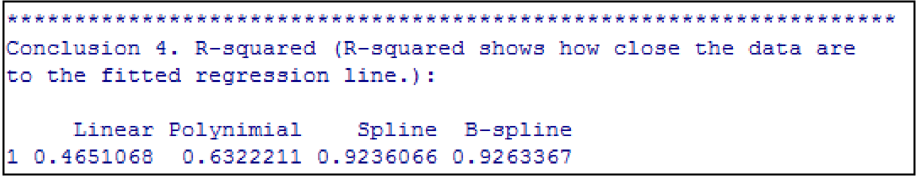
```
Conclusion 3. Residuals: The residuals regressions values are:

       Linear Polynomial       Spline     B-spline
1   6505.1176  3647.4890   465.00869    60.64202
2   1491.4103  -223.1669 -1053.33261  -401.41916
3    246.7029  -488.1159   592.99324   859.80378
4  -1910.0044 -1828.3579   147.58528  -366.54990
5  -3013.7118 -2278.8930  -334.52426  -976.11918
6  -2557.4191 -1332.7211   409.13011   720.30670
7  -3295.1265 -1743.8423   107.39541  1066.91021
8  -3821.8338 -2107.2566  -558.08263  -477.84617
9  -2754.5412 -1039.9640 -1164.93726 -2165.75905
10  3508.7515  5060.0356  2406.43907  1863.77601
11  3275.0441  4499.7421   458.89477  1042.97199
12  -298.6632   436.1556 -2462.19381 -1786.93387
13   489.6294   571.2759    46.57633   267.40466
14   504.9221  -229.8967   936.64403   546.97844
15   782.2147  -932.3625   635.36946  -299.42887
16   847.5074 -2010.1213  -632.96583    45.26240
```



```
            The regression which has the smallest error is:
             [1] "B-spline regression" "Poly regression"
             [3] "Linear regression"   "Spline regression"
             [5] "Spline regression"   "Spline regression"
             [7] "Spline regression"   "B-spline regression"
             [9] "Poly regression"     "B-spline regression"
            [11] "Spline regression"   "Linear regression"
            [13] "Spline regression"   "Poly regression"
            [15] "B-spline regression" "B-spline regression"

The number of the smallest regression errors for evrey predicted value is:
smallestErr
B-spline regression   Linear regression     Poly regression   Spline regression
                  5                   2                   3                   6
******************************************************************
```

Source: own results generated by `Regression.R` function

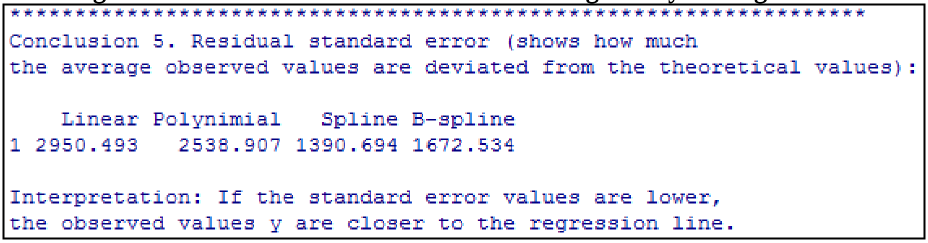
Fig.A.6. The R-squared in the case of regressions for the annual number of unemployed



```
******************************************************************
Conclusion 4. R-squared (R-squared shows how close the data are
to the fitted regression line.):

     Linear Polynimial    Spline  B-spline
1 0.4651068  0.6322211 0.9236066 0.9263367
```

Source: own results generated by `Regression.R` function

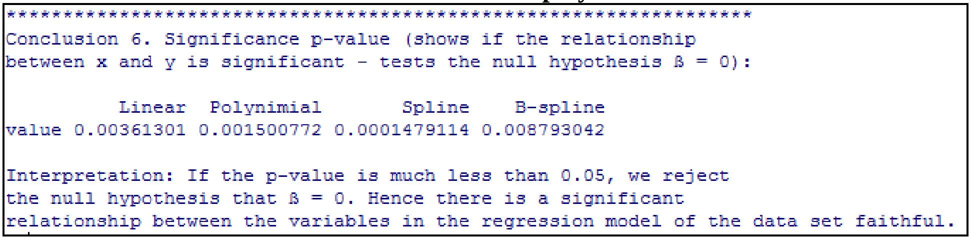
Fig.A.7. The standard errors of the estimates given by the regressions



```
******************************************************************
Conclusion 5. Residual standard error (shows how much
the average observed values are deviated from the theoretical values):

    Linear Polynimial   Spline B-spline
1 2950.493   2538.907 1390.694 1672.534

Interpretation: If the standard error values are lower,
the observed values y are closer to the regression line.
```

Source: own results generated by `Regression.R` function

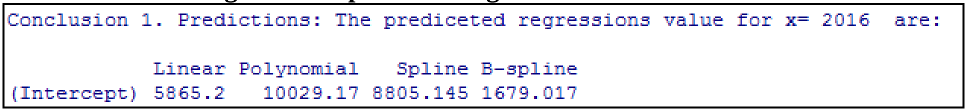
Fig.A.8. The exact level of significance given by the regressions regarding the yearly number of unemployed



```
******************************************************************
Conclusion 6. Significance p-value (shows if the relationship
between x and y is significant - tests the null hypothesis ß = 0):

          Linear  Polynimial       Spline    B-spline
value 0.00361301 0.001500772 0.0001479114 0.008793042

Interpretation: If the p-value is much less than 0.05, we reject
the null hypothesis that ß = 0. Hence there is a significant
relationship between the variables in the regression model of the data set faithful.
```

Source: own results generated by `Regression.R` function

Fig.A.9. The predicted regressions value for x=2016



```
Conclusion 1. Predictions: The prediceted regressions value for x= 2016  are:

            Linear Polynomial    Spline B-spline
(Intercept) 5865.2    10029.17 8805.145 1679.017
```

Source: own results generated by `Regression.R` function

Apendix B. Sample R code

The source code of the linear regression computation - excerpt

```
Regression<-function(X,Y,p)
{
library(splines)
cat("##################################################","\n")
cat("0. DATA INPUT:","\n")
cat("X=",X,"\n")
cat("Y=",Y,"\n")
cat("p=",p,"\n")
cat("\n")
cat("##################################################","\n")
cat("I. LINEAR REGRESSION:","\n")
cat("\n")
cat("I.1. The linear regression COEFFICIENTS are: \"Intercept\"&\"X\"","\n")
CoefLinearRegression<- lm(Y~X)
print(CoefLinearRegression)
cat("Summary:","\n")
print(summary(CoefLinearRegression))
cat("\n")
cat("I.2. The PREDICTED linear regression values are:","\n")
LinearRegression<-predict(CoefLinearRegression)
print(LinearRegression)
cat("\n")
cat("I.3. The PREDICTED linear regression value for x=",p," is:","\n")
ylinearRegression<-
CoefLinearRegression$coefficients[1]+CoefLinearRegression$coefficients[2]*p
print(ylinearRegression)
cat("\n")
cat("I.4. The linear regression RESIDUALS are:","\n")
ResidualLinear<-Y-LinearRegression
print(ResidualLinear)
PlotDataLinear<-plot(X,Y, main="The REGRESSION reprezentations",
xlab="Independent Variable", ylab="Dependent Variable")
points(X,Y,pch=20)
PlotLinearRegression<-abline(CoefLinearRegression,col=28)
cat("\n")
cat("################################################","\n")
…
```

The source code for computation of the residuals and its frequency - excerpt

```
…
cat("Conclusion 3. Residuals: The residuals regressions values are:","\n")
cat("\n")
for(i in 1:length(X)){
ResidualLinear[i]<-Y[i]-LinearRegression[i]
residualP[i]<-Y[i]-ypolyRegression[i]
residualS[i]<-Y[i]-ysplineRegression[i]
residualBS[i]<-Y[i]-yBsplineRegression[i] }
table2<-data.frame(ResidualLinear,residualP,residualS,residualBS)
colnames(table2) <- c("Linear","Polynomial","Spline", "B-spline")
print(table2)
#the smallest positive error:
mini<-numeric()
for (i in 1:length(X)){
mini[i]<-
min(abs(ResidualLinear[i]),abs(residualP[i]),abs(residualS[i]),abs(residualB
S[i]))}
#print(mini)
cat("\n")
cat("The regression which has the smallest error is:","\n")
smallestErr<-vector()
for (i in 1:length(X)){
if (mini[i]==abs(ResidualLinear[i])) {smallestErr[i]<-("Linear regression")}
else if (mini[i]==abs(residualP[i])) {smallestErr[i]<-("Poly regression")}
else if (mini[i]==abs(residualS[i])) {smallestErr[i]<-("Spline regression")}
else if(mini[i]==abs(residualBS[i])) {smallestErr[i]<-("B-spline
regression")}
}
print(smallestErr)
…}
```