Economic Insights – Trends and Challenges	Vol.VI(LXIX) No. 1/2017	41 - 50

Classic and Modern in Regression Modelling

Cristian Marinoiu

Petroleum-Gas University of Ploieşti, Bd. Bucureşti 39, 100680, Ploieşti, Romania
e-mail:marinoiu_c@yahoo.com

Abstract

Regression models are one of the most important sections of the classical mathematical statistics, theoretical results obtained in this area are truly impressive. At the same time, their area of applicability is very wide. It is a much known fact that technical and economic sciences, sociology, biology, psychology, genetics are just a few examples that benefit from the advantages of the regression modelling. In this paper we review some layouts of machine learning inspired by the regression modelling with the intention of highlighting the extraordinary contribution of this concept of mathematical statistics and creating one of the most interesting branches of modern computer science, called data science.

Keywords: *regression; regularization; Ridge regression; LASSO; elasticNet; PCR; PLS.*

JEL Classification: *C51; C40.*

Regression Problem and Regression Models

Let X and Y be two random variables. We want to find a function $f(x)$ so that the variable Y can be approximated by$f(X)$. A measure of this approximation is given by the mean squared error (Mean Squared Error), meaning$MSE\big(f(X)\big) = E(Y - f(X))^2$, where by E(X) we denote the mean of the random variable X .

The best choice is the function $f(x)$ which minimizes the above relation. In (Hastie, Tibshirani and Friedman 2009, p.18) it is shown that the function which minimizes $MSE(f(X))$ is the conditional mean of Y given $X = x$, meaning $f(x) = E(Y/X = x)$, also known as the regression function. In practice, finding the regression function $f(x)$ starting from its definition is difficult because, very often, we do not know the conditional probability density function $g(y/x)$ of Y given $X = x$ (Marinoiu, 2015, p. 21). At the same time, even if the theory gives us the possibility to estimate the regression function $f(x)$, starting from n available observations $(x_i, y_i), i = 1, \cdots n$ for the variables X, Y, these estimations do not present a practical interest, because of the low convergence rate or of the too higher variability of the estimated function (Hastie, Tibshirani and Friedman 2009, p.18). A viable alternative to estimate the regression function is the regression model.

Linear Regression Models

The most known and used regression models are linear regression models. For the simple linear regression model it is assumed that the relation between the random variables Y and X can be expressed by a linear relation as (Montgomery, Peck and Vining, 2012):

$Y = \beta_1 + \beta_2 X + \varepsilon,$

where the parameters $\beta_1$and β_2 are unknown parameters, and ε este is the additive error of the model.The natural generalization of the simple linear regression model is the multiple linear regression model, which assumes that between the random variable Y (also called dependent or response variable) and the random variables $X_{1,}X_{2,\dots,}X_p$ (also called independent, regressor, predictor or explanatory variables) there is a linear relation in the unknown parameters $\beta_{1,}\beta_{2,}\cdots,\beta_p$ under the form:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

where ϵ is the additive error of the model.

Assuming that for the variables (Y , $X_{1,}X_2,\dots X_p$) n observations $y_i,x_{i,1}x_{i,2},\cdots,x_{i,p}$ are available, the relation (1) can be written:

$$y_i = \sum_{j=1}^{p} \beta_i x_{i,j} + \varepsilon_i,\ i = 1,2,\cdots,n \quad (2)$$

The relation (2) can be expressed in a more convenient way in matrix notation, meaning:

$$y = X\beta + \varepsilon\ , \quad (3)$$

where:

y is a $nx1$ vector which contains the values of the n observations y_i of the dependent variable X is a nxp matrix also called design matrix. The column j represents the vector which contains n observations $x_{i,j}$ of the regressor variable X_j . Because the considered model is with interception, the first column of the matrix is a vector with all n components equal to 1;

β is the $px1$ vector of the unknown parameters $\beta_{1,}\beta_{2,}\cdots,\beta_p$;

$\mathcal{E}$ is the $nx1$ vector of the errors.

Solving the Linear Regression Model by the Least Squares Method

The most used method in order to estimate the unknown parameter β is the least squares method. According to this method the least squares estimator $\hat{\beta}$ of the parameter β is the value which minimizes the sum of the squares errors $S(y,\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \|\epsilon\|^2 = \|y - X\beta\|^2$, where $\|x\|$ is the euclidean norm of the vector x.

In short, the problem to be solved is $\min_\beta \|y - X\beta\|^2$,

and the solution to the above problem is

$$\hat{\beta} = arg\ \min_{\beta} \|y - X\beta\|^2$$

The minimum problem has an unique solution if and only if the columns of the matrix X are linearly independent or, equivlently, the linear regression model (3) verifies the assumption

H1: $p \le n$ şi $rang(X) = p$ (matrix X has maximum rank).

In this case, denoting by X^T the transposed matrix X, we obtain $\hat{\beta}$ as solution of the system of normal equations

$$X^T X\beta = X^T y \tag{4}$$

namely $\hat{\beta} = (X^T X)^{-1} X^T y$.

The estimator $\hat{\beta}$ is unbiased, meaning $E(\hat{\beta}) = \beta$.

The multiple linear regression model can be solved both analytically, by the least squares method (as above) and by using the gradient descendent method.

If, in addition, the following assumption is true

H2: the errors are uncorrelated, with zero mean and constant variance σ^2 (the homoscedastic case),

an unbiased estimator of the unknown variance σ^2 is $\hat{\sigma}^2 = S(y, \hat{\beta})/(n-p)$.

The central result of the linear regression models is given by the Gauss-Markov theorem, which shows that in conditions of respecting the assumptions H1 and H2, in the class of the unbiased and linear estimators in the observations $y_1, y_2, \cdots, y_n$, the least squares estimator $\hat{\beta}$ is optimal, meaning that it has the lowest variance and it is unique with this property (Cornillon and Lober , 2007). It's said that the least squares estimator $\hat{\beta}$ is BLUE (the Best Linear Unbiased Estimator) and we will note it by $\hat{\beta}^{BLUE}$.

Using Linear Regression Model for Prediction

Let be x_0 any point different from the observed values of the regressor variables $X_1, X_{2,\cdots,} X_p$ and y_0 the true but unknown value predicted in this point for the dependent variable Y, so that the assumptions of the proposed regression model are verified. Therefore,

$y_0 = x_0^T \beta + \varepsilon_0$, ε_0 being the error.

The prediction obtained with the help of the regression model in the point x_0 is $\hat{y}_0 = x_0^T \hat{\beta}$

The prediction error is defined by $PE = E(y_0 - \hat{y}_0)^2$ and can be written under the form (Hastie, Tibshirani and Friedman 2009, p. 223):

$$PE = Bias^2(\hat{y}_0) + Var(\hat{y}_0) + \sigma^2 \tag{5}$$

where $Bias(\hat{y}_0)$ =E$(\hat{y}_0) - y_0$, $Var(\hat{y}_0) = E\left(\hat{y}_0 - E(\hat{y}_0)\right)^2$, σ^2 is the irreducible error of the model (the variance of the errors of the linear regression model).

Given that, in addition to the assumptions H1, H2, the following assumption is also respected

H3: the errors ε_i are normally distributed with zero mean and constant variance σ^2, $(\epsilon_i \sim N(0, \sigma^2))$,

a $100(1-\alpha)$ percent prediction interval can be built for the predicted value y_0

$$\left(\hat{y}_0 - \hat{\sigma} t_{n-p;1-\alpha/2}\sqrt{x_0^T (X^T X)^{-1} x_0 + 1} \le y_0 \le \hat{y}_0 - \hat{\sigma} t_{n-p;1-\alpha/2}\sqrt{x_0^T (X^T X)^{-1} x_0 + 1}\right),$$

where $t_{n-p;1-\alpha/2}$ is the $1-\alpha/2$ quantile of Student's t distribution with $n-p$ degrees of freedom (Montgomery, Peck and Vining, 2012, p.104).

Bias-Variance Tradeoff

In the case of a supervised learning model, such as the regression model, it is necessary to ask two questions:

1. to what extent the proposed model captures the basic trend of training data? (is the proposed model adequate for the available training data?)
2. to what extent the model produces a quality prediction for new data that were not used in its construction? (does the proposed model have the power to generalize?)

We can answer these two questions by analysing carefully the relation (5) which represents the decomposition formula of the prediction errorPE. The dependence of the prediction error PE as well as of the values $Bias^2(\hat{y}_0)$ and $Var(\hat{y}_0)$ on the model complexity is suggestively illustrated in figure 1. The conclusion is obvious: the models which have small complexity are characterized by a high deviation $Bias^2(\hat{y}_0)$ and a small variance $Var(\hat{y}_0)$ and those of a big complexity are characterized by a small deviation and a high variance. So, it is impossible to have a model which has simultaneously a small bias and a small variance. From the theoretical point of view, the optimal level for the complexity of a model is given by that level for which the increasing of the variance is equivalent to the decreasing of the squared deviance (Fortmann-Roe, S., 2012). Basically a model which has the level of complexity much higher than this optimal level is affected by the phenomenon of over-fitting and those for which the level of complexity is far below that optimal level are affected by the phenomenon of under-fitting. An over-fitted model represents the training data very faithfully, reflecting also minor issues, nonessential data (noise) and, on the contrary, an under-fitted model is not adequate enough for the training data. Because on both extreme situations the model does not allow quality predictions it is necessary to realize a compromise between bias and variance. Since there are no analytical methods for the determination of an optimal level of complexity, in practice, it is approximated by the minimum point of the total prediction error PE .

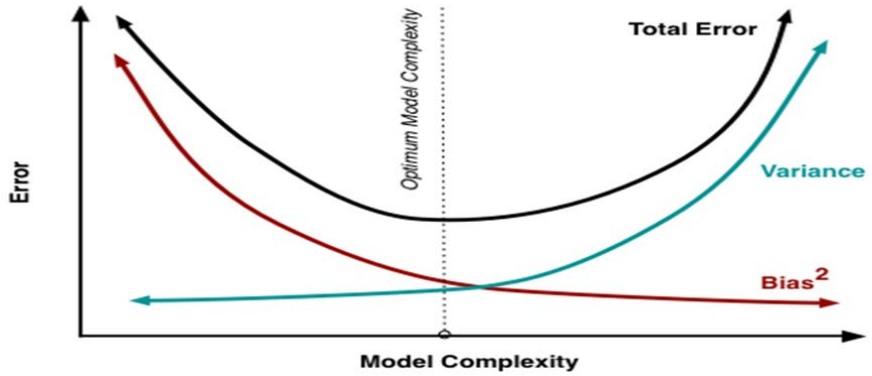

Fig.1. The graphical representation of the curves PE (Total Error) , $Bias^2(\hat{y}_0)$ and $Var(\hat{y}_0)$ in relation with the complexity of the models

Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

In recent years, the study of the regression models with a high degree of complexity, in which, typically, the number of variables p exceeds the number of records, n has seen a growing interest. This situation is frequently encountered, for example, in modelling problems related to image processing and in genetics (microarray analysis). These real problems have spurred new research that led to the development of performing regression algorithms adapted to new challenges. The idea that a good model is the model that succeeds (Biernat and Lutz, 2016) "to preserve the maximum of information contained in data with a minimum of variables" is generally accepted. Reducing the number of explanatory variables of a large regression model

is an important goal in modelling because it avoids nuisance phenomena such as multicolinearity and over-fitting and, at the same time, it leads to a model easily to interpret.

Examples of techniques that can accomplish this goal include:

- selecting a limited number of important variables using classical statistical techniques: forward selection, backward selection, stepwise selection, all possible regressions, all based on statistical tests;
- using new "artificial variables" as a linear combination of the original variables of the model;
- selecting a limited number of important variables using regularized regression models.

In the following sections we make a brief presentation of the methods listed above.

The Selection of „The Best" Regression Model

By selecting only some of the predictor variables that we have available we obtain a submodel with a prediction variance lower than that calculated for the initial regression model, but, at the same time, with a higher bias (Miller, 1990 p. 6). In this sense, a „good" model must be chosen so as to achieve a reasonable compromise between the bias and prediction variance. Among the classical solutions developed in this direction we mention the following methods: all possible regressions, forward selection, backward selection, stepwise selection (Afifi and Azen, 1972, p.129; Draper and Smith, 1966, p.162)).

All possible regressions method generates all $2^p - 1$ submodels, p being the number of predictor variables of the initial model, and orders them according to different criteria., among which we may place adjusted R^2 or C_p statistics (Javaras and Vos, 2002). For example , a „good" model can be chosen between those which have a high adjusted R^2value. As we can see, for each new variable introduced in the model, the computing effort doubles, the result being a very high computing cost for $p > 10$.

The *backward selection method* considers, initially, all potential variables included in the model. Then, at each step of the algorithm, the variable in the model whose elimination would get the lowest increase in the Residual Sum of Scores (RSS) is detected. We remove this variable only if the RSS increase is not "too high", otherwise the removing can be blocked. The process continues until a single variable remains in the model or until the stopping criterion is accomplished. When the set of the initial variables is large and we want to select a model with a small number of variables, the high computing time in order to perform the selection limits the usefulness of the method. An alternative method of backward selection may be the *forward selection method*.

The *forward selection method* considers that the initial set of chosen predictor variables is empty. At each step of the algorithm the variable whose insertion in the model would produce the lowest increase of RSS value, is detected, from the set of variables considered. If, by adding this variable, the RSS did not diminish "enough" the variable cannot be introduced into the model, otherwise it is introduced. The process continues until the stopping criterion has been accomplished or all potential variables have been introduced in the model.

The expressions "too much" or "enough" have a very precise quantitative correlation by comparing the respective values with the adequate quantiles of Fisher distribution.

Generally, the use of the forward selection method provides reasonable solutions in terms of computing time compared to the backward selection method. However, a major disadvantage of this method is that it ignores the effect of introducing a new variable in the model on the role of existing variables in the model.

The *stepwise regression method* improves the algorithm of the forward regression method as follows: after the introduction of a variable in the model, with the exception of the first variable, it is tested whether any of the previously introduced variables can be deleted without increasing too much the RSS value. In this way, through the test, we can see if a variable, considered the most suitable to enter the model at some point, became insignificant in terms of its contribution to the RSS value, because of its relation with the variables introduced later into the model.

Regression Models Based on Dimensionality Reduction

The idea behind this type of models is that in a large regression model we can determine a number of variables called latent factors $Z_1, Z_2, \cdots, Z_m$, $m < p$ which can explain the largest part from the dependent variable variance y. In such a model the m new explicative variables $Z_1, Z_2, \cdots, Z_m$ are obtained as linear combinations of the initial variables $X_1, X_2, \cdots, X_p$. Starting from the model (1), we obtain (Hastie, Tibshirani and Friedman, 2009):

$$Y = \theta_1 Z_1 + \theta_2 Z_2 + \cdots + \theta_m Z_m + \varepsilon \quad (6)$$

where $Z_k = \sum_{j=1}^{p} \alpha_{j,k} X_j$, $k = 1,2, \cdots m$.

By stating the values $\alpha_{j,k}$, the regression model (6) can be classically solved, through the least squares method.

The regression model (6) can be written as

$$Y = \sum_{k=1}^{m} \theta_k Z_k + \epsilon = \sum_{k=1}^{m} \theta_k \sum_{j=1}^{p} \alpha_{j,k} X_j + \epsilon = \sum_{j=1}^{p} \sum_{k=1}^{m} \theta_k \alpha_{j,k} X_j + \varepsilon$$

or equivalently,

$$Y = \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon, \quad (7)$$

where

$$\beta_j = \sum_{k=1}^{m} \theta_k \alpha_{j,k}. \quad (8)$$

From the relations (7) and (8) we obtain the following interpretation: solving the regression model (6) is equivalent to solving the initial model (1) under the restrictions (8).

For a particular choice of the weights $\alpha_{j,k}$, the model with reduced size (6) offers better results that the initial model (1). There are two main models which belong to this class of regression models (Hastie, Tibshirani and Friedman 2009, pp. 79-80): Principal Component Regression (PCR) and Partial Least Squares (PLS).

The PCR model is obtained when $Z_1, Z_2, \cdots, Z_m$ are the first m principal components. In this case we can assume that the directions in which the predictive variables X_j have the higher variation (directions indicated by the first m principal components) are those which reflect their association with the dependent variable Y. The algorithm is based on the use of the Singular Value Decomposition(SVD) of the matrix $X^T X$, where X is the design matrix from the relation (3). If the assumption is correct the PCR model has the following advantages (Michy, 2016) :

- dimensionality reduction, if the number m of used principal components is much smaller than p;
- obtaining a model where multicolinearity or almost multicolinearity is certainly missing (principal components are uncorrelated);
- reducing the risk of over-fitting by using a small number of predictor variables.

The PCR model ignores the correlation degree between the new variables $Z_1, Z_2, \cdots, Z_m$ and the dependent variable Y. In PLS model we also take into account this aspect. Practically, the

latent factors $Z_1, Z_2, \cdots, Z_m$ are extracted successively by an iterative process which maximizes the correlation between the variables Z_i and the dependent variable Y. The used algorithm, introduced by Wold (1975), is called NIPALS (Nonlinear Iterative Partial Least Squares) and it is based on the Singular Value Decomposition (SVD) of the matrix X^TY.

Regularized Regression Models

By regularizing a regression model we understand the process by which new information is added to the model in order to solve an ill-posed problem or to avoid the over-fitting of the model.

In the following we make a brief presentation of four techniques of regularized regression: Ridge regression, LOSS (Least Absolute shrinkage and Selection), Adaptive LOSS and ElasticNET.

Ridge regression model

Today there are a lot of practical problems that occur frequently (e.g. image processing, microarray regression and gene selection etc.) where the number of variables p is strictly higher than the number of observations n , i.e. H1 assumption is not satisfied. In this situation a part of the model variables can be correlated, with the result that the matrix X has collinear or nearly collinear columns and implicitly the matrix X^TX of the normal equation system (4) is not invertible. As a result:

- There is not a unique solution obtained for solving the regression model;
- The model obtained is unstable, any easy change of the inputs (caused, for example, by the rounding errors) produces great changes in the final result;
- Mean Squared Error (MSE), total variance and mean length of the estimator $\hat{\beta}^{BLUE}$ take very high values which causes a high variance of the model (Vinod and Ullah, 1981).

A solution to prevent the regression coefficients $\hat{\beta}_i$ to take uncontrolled large values is regulating their parameters β_i for example constraining them to have a smaller length than an arbitrary positive value c.

In practice we are looking for a value that minimizes, as in the case of the classical regression model, the sum of the squares of the errors $S(y,\beta) = \sum_{i=1}^{n} \varepsilon_i^2$, with the added constraint $\|\beta\|^2 \leq c$. Mathematically, we must solve the problem,

$\min_\beta \|y - X\beta\|^2$ with the restriction $\|\beta\|^2 \leq c, c > 0.$

The problem is equivalent to the minimum problem obtained by penalizing the sum of squared errors, the penalty function being the Euclidean norm (l_2) of the parameter β:

$\min_\beta \|y - X\beta\|^2 + k\|\beta\|^2$, *k>0.*

which has the unique solution $\hat{\beta}_k^{Ridge} = (X^TX + kI_n)^{-1}X^Ty$. In this case, the matrix X is standardized, and the variable y is centred.

The Ridge Estimator was invented by (Hoerl and Kennard, 1970) as a solution to the regression problems that do not satisfy H1 hypothesis. Unlike the least squares estimator, the Ridge estimator is unbiased, but it generally has lower prediction error. In this way (Vinod and Ullah, 1981) , there is always a value of the k parameter so $MSE(\hat{\beta}_k^{Ridge}) \leq MSE(\hat{\beta}^{BLUE})$. The k is called shrinkage parameter: when $k \to 0, \hat{\beta}_k^{Ridge} \to \hat{\beta}^{BLUE}$ and when $k \to \infty, \hat{\beta}_k^{ridge} \to 0$

LASSO (Least Absolute Shrinkage and Selection)

Ridge regression model produces regression coefficients shrunken toward zero value, but never equal to zero. This can be considered as a limitation of Ridge regression model, which does not provide a selection of the variables entered in the model, while maintaining a model with many variables and therefore difficult to interpret. Tibshirani (1996) replaces Euclidean norm l_2 used in Ridge regression model as a penalty function, by the norm l_1 (also called Manhattan distance or city-block distance or taxi distance), obtaining LASSO model. In LASSO model, a part of the coefficients take zero value thereby achieving a model with fewer variables.

The norm l_1 of the parameter β is defined by:

$$\|\beta\|_{l_1} = |\beta_1| + |\beta_2| + \cdots + |\beta_p|.$$

The estimators of the parameters β_i is obtained by solving the minimum problem:

$\min_\beta(\|y - X\beta\|^2 + k\|\beta\|_{l_1})$, *k>0.*

The problem does not accept an analytical solution but can be solved by using algorithms of the quadratic programming. LARS (Least Angle Regression) is a powerful algorithm developed specifically to solve this problem in Bradley et al. (2004). The main disadvantage of this method is that for $p > n$ the method limits the maximum number of the selected variable to n, i.e. to the number of observations available. Also, if several variables are correlated, practically forming a group, the method tends to select only one variable in the group and not the whole group.

Adaptive LASSO

In Zou (2006) a modified version of LASSO algorithm is introduced, called Adaptive LASSO, in which the coefficients β_j are found by solving the minimum problem:

$\min_\beta(\|y - X\beta\|^2 + k\sum_{j=1}^{p} \widehat{w}_j\beta_j)$, *k>0,*

where $\widehat{w}_j$ are weights dependent on data. These weights have the role to penalize differently each coefficient. As LASSO, the Adaptive LASSO algorithm permits the selection of a submodel of the model (1) by canceling some of the coefficients. In addition, it is shown that for an appropriate choice of weights $\widehat{w}_j$ (for example $\widehat{w}_j = 1/\hat{\beta}_k^{ridge}$), the estimators obtained with Adaptive LASSO algorithm, in contrast with LASSO algorithm, have oracle type properties, meaning that it work "as well as if the correct submodel is known" (Fan and Li, 2001) , namely (ZOU 2006, p.1418):

- identifies the right submodel: by noting $A = \{j: \hat{\beta}_j \neq 0\}$, we have $A = \{j: \beta_j \neq 0\}$;
- has the optimal estimation rate: $\sqrt{n}(\hat{\beta}_A - \beta_A)$ tends in distribution to the normal distribution $N(0, \sigma^*)$, where σ^* is the covariance matrix knowing the true submodel.

ElasticNet

The Ridge Regression method is very useful when the variables of the model are highly correlated, yet not having the ability of the LASSO method to eliminate some variables in the model. The LASSO method also has at least two limitations specified in the paragraph where the method was described. The ElasticNet method proposed in Zou and Hastie (2005) performs the regularizations and selection of variables simultaneously and manages to cancel the disadvantages of the two methods and only keep their advantages. ElasticNet method uses as penalty function a linear combination between the norms l_2 and l_1 already used with this

purpose in the Ridge regression method and in LOSS method. Naturally, the problem proposed to be solved in this case is the following one:

$$\min_{\beta}(\|y - X\beta\|^2 + k_1\|\beta\|^2 * k_2\|\beta\|_{l_1})\ k_1, k_2 > 0,\ k_1 + k_2 = 1$$

or equivalent

$$\min_{\beta}(\|y - X\beta\|^2 + (1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_{l_1}),\ 0 \le \alpha \le 1$$

The parameter α adjusts the degree of closeness to Ridge regression method or LASSO method. At the limit, for $\alpha = 0$ the method reduces to Ridge regression, and for $\alpha = 1$ to LASSO method.

Solving the problem of the minimum is achieved with LARS-EN algorithm proposed in Zou and Hastie (2005), which is based on the algorithm LARS.

Conclusions

Despite the passing of time, regression modelling remains a timeless concept and a tool permanently adaptable in order to successfully cope with problems that require new knowledge extraction from a growing data volume. Emerged in the context of the evolution of mathematical statistics, regression models became, in the digital age, the starting point or the source of inspiration for a large part of the specific techniques in a field which develops constantly, called data science. The modern data modelling techniques presented in the last part of this paper represent only a few data mining techniques which have as a source of inspiration the linear regression model.

References

1. Afifi, A.A. and Azen, S., P., 1972. *Statistical Analysis* , A computer oriented approach, Academic Press, New York, London.
2. Biernat, E. and Lutz, M., 2016, *Data Science:fondamenteaux et etudes de cas-Machine Learning avec Python et R,* Groupe Eyrolles, Paris.
3. Bradley, E., Hastie, T., Johnstone, I. and Tibshirani, R., 2004. *Least Angle Regression*, The annals of Statistics, Vol. 32, No.2, 407-499, Institute of Mathematical Statistics.
4. Cornillon P. A. and Lober E. M, 2007. *Regression-Theory et applications*, Springer–Verlag France, Paris.
5. Draper, N.R. and Smith, H., 1966. *Applied regression analysis*, John Wiley and Sons, Inc, New York, London, Sydney, First edition.
6. Fan, J. and Li, R., 2001. Variable selection via penalized nonconcave likelihood and its oracle properties, *Journal of the American Statistical Association*, Vol.96, No.456, December 2001, Theory and Methods.
7. Fortmann-Roe, S., 2012. *Understanding the Bias-Variance Tradeoff,* available at http://scott.fortmann-roe.com/docs/BiasVariance.html, [Accessed on 12.03.2017].
8. Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning, Data mining, inference and prediction*, Second edition, Springer Series in Statistics.
9. Hoerl, A. E. and Kennard , R. W., 1970. Ridge regression: biased estimation for nonorthogonal problem, *Technometrics*, Vol. 42, No. 1, pp. 55-67.
10. Javaras, K. and Vos, W., 2002. Introduction to statistical modeling, *ModellingLecture3- Linear Models(continued*), [online via internal VLE], University of Oxford. Available at: https://www.stats.ox.ac.uk/pub/bdr/IAUL/. [Accessed on: 13.03.2017].
11. Marinoiu, C., 2015. *Modele de regresie liniară*, Petroleum-Gas University of Ploiesti Publishing House.

12. Michy A, 2016. *Performing Principal Components Regression (PCR) in R* , available at http://www.quantide.com/performing-principal-components-regression-pcr-in-r/ [accessed on 15.03.2017].
13. Miller, A.J., 1990. *Subset selection in regression*, Chapman and Hall, London , New York, Tokyo, Melbourne, Madras.
14. Montgomery, D., Peck E. and Vining, G., 2012. *Introduction to linear regression analysis,* John Wiley & Sons, Inc., Publication, Fifth Edition.
15. Tibshirani, R., 1996. Regression Shrinkage and Selection via LASSO, *Journal of the Royal Statistical Society*, Serie B (Methodological), Volume 58, ISSUE 1 , 267-288.
16. Vinod, H.D. and Ullah, A., 1981. *Recent advances in regression methods,* Marcel Dekker , Inc., New York.
17. Zou, H. and Hastie, T., 2005. Regularization and variable selection via Elasic Net, *Journal of Royal Statistical Society*. B, Part 2, pp.301-320
18. Wold, H., 1975. *Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach*, Perspective in Probability and Statistics, In honor of M.S.Bartlett, pp. 117-144.
19. Zou, H., 2006. The Adaptive LASSO and its Oracle Properties, *Journal of the American Statistical Association*, Theory and Methods 101, NO. 476, pp. 1418-1429.