CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Jana Vorlíčková

Least Absolute Shrinkage and Selection Operator Method

Bachelor thesis

Prague 2017

Author: Jana Vorlíčková
Supervisor: RNDr. Michal Červinka, Ph.D.

Academic Year: 2016/2017

Bibliographic note

VORLÍČKOVÁ, Jana. *Least Absolute Shrinkage and Selection Operator Method.* Prague 2017. 34 pp. Bachelor thesis (Bc.) Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor RNDr. Michal Červinka, Ph.D.

Abstract

The main intention of the thesis is to present several types of penalization techniques and to apply them in economic analyses. We focus on penalized least squares, with a main topic being the lasso. The penalization methods are commonly employed to data sets with a relatively large number of the variables as compared to the sample size. These methods simplify the model by shrinkage of the estimates of the coefficient of the irrelevant variables towards zero or they put these estimates equal to zero, i.e. they produce a sparse solution. Namely, we present the following methods: ridge regression, best subset selection problem, lasso and elastic net. We discuss several applications of the lasso in the current economic and finance research and hence present the lasso in more details. In the practical part of the thesis, we analyze a real economic data using the elastic net, the ridge regression, the lasso and the ordinary least squares method. We use the mean squared error as the measure of performance of the respective method. The penalized least squares methods surpass the ordinary least squares method, with the elastic net being the best performing method.

Keywords

penalized least squares, lasso, elastic net, ridge regression, penalization techniques in economics

Abstrakt

Cílem této práce je diskutovat význam a možné použití penalizačních metod pro lineární model v ekonomii. Penalizace se používá zejména při odhadování parametrů modelu s vysokým počtem proměnných. V této práci se přitom soustředíme na penalizaci nejmenších čtverců, která je nejčastěji používanou ztrátovou funkcí. Penalizační metody zjednodušují model a to tak, že odhady koeficientů u nedůležitých proměnných zmenšují či vynulují, tj. produkují tzv. řídké řešení. Jmenovitě jsou představeny následující metody: hřebenová regrese, úloha výběru nejlepší podmnožiny, lasso metoda a elastická síť. Nejvíce prostoru v teoretické části je věnováno lasso metodě; představeno je i její současné užívání v ekonomických a finančních analýzách. V praktické části práce analyzujeme reálná ekonomická data pomocí elastické sítě, hřebenové regrese, lasso metody a klasické metody nejmenších čtverců. Získané výsledky srovnáváme pomocí střední čtvercové chyby. Ukazujeme, že penalizační metody poskytují lepší výsledky než metoda nejmenších čtverců, přičemž nejlepší výsledky jsou získány s pomocí elastické sítě.

Klíčová slova

penalizované nejmenší čtverce, lasso, elastická síť, hřebenová regrese, penalizaγní metody v ekonomii

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, 16 May 2017

Signature

Acknowledgment

I am grateful especially to the supervisor RNDr. Michal Červinka, Ph.D. for his time and constructive remarks that improved the quality of the thesis. I would also like to thank Tomáš Masák for his helpful comments. Finally, I am thankful to my family and friends for their support during the studies.

Bachelor thesis proposal

Author	Jana Vorlíčková
Supervisor	RNDr. Michal Červinka, Ph.D.
Topic	Least Absolute Shrinkage and Selection Operator Method

The lasso (Least Absolute Shrinkage and Selection Operator) [1] is a method used to estimate important variables in models which work with high dimensional data. The lasso is a penalized regression technique which uses l1-norm (absolute value) penalization and it is based on minimization of the least-squares objective function which includes l1-penalty term. This technique performs both regularization and variable selection. We introduce related penalization techniques, namely, the so-called best subset selection method, ridge regression method and elastic net method.

The main goal of this bachelor thesis is to illuminate application of the lasso method when analyzing real economic data. We will employ the R software for numerical experiments. We shall compare the lasso estimator with several other types of estimators based on minimization of mean squared error.

Outline

- Introduction
- Lasso regression method
- Modifications of lasso and theoretical comparison
- Data analysis and numerical comparison
- Conclusion

Bibliography

[1] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* Vol.58, No.1, 267-288, 1996.

[2] Jacob, L., Obozinski, G., and Vert, J.P. Group lasso with overlap and graph lasso. *Proceeding ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[3] Buehlmann, P., Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications0.* Springer Berlin Heidelberg, 2011.

[4] Hastie, T., Tibshirani, R., Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations.* Chapman & Hall/CRC Monographs on Statistics& Applied Probability, 2015.

[5] Belloni, A., Chernoyhukov, V., Hansen, Ch. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic perspectives*, vol.28, no.2, 2014.

[6] Zou, H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, Volume 101, Issue 476, 2006.

Contents

Notation

$\mathbf{X}_i$	vector of data inputs (independent variable) of length n
$\mathbf{X}$	$(n \times k)$-matrix of independent variables, where $\mathbf{X}_j$, $j = 1, \ldots, k$, are columns of $\mathbf{X}$
β	vector of coefficients of length k
$\mathbf{y}$	dependent variable, vector of length n
$\mathbf{r}$	vector of residuals of length n
$\mathbf{X}^{\mathsf{T}}$	transpose of a matrix $\mathbf{X}$
$\mathbf{X}^{-1}$	inverse of a matrix $\mathbf{X}$
$\mathbf{X}^{-}$	pseudoinverse of a matrix $\mathbf{X}$
x^{+}	$\max\{0, x\}$
$sign(x)$	is equal to -1 if $x < 0$, to 1 if $x > 0$ and to 0 if $x = 0$
$\mathbb{I}_A$	indicator function of a subset A
$\mathbf{I}_k$	$(k \times k)$-identity matrix
$\mathrm{E}\mathbf{X}_i$	expected value of $\mathbf{X}_i$
$\mathrm{Var}(\mathbf{X}_i)$	variance of $\mathbf{X}_i$
$\lVert \cdot \rVert_p$	ℓ_p-norm

1 Introduction

Data sets with a relatively large number of the variables as compared to the sample size are becoming increasingly available and more common in empirical economics and finance. When analyzing such data, standard model‘s interpretability is plummeting due to this high dimensionality. Recently, the econometricians and statisticians have developed various ways of dealing with high dimensional data. However, those methods still have not been commonly implemented to the economic research tools.

The aim of the thesis is to introduce the econometric method of penalized ordinary least squares, namely the lasso, the ridge regression and the elastic net, with main topic being the lasso. The lasso is a widely used method in high dimensional data analyses in natural sciences. Nevertheless, it can be helpful in the economic research. This technique is able to provide a sparse model, i.e. it selects the only relevant variables, and it creates the sub-model. The sparse sub-models are more transparent and the interpretation is easier then in the case of the full model. Consequently, we examine the performance of those methods on the real data. We have a data set collected by Trend in International Mathematics and Science Study (TIMSS). We include the comparison to the OLS estimator using mean squared error as the relevant statistical criterion.

The thesis is structured as follows: In Chapter 2 we introduce the lasso application in economic and finance research. We acquaint the reader with the current application of the lasso on particular examples. In the end of this chapter, we provide the general mathematical background, which is needed in the theory behind the lasso. The following two chapters acquaint the reader with the mathematical theory behind the penalization techniques. We describe the best subset selection, the ridge regression and the lasso in more details. As the main topic of the thesis, the lasso is elaborated at most. Finally, we evaluate the performance of the penalization methods on the real behavioral microeconomic data in Chapter 5.

2 Lasso in economics and finance

In this chapter, we first give a literature review of application of the lasso in economic and finance research. We summarize the main thought of economists, which takes into consideration the penalized least squares methods as a new econometric technique. We focus on the lasso (Tibshirani, 1996), which is a regularization procedure that shrinks regression coefficients toward zero and it simplifies the model by variable selection. The detailed description of the lasso is provided later in Chapter 4. Secondly, we review the mathematical background, which covers the basic definitions and the derivation of the OLS estimator. This theory is needed in Chapter 3 and 4.

2.1 Literature review

The high dimensional data set are becoming more and more frequently analyzed in the economic and finance research. Economists state that the high dimensional data arise through a combination of two phenomena (Belloni et al., 2014). Firstly, many different attributes per observation are available, and the set of these characteristics creates high dimensional data sets. For example, the US Census (or the Census of another country), the Current Population Survey, the Survey of Income and Program Participation, and the National Longitudinal Survey of Youth collect information on hundreds of individual characteristics. Thus, it is possible to obtain thousands or tens of thousands of available variables per observation. Secondly, even when the number of available variables is not large, researchers rarely know which variables are the underlying ones to create the model of interest, which is omitting the insignificant variables. Researchers are thus faced with a large set of potential variables formed by different ways of transforming and interacting the relevant variables (Belloni et al., 2014).

When we have to deal with high dimensional dataset in economic context, lasso is frequently present. The intention of economists is to develop a simple and generally understandable model and the penalization methods have proved to be very helpful in achieving this goal (Caner and Zhang, 2013). The family of selection techniques are particularly more desired because they accurately return a sparse model.

These techniques are becoming popular mainly due to the increasing amount of available data, e.g. analysts attempting to change the attitude of the econometric analyses of consumer demand estimation and or integrate new methods. In markets selling complex products, such as notebooks or automobiles, an individual‘s preferences could relate to a high-dimensional set of characteristics describing these products. Despite this complexity, researchers often assume that the dominant influences on consumers‘ decisions can be described by a relatively small set of relevant characteristics implicitly selected as part of the empirical study. The penalization techniques as the lasso can be used as an effective tool to analyze demand with high dimensional product characteristics. It provides a selection of important attributes of customers‘ decision (Gillen et al., 2014). These authors have introduced a complex model, where they have performed a selection via a triple-lasso, i.e. the lasso was applied three times in the model. Successively the authors have used the lasso for the selection of explanatory controls, treatment controls and instrumental variables.

Once more the instrumental selection and the estimation of simultaneous equations model by the lasso were demonstrated on the three extra examples (Belloni et al., 2014). The authors have analyzed data from different economic topics adopted from the empirical literature: estimating the effect of eminent domain on house prices, estimating the effect of abortion on crime, and estimating the effect of institutions on economic output. The estimation of the effect of the institution is more complicated because there can be simultaneity between output and institution, i.e. higher incomes may lead to the more developed institutions and vice verse. The application of the lasso on a simultaneous equation model consists of running the lasso on each of the three reduce-form equations. They emphasized that the lasso targets prediction and selection of relevant variables, not learning about specific model parameters. Firstly, they did dimensional selection by researchers‘ intuition and subsequently they applied the lasso. Their suggestion is that the high dimensional techniques may usefully complement selection using subject matter expertise to choose instruments and strengthen researchers‘ ability to draw useful conclusions from data. They claim that these techniques can add rigor to these exercises and thus potentially strengthen the plausibility of conclusion drawn in applied economic papers.

Despite the fact that the lasso was originally introduced for linear regression problems, this approach has been applied to time series analysis by several authors mainly in case of autoregressive model (Wang et al., 2007) or (Nardi and Rinaldo, 2011). Obviously there are other modifications of the lasso, which were introduced especially for the economic purposes. One of them is the application of the lasso to high dimensional regression with a possible change-point (Lee et al., 2016). They have analyzed the relationship between economic growth rate and the initial GDP with other covariates. They have not only examined the lasso performance, but also provided a discussion about the oracle properties and the asymptotic behavior for this special case. We do not discuss these characteristics of the lasso in this thesis but for the interested readers we recommend the article about the adaptive lasso (Zou, 2006) and a detailed work (Bühlmann and Van De Geer, 2011). Other lasso-based methods exist and they are steadily emerging in recent econometric research, e.g. (Schneider and Wagner, 2012) performed the adaptive lasso to catching growth determinants. In finance, the lasso quantile regression have been used to measure a causal relationship between hedge funds and a benchmark assets at different quantiles. Subsequently, the reduction of dimension has been provided by the lasso and the author have implemented Value-at-Risk (VaR) as a portfolio risk measure. In this case we can see the connection between commonly used econometric method and the lasso (Nasekin, 2013).

However, we do not detect any article where the behavioral microeconomic data were analyzed, so it is the intention of this thesis to make the variety of economic data types estimated by the lasso more complete.

There is a widespread presence of other techniques for handling the high dimensional data sets in economic analyses. An article written by three french authors (Epprecht et al., 2013) compares two of those methods - the lasso and the autometrics. The autometrics carries out a set of statistical tests like F-test, t-test or test for heteroskedasticity and it finds the best model based on the results of those tests. In the simulations, the autometrics performs better than the lasso. However, on real data, the lasso provides better results. As the set of techniques used to provide a sparse high dimensional model is various, we suggest a useful summary in (Varian, 2014) or (Fan et al., 2011) for an interested reader. Mostly, both articles summarize the theoretical background of the high dimensional techniques and, regretfully, they have refrained from covering more illustrative empirical examples. The practical example can be found in (Bai and Ng, 2008). The authors have compared performance of several techniques for high dimensional data sets in forecasting economic time series.

As a conclusion, the lasso is slowly getting a place between popular econometric tools. The number of problems requiring analyses of high dimensional data is growing, therefore we can expect the increasing popularity of penalization methods. Moreover, the strong advantage of the lasso lies in its sparsity and this property can take contribution to economic analyses in the form of uncluttered models.

We summarize the application of the lasso in the economic and finance research in this section. In sequel, we present the penalized least squares with concentration to the lasso. Firstly, there is a description of the theoretical background which is consequently supported by the application on the real data.

2.2 Mathematical background

This section summarizes and explains mathematical terms, which are needed for understanding of the theory behind the topic of this thesis. We also remind the derivation of the estimator of coefficients β using the ordinary least squares method (OLS) (Zvára, 2008). The reader familiar with these topics can readily skip this part.

During the whole thesis, the matrix notation will be used for simplicity. We recall the definitions of *pseudoinverse* of a matrix, *orthogonal* matrix and *singular* matrix.

Definition 1 (Pseudoinverse of a matrix)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times k}$ be a real matrix. Then matrix $\mathbf{A}^{-}$ is a pseudoinverse of $\mathbf{A}$ whenever

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}.$$

Definition 2 (Orthogonal matrix)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times k}$ be a real matrix. Then matrix $\mathbf{A}$ is called orthogonal whenever

$$\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{I}.$$

Definition 3 (Singular matrix)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times k}$ be a real matrix. Then matrix $\mathbf{A}$ is called singular whenever

$$\det\mathbf{A} = 0,$$

where $\det\mathbf{A}$ is a determinant of a matrix $\mathbf{A}$. Otherwise it is called an invertible matrix.

The other term which we will encounter is *ℓ_p-norm.* Generally, the norm is a function, which describes in certain (possibly abstract) sense the length, size, or extent of the object.

Definition 4 (ℓ_p-norm)**.** Let $\mathbf{x} \in \mathbb{R}^n$ be a real vector and $p \geq 1$ be a real number. Then

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

is called ℓ_p-norm of vector $\mathbf{x}$.

For our purposes, ℓ_1-norm and ℓ_2-norm are important.

Method of the ordinary least squares

Suppose we have data $\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, $\mathbf{X}_i \in \mathbb{R}^k$, $i = 1, \ldots, n$, and $\mathbf{X} \in \mathbb{R}^{n \times k}$. We assume without loss of generality that they are centered, i.e. no intercept is needed, and suppose we have the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$. Now we would like to estimate parameters β by the OLS method, formulating the problem as

$$\hat{\beta}^{OLS} \in \underset{\beta \in \mathbb{R}^k}{\arg\min} \; \|\mathbf{Y} - \mathbf{X}\beta\|_2^2. \tag{2}$$

The method of ordinary least squares is a procedure where the minimum of sum of squares is found. We denote $S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\mathsf{T}(\mathbf{Y} - \mathbf{X}\beta)$, $\beta \in \mathbb{R}^k$, and we take the derivatives of S with respect to β,

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}^\mathsf{T}(\mathbf{Y} - \mathbf{X}\beta).,$$

Furthermore, we set the gradient of S equal to the vector of zeros and we obtain the system of normal equations

$$\mathbf{X}^\mathsf{T}(\mathbf{Y} - \mathbf{X}\beta) = 0. \tag{3}$$

Even if $\mathbf{X}^\mathsf{T}\mathbf{X}$ is not invertible we obtain a solution to normal equations,

$$\hat{\beta}^{OLS} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-}\mathbf{X}^\mathsf{T}\mathbf{Y}. \tag{4}$$

However, the pseudoinverse need not be unique, thus the obtained solution is not unique. In the case, when we have a full-rank linear model then $(\mathbf{X}^\mathsf{T}\mathbf{X})^- = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, and there exists a unique solution to normal equations (3) defines as

$$\hat{\beta}^{OLS} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}. \tag{5}$$

In the following chapter we introduce several types of estimator which are based on penalization of the OLS objective function.

3 Selection and shrinkage methods for linear models

The usual way how to estimate the coefficients in linear models is by the ordinary least squares method (OLS). This method is simple, easy to understand and very useful for a certain type of data. If we have model with negligible number of parameters relative to the number of observations the OLS method returns satisfactory results. However, if k (a number of explanatory variables) is large enough and, especially, when larger than n (a number of observations), the model estimated by the OLS method becomes very complex. The complex model containing so many variables is hardly interpreted due to the fact the OLS method typically produces non-zero estimates of coefficients. Moreover, in the unrestricted model, where selection is not provided, we can meet the problem of overfitting. Overfitting occurs when we have a complex model, e.g. having too many covariates relative to the number of observation, than it is desired, which leads to poorer predictive performance because of the estimate is excessively affected by noise in data. Nowadays, it is a common problem in statistical learning methods (Friedman et al., 2001).

One possibility is to deal with overfitting by penalization methods. The objective function (2) is penalized by the restricted ℓ_p-norm. The ℓ_p-norm is controlled by the parameter $t > 0$, which is chosen in advance. Generally, the problem can be written as

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^k:\ \|\beta\|_p \leq t}{\arg\min} \ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2. \tag{6}$$

For $p > 1$, we have a strictly convex problem, and we obtain the unique solution of (6). The estimates of coefficients are shrunk toward zero. The amount of shrinkage depends on the magnitude of the tuning parameter t. We present in detail the ridge regression (Hoerl and Kennard, 1970) where the l_2-norm is employed. It partially solves overfitting but it does not make the model simpler, i.e all parameters are active. The selection is provided for $p = 0$ [1]. This problem is a so-called best subset selection. Unfortunately, it is a multi-step method. The solving of this method takes a high computation cost. The interesting problem arises in case of $p = 1$. The ℓ_1-penalized problem is still convex, and at the same time this is the largest value of p which allows the variable selection. This method is called the lasso (Tibshirani, 1996) and it is the topic of Chapter 4.

3.1 Ridge regression

Suppose a data sample of size n, and k is the number of explanatory variables. The OLS estimator $\hat{\beta}^{OLS}$ is uniquely defined only if $\mathbf{X}^\mathsf{T}\mathbf{X}$ is an invertible matrix. Therefore, necessarily the same inconvenience of non-unique estimates arises in high dimensional case, where $k \gg n$. The problem of singularity can be solved by a so-called ridge regression method (Hoerl and Kennard, 1970). For economists, the ridge regression can be a useful instrument when they have to deal with multicollinearity, i.e. the explanatory variables are highly correlated, thus they are close to the linear relationship. Consequently, the design matrix is close to singularity.

The ridge estimator is defined as a solution to the constrained minimization problem. The typical OLS objective function is constrained by the ℓ_2-penalization, which shrinks coefficients towards zero,

$$\hat{\beta}_{ridge} \in \underset{\beta \in \mathbb{R}^k:\ \|\beta\|_2 \leq t}{\arg\min} \quad \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \tag{7}$$

where t is a positive tuning parameter. Another way how to write the ridge regression problem is

$$\hat{\beta}_{ridge} \in \underset{\beta \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}, \tag{8}$$

where $\lambda > 0$ is a chosen parameter that controls how fast the coefficients are shrunk towards zero, i.e. the larger the value of λ the greater the effect of

[1] Definition 4 does not allow us this options as this is not a norm by the definition. We denote $\|\mathbf{x}\|_0$ as sum of nonzero elements $x_i, i = 1, \ldots, n$, of vector $\mathbf{x} \in \mathbb{R}^n$. In practice, this ℓ_0-penalization tells us a number of non-zero estimates of coefficients.

shrinkage. The ridge estimator can be written in closed-form expression

$$\hat{\beta}_{ridge} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}. \tag{9}$$

Note that due to a positive constant λ added to the diagonal of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, it handles the problem of singularity as the eigenvalues become nonzero.

The ridge estimator is not unbiased, however, the lower variance of $\hat{\beta}_{ridge}$ is obtained in contrast to the variance of $\hat{\beta}^{OLS}$. Moreover, it was proved that for each particular problem there exists a suitable choice of λ for which the mean squared error (MSE) is reduced as a result of the trade-off between the lower variance and the biased estimator (Hoerl and Kennard, 1970).

The ridge regression is an example of the estimator from the family of the continuous shrinkage methods. The multi-step selection method is introduced in the following section.

3.2 Best subset selection

The best subset selection is another method which can simplify the original model. Unlike the ridge regression, the best subset selection chooses only some variables to interpret the model. The principle of this method is based on selection of the subset from the original set of variables and the general approach is to pick the smallest subset that fulfills certain statistical criteria e.g., the largest adjusted R^2, the smallest possible MSE, etc. The reason why we would like to use only a subset of variables instead of a full set is because the chosen sub-model might actually estimate the coefficients β and predict responses with the smaller variance than the model using all explanatory variables. The sub-model also results in more transparent model, i.e. the interpretation is easy, rather than in case of the model with a full set of active variables.

The best subset selection provides an optimal solution. However, the finding of this solution is a very computationally demanding problem. Solving the problem accounts for estimation of 2^k sub-models and afterwards evaluation of each of them and selection of the best one, which fulfills the chosen criteria. Unfortunately, there is not any straight way of computation, which reduces the computation cost enough, especially for $k \gg n$ (Friedman et al., 2001). It is a reason why the statisticians have been forced to create new selection algorithms like the stepwise subset selections: the forward subset selection or the backward subset elimination (Friedman et al., 2001), which serves as a partial replacement of the absence of the optimal best subset selection algorithm. We do not provide here the precise description of these algorithms because it is beyond the scope of this thesis. Instead this thesis will mainly focus on the alternative, which is called the lasso. While the form of selection performed by best subset selection is a multi-step method, in the next chapter, we will see that the last one mentioned - the lasso, which is a one-step procedure.

4 Lasso

The lasso is an acronym for "least absolute shrinkage and selection operator" (Tibshirani, 1996). This acronym comes from its functionality that it does not only shrink coefficients towards zero but it also provides a selection of the significant covariates.

The lasso estimator is defined as

$$\hat{\beta}_{lasso} \in \underset{\beta \in \mathbb{R}^k:\ \|\beta\|_1 \leq t}{\arg\min} \ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2, \tag{10}$$

where $t > 0$ is a chosen tuning parameter. Substituing ℓ_2-penalty by ℓ_1-penalty in (7), we obtained the constrained minimization problem. Just as in case of the ridge regression, the lasso estimator (10) can be rewritten to an unconstrained minimization problem,

$$\hat{\beta}_{lasso} \in \underset{\beta \in \mathbb{R}^k}{\arg\min} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{11}$$

where $\lambda > 0$. There is not a one-to-one relationship between the parameter λ and the tuning parameter t for all values of $t > 0$. When t is chosen large enough that the OLS estimator lies inside the constraint set, i.e. constraint in (10) is not active. We are not able to find a corresponding non-zero value of λ, which gives the non-penalized OLS estimator. The OLS estimator is obtained in (11) for $\lambda = 0$. Thus, the relationship holds if and only if the constraint $\|\beta\|_1 \leq t$ is active.

The estimator of the ridge regression and the lasso is illuminated in Figure 1. We use the form of estimator (7) and (10) to depict the figure.

To simplify the problem, we consider only two parameters β_1 and β_2. The shaded areas are the constraint sets for the lasso and for the ridge regression, respectively. The elliptic contours are contours of sum of squares error with the OLS estimator in the center. Considering the ridge regression, the problem is strictly convex. The ℓ_2-geometry creates the constraint set in the shape of the disk and the ℓ_1-geometry creates the convex diamond with edges. The solution of both methods is the point where the elliptical

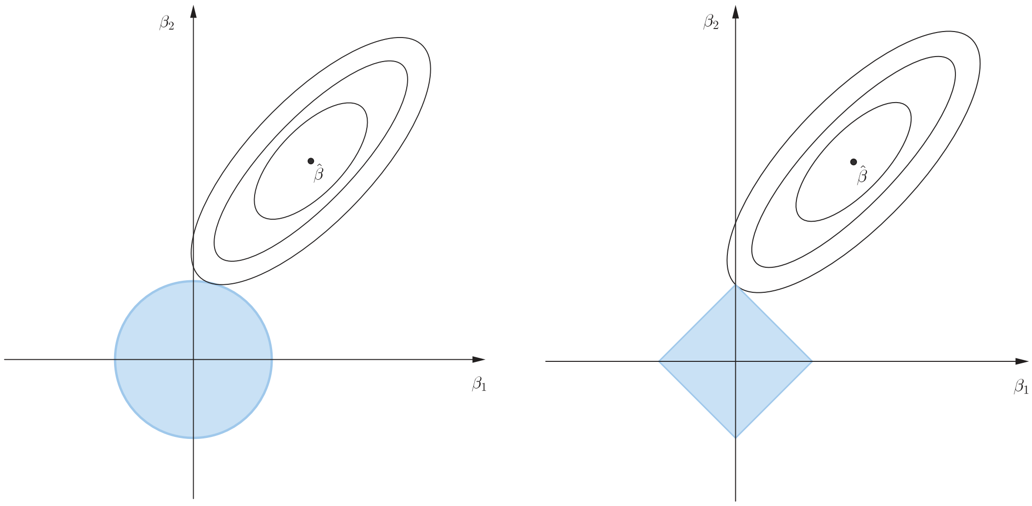


Figure 1: Illustration of two dimensional case of estimation for the lasso (right) and the ridge regression (left).

contours touch the constraint set. Note that the ridge estimate can be shrunk arbitrarily close to zero, but not exactly to zero due to the shape of constraint set (Hastie et al., 2015). Observe that the lasso estimate can be set to zero. Consider the situation when the contours touch the vertex then one coefficient is exactly zero. In Figure 1, β_2 is chosen as the only relevant parameter in our model.

The situation for higher dimensions is more complicated but it is based on the same principle. The diamond becomes a polyhedron, so with a higher number of estimated coefficients, the number of possibilities that some of the estimates are truncated to zero also increases. A three-dimensional case is depicted in (Tibshirani, 1996).

The lasso leaves only some parameters active and so does the optimal method best subset selection. While both provide a sparse solution, i.e. they reduce the number of explanatory variables in the final model, the lasso also shrinks estimates towards zero. Thus, they are not equivalent to the corresponding OLS estimates of the chosen sub-model. Unlike the replacement methods, e.g. forward or backward selection, the lasso takes an advantage in terms of continuous shrinkage of the coefficients towards zero. As a consequence, its asymptotic properties can be easier studied than if we are exploring these characteristics in case of the multi-steps methods (Bühlmann and Van De Geer, 2011). The continuity and mainly the convexity result in the existence of computationally efficient algorithms in contrast to the computation cost of the problem of the best subset selection.

Similarly to the ridge regression the lasso deals with overfitting by deriving a biased estimator but with possibly lower variance than the variance of the OLS estimator. It can lead to a lower expected prediction error in contrast to the result which we obtain when we use the OLS method. However, this does not hold in each situation. The particular examples are mentioned in (Hansen, 2016).

Secondly, the lasso is better than OLS for the purpose of interpretation. With a large number of independent variables, we often would like to identify a smaller subset of these variables that exhibit the strongest effects. The sparsity of the lasso is mainly counted as an advantage due to a simpler interpretation, but it is important to highlight that the lasso is not able to select more than n variables. This restriction does not cause any problem when $k < n$, but suppose we have a high dimensional case, $k \gg n$, and more than n relevant variables are required to be included in the model. In such a case when we employ the lasso, we do not obtain the right model.

Another potential problem occurs when we have a set of variables where are groups of highly correlated variables. The lasso tends to choose only one of these variables and it does not choose the whole group and as a consequence we obtain the misleading results.

In the following chapter we discuss the closed-form expression of the lasso estimator, which exists only under special assumptions. Consequently, the properties of the lasso estimator shall become clearer from this formula.

Estimator	Formula
Best subset (size M)	$\hat{\beta}_i^{OLS} \cdot \mathbb{I}_{(\lvert\hat{\beta}_i^{OLS}\rvert \geq \lvert\hat{\beta}_{(M)}^{OLS}\rvert)}$
Ridge	$\hat{\beta_i}^{OLS}/(1+\lambda)$
Lasso	$sign(\hat{\beta}_i^{OLS})\Big(\lvert\hat{\beta}_i^{OLS}\rvert - \frac{\lambda}{2}\Big)^+$

Table 1: The formulas for estimators of β_i provided the orthonormal design case. Each estimator is expressed in terms of the OLS estimator ${\beta_i}^{OLS}$ (Friedman et al., 2001, table 3.4).

4.1 Orthonormal design case

Generally, there does not exist any explicit closed-form solution for the lasso. However, in the special case of an orthogonal design, the lasso estimator can be explicitly expressed as a closed formula in terms of the OLS estimator. Likewise the ridge estimator and the best subset estimator can be written as a simple coordinate-wise transformation of the OLS estimator. Comparing these adjusted formulas of the penalized OLS estimators to the OLS estimator, we outline the mutual relationship, and this makes the approach of those penalization methods more comprehensible.

In the following lemma we state the lasso estimator for the orthogonal case. The proof is left as an easy exercise to the reader in available literature, thus, we leave it for reader, too. Afterwards we explain its relationship to the OLS estimator.

Lemma 4.1 (Tibshirani, 1996). *Let* $\mathbf{X}$ *be an orthogonal matrix. Then* $\mathbf{X}^\mathsf{T}\mathbf{X}$ *is a diagonal matrix and the solution of the optimization problem:*

$$\min_{\beta \in \mathbb{R}^k} \Big\{ (\mathbf{Y} - \mathbf{X}\beta)^\mathsf{T}(\mathbf{Y} - \mathbf{X}\beta) + \lambda \sum_{i=1}^{k} |\beta_i| \Big\} \tag{12}$$

can be written as

$$\hat{\beta}_i = sign(\hat{\beta}_i^{OLS})\Big(|\hat{\beta}_i^{OLS}| - \frac{\lambda}{2}\Big)^+, \qquad i = 1, ..., k, . \tag{13}$$

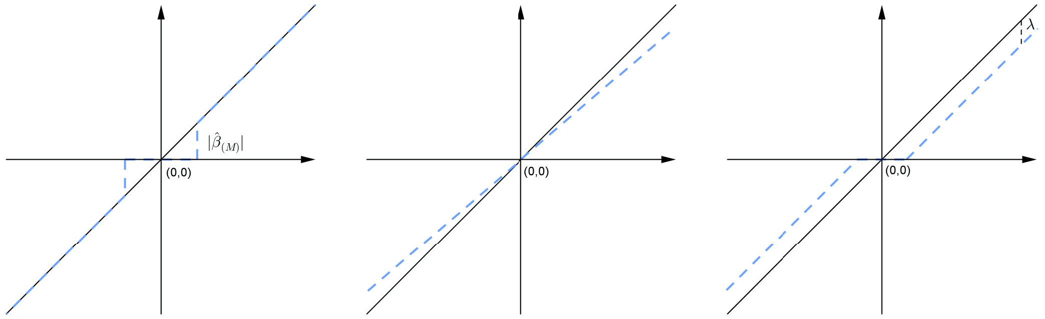
The estimators which are derived under the condition of orthonormal design case are summarized in Table 1. Looking at the table, it can be clearly seen that if λ is zero, the ridge estimate is equal to $\hat{\beta}^{OLS}$. Assume $\lambda > 0$ then the ridge regression provides a proportional shrinkage. The shrinkage becomes stronger with increasing λ. Further for $\lambda \to \infty$, the ridge estimate is shrunk towards zero. The lasso estimate is equal to zero if $\frac{\lambda}{2} \geq |\tilde{\beta}_i|$, i.e. for really small absolute values of the OLS estimates. Otherwise, $\lambda > 0$ causes shrinkage of coefficients towards zero and the sign of the lasso estimate corresponds to the sign of the OLS estimate.

The table is accompanied by Figure 2, where the estimators are depicted graphically. The estimators are drawn by broken blue lines. The black line shows the OLS estimator for reference. The best subset estimator is the most comprehensible if we concentrate on panel (a) of Figure 2. It provides so-called hard thresholding and it keeps the selected OLS estimate unchanged. Thus, the best subset estimator accurately copies the OLS estimator since the value of $|\hat{\beta}^{OLS}_{(M)}|$, which is chosen in advance. More precisely, M, a size of the subset, is chosen by the corresponding technique. This procedure is called hard tresholding due to this jump. Likewise, the characteristics of the lasso and the ridge regression are as well observable on the graphs (Friedman et al., 2001). It is apparent that the ridge regression does not provide a variable selection but it only shrinks estimates of coefficients, i.e. the model still contains the full set of variables, which does not make it easier to understand and to interpret.

In this section, we outlined the geometric representation of the particular penalization methods. We introduce the methods of the estimation of the lasso parameters in the next section.

4.2 Solution of the lasso

The number of non-zero coefficients in any lasso solution is not larger than n even if we consider the high dimensional case, $k \gg n$ (Tibshirani et al., 2013). Generally, when the number of nonzero estimates is not known in advance, we start with the estimation of the penalization parameter λ and



(a) Best Subset

(b) The Ridge Regression

(c) The Lasso

Figure 2: Comparison of the shrinkage and selection behavior of the estimators

then the appropriate algorithm is applied.

We do not want to choose λ too small due to the insufficient restriction. On the other hand, when λ is very large, the restriction is stronger than it is desired. We handle the problem of the optimal value of λ by the application of chosen criteria. In Tibshirani (1996) three methods of the estimation were introduced. However, there exist another criteria, which are used to estimate λ (Bühlmann and Van De Geer, 2011). These criteria are based on the degrees of freedom of the lasso and the minimization of prediction error. The parameter can be estimated by cross-validation, Akaike Information Criteria (AIC) and Bayes Information Criteria (BIC) (Zhang et al., 2010) or by the method based on Stein's unbiased estimate of risk (Tibshirani, 1996). Obviously, we can be estimated t , which is used in the constrained form (10), instead of λ. It returns very similar results.

We use a particular statistical software, which is able to compute the lasso solution by the selected algorithm. One of them is the lasso modification of Least Angle Regression (LARS), which is a model selection algorithm (Efron et al., 2004). It takes advantage of the lasso geometry and this makes easier to understand how the algorithm works. However, there exist several more computationally efficient algorithms, which are included in statistical packages as prepared functions. Using the statistical packages, it is easy to apply different algorithms for users. Thus, it is up to a particular problem, e.g. types of data, which one is the suitable choice.

We are not going to describe all these methods and algorithms in detail, we present only the main idea behind the cross-validation as it shall be employed in the practical part.

4.2.1 Cross-validation

The procedure known as a cross-validation is often used in practice in order to estimate the best value of λ. We would like to find λ which gives us the model with the lowest mean squared error (MSE). Firstly, let us recall the definition of the mean squared error.

Suppose we have

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon, \tag{14}$$

where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2 > 0$. The mean squared error is defined as a squared difference between true values of $\mathbf{Y}$ and fitted values $\hat{f}(\mathbf{X}) = \hat{\mathbf{Y}}$,

$$\mathrm{MSE} = \mathrm{E}[\mathbf{Y} - \hat{f}(\mathbf{X})]^2. \tag{15}$$

The mean squared error can be expressed as a sum of the variance of the estimator and its bias. This expression corresponds to already mentioned statement about the minimization of the trade-off between the variance and the bias, i.e. when the estimator is unbiased, it does not imply that it has the lowest possible MSE.

In Algorithm 1 (Efron and Tibshirani, 1994), the procedure of cross-validation is depicted. We create artificial training and test sets by splitting up the data set to K groups. Typical choice of K is 5 or 10, where one group is chosen as a test set and the remaining $K-1$ groups form the training set (Hastie et al., 2015).

The averaging of the K estimates of mean squared error (step 3), gives us a cross-validation error curve. The estimate of λ is chosen by the point,

Algorithm 1 K-fold cross-validation

```
1. Split the data into $K$ roughly equal-sized parts.
2. For the $k$th part, fit the model to the other $K - 1$ parts of the data, and calculate the mean squared error of the fitted model when predicting the $k$th part of the data.
3. Do the above for $k = 1, 2, \dots, K$ and average the $K$ estimates of mean squared error.
```

where the cross-validation error curve reaches the minimum. This choice often results in insufficient regularization, meaning too many variables stay active. This is so-called screening property (Bühlmann and Van De Geer, 2011). Therefore our best cross-validated model estimated in Subsection 5.3.1 will be based on the one standard error rule choice of λ, i.e. we take the most regularized model whose error is within one standard error of the minimal error. This rule is able to choose the simplest model whose accuracy is comparable with the best model, i.e. the one estimated for minimum of the cross-validation curve. (Friedman et al., 2001). The cross-validation will be used to analyze the real data in the following chapter.

4.3 Elastic net

Before the analysis, one of the lasso modifications, namely the elastic net (Zou and Hastie, 2005), is presented. The elastic net is able to provide a spare solution, and it eliminates the problems, that occur when the lasso is used on data sets with the highly correlated variables. The correlation is commonly present in analyses of economic and finance data (Persons, 1910). Therefore, we expect that the elastic net should have a convenient properties to the application in economics.

In Section 3.1, we have described the ridge regression, which is an efficient instrument for solving the problem of highly correlated variables, whereas

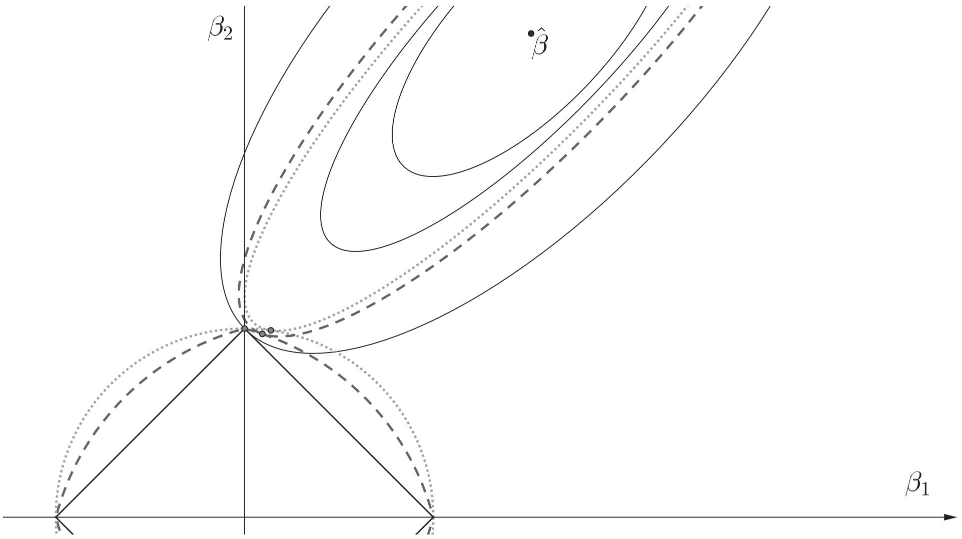


Figure 3: Illustration of the elastic net estimator for $\alpha = 0.5$

the lasso does not perform as well. On the other hand, the lasso gives us a sparse model, which we want to reach for the sake of interpretation. The combination of ℓ_1-penalty and ℓ_2-penalty conveys the desired properties of both methods to the new one - the elastic net (Zou and Hastie, 2005).

The elastic net estimator is defined as

$$\hat{\beta}_{EN} \in \underset{\beta \in \mathbb{R}^k}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \Big[\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \Big] \right\}, \tag{16}$$

where $\alpha \in [0, 1]$ is the parameter that can be varied. When $\alpha = 0$, it reduces to the squared ℓ_2-norm, and with $\alpha = 1$, it reduces to the ℓ_1-norm, corresponding to the lasso penalty. The regularization weight is controlled by $\lambda > 0$.

The penalization is a convex combination of the lasso and ridge penalty, and due to ℓ_2-penalty it is a strictly convex problem for $\alpha > 0$, i.e. a unique solution exists regardless of the correlation between independent variables. Figure 3 compares the constraint sets for the lasso (———), the elastic net (-----) and the ridge regression (⋯⋯). We mark the points where the elliptical contours of sum of squares touch the sets and these points denote the solution for particular methods. Figure 3 is a clear illumination of connection the lasso and the ridge regression characteristics described above. We highlight only the most important approach of the elastic net, and for more detailed discussion about the elastic net we refer the reader to (Zou and Hastie, 2005) or (Hastie et al., 2015).

In the following chapter, we analyze the real economic data by the penalization techniques, which we presented in this chapter and in Chapter 3.

5 Application on real data

We have already presented some application of the penalization methods in the economic research in Section 2.1. In this chapter, we show an application of the penalization methods on a real data set. We have been unable to recognize any article, which would evaluate the performance of the lasso method on the behavioral microeconomic data. Hence we have decided to fill this gap in econometric analyses. Moreover, we were unable to find any paper, which would compare different penalization techniques, namely the lasso, the ridge regression and the elastic net, on an economic data set.

Our data set is a part of the broader data set collected by Trends in International Mathematics and Science Study (TIMSS). The data are publicly available on the official webpage of TIMSS[2].

We proceed with applying the lasso, the elastic net and the ridge regression, and we compare the resulting performance using the mean squared error of each method in Section 5.3. However, we propose alternative and more complex models, which are expected to improve the performance at the cost of increasing complexity. These methods are based on theoretical concepts, which are beyond the scope of this thesis and therefore the methods will be only briefly discussed.

5.1 Data description

Our data set was collected by TIMSS in 2007. TIMSS provide a collection of a board data set across almost 60 countries regularly every 4 years. The students of 8th grade or 4th grade are tested in their skills in mathematics and science. The data set is accompanied by additional information about the background of schools, teachers and students. TIMSS provides a data set, which consists of several more specified data subsets. We do not focus on the students‘ performance examined in the main data subset with test

[2] https://timssandpirls.bc.edu/TIMSS2007/idb_ug.html

scores, but we focus on the data subset, which collects the information about the background of students. This data subset describes student‘s relationship to particular subjects, how the local environment suits the student‘s performance, and it includes the information about student‘s‘parents. We use the observations collected in the Czech Republic. The number of students who are participating in the poll is $n = 4854$. We have to consider that the questionnaires were answered by children. It follows that they did not complete them properly, so we were forced to reduce the date set. Further, some questions may be asked only in several classes. Therefore, we do not have complete information about our sample. We decide to use only observations, where no values are missing, therefore, after selecting only the completely filled questionnaires, the number of observations is reduced to $n = 2561$. As alternative way to omitting the observations, we propose an application of the particular method for dealing with missing data.

Originally, the data set contains 383 explanatory variables. We omit the variables where all values were missing (NA) and the variables which include identification information like the student‘s ID, country ID, a language of testing, etc. Those variables are either constant or they do not bring us any additional valuable information. Finally, we exclude variables describing the partial weight of observations, e.g. school weight factor. We only leave the total weight of each student in the data set. This variable is taken in consideration in the final model.

We select a student‘s academic expectation as our dependent variable. The students answered the question which level of education they expect to achieve. The students had option to chose one of 6 possible answers: 1 - finish upper secondary education (ISCED 3); 2 - finish post-secondary, non-tertiary education (ISCED 4); 3 - finish first stage of tertiary education (practically oriented, ISCED 5B) ; 4 - finish first stage of tertiary education (theoretically oriented, ISCED 5A, First degree) ; 5 - beyond (ISCED 5A, First degree); 6 - I do not know. The answers are denoted by numbers that increasingly represents the expected level of education. We will not provide a detailed description of the independent variables in this thesis. However, a short comment is given in Appendix. Moreover, the individual variables are marked by an acronym and the system of labeling is well explain in TIMSS 2007 User Guide (Foy and Olson, 2009).

Despite the fact the dependent variable is a factor (categorical) variable, we do not use the multinomial regression because of the complexity of the theory this technique is based on. We would like to demonstrate the selection abilities and compare the mean squared errors of our three methods and for these purposes the penalized OLS method is more than sufficient. The main goal is to show how the penalization methods work, and the linear model is a satisfactory tool for these analyses. However, for future analyses, we propose a more complex model, which better suits to the nature of the variables. The dependent variable is a factor variable, hence the multinomial regression can be a better instrument than basic linear regression. We suspect the improvement can be achieved by using some of the lasso modification. Instead of the elastic net or the basic lasso we suggest to employ the group lasso (Yuan and Lin, 2006), which is able to select correlated groups of variables. Many of our covariates are also factor variables, thus it seems as very reasonable choice. However, the theory behind this model is beyond scope of this thesis and the author‘s skills.

5.2 Analyses

We analyze the data in software **R** using the package `glmnet` (Friedman et al., 2009). Firstly, we compare the performance of the lasso, the ridge regression and the elastic net on the real data set. We do not provide a simulation study since it has been already published in (Zou and Hastie, 2005). We estimate only a tuning parameter λ for different values of parameter α, especially for values $\{0, 0.1, \ldots, 0.9, 1\}$. It follows from (16) that those values of α include not only several elastic net estimators, but also the lasso estimator and the ridge estimator. This is the reason why we distinguish the methods only by an appropriate value of α on the description of the graphs etc.

Generally, the performance of estimators can be evaluated using different criteria (Lehmann and Casella, 2006). It depends on the statistician‘s preferences, i.e. what is the important criteria in a particular problem, and what is the purpose of the study. We choose the mean squared error as the statistical criterion for the reason that it is well understandable for the most economists. Moreover, we can evaluate the trade-off between the high variance and the biased estimator comparing the results to MSE obtained from the estimation of linear model by OLS. Based on the results we show the lasso solution, the ridge solution and solution for the best performing α of the elastic net.

5.2.1 MSE

We randomly divide the data set to training and test sets. The train set represents $0.66n$ observations, where $n = 2561$, and the rest of the observations is included to the test set. We run the cross-validation on our training set 500-times considering it is a (pseudo)random process. We fit the models using the function `cv.glmnet`.

We proceed with comparison of the responses to our test set and we calculate MSE for each model. Further, we estimate the linear model by the simple OLS method. We keep all 205 variables in the model. Thus, we obtain $\text{MSE}_{OLS} = 35.54381$. Whereas, MSE for the penalized OLS method lies beeween 2.6 and 3.1. This gives us an empirical proof that the penalization techniques bring us some benefit in these analyses. It follows the penalization methods are a suitable choice for this data type. The cross-validation curves are illuminated in Figure 4. We can see, that MSE of the penalization techniques is significantly lower than MSE_{OLS}, which supports the statement from the previous chapter than the trade-off between an unbiased estimator and a magnitude of variance can be prevailed by the low variance. However, this is not a rule. In certain situation, the OLS estimator can actually outperform the penalized OLS estimators (Hansen, 2016). Notice, that all these penalization estimators reduce to the OLS estimator for $\lambda \longrightarrow 0$, but this is not depicted in Figure 4. Figure 4 illuminates only part of the cross-validation curve.

We summarized our results using the box plots in Figure 5. Overall, the elastic net outperforms both the lasso and the ridge regression. We suggest the value $\alpha = 0.1$ as the optimal choice. The reasons are following, the choice is based not only the lowest mean or the lowest median but considering the variability as well. It makes sense that the option with lower interquartile range (IQR) is more trustworthy than the option with higher IQR. This option, $\alpha = 0.1$, satisfies well both of those criteria. Moreover, the lasso ($\alpha = 1$) prevails over the ridge regression ($\alpha = 0$) when we decide to opt for the criteria of IQR as the crucial one. In the following section we estimate the model with the suitable choice of parameters λ and α based on the results obtained in this section.

5.2.2 Results

Using the λ chosen by cross-validation, we are able to introduce the final sparse model. Firstly, Figure 6 shows the whole lasso coefficient path. The single lines represent individual variables. The number on the top indicates the number of active variables for each value of λ (horizontal axis). The size of coefficient is demonstrated on vertical axis. The elastic net coefficient path looks similarly, on the other hand the ridge regression coefficient path is not so interesting because no variable selection is provided. Thus, all estimates are active during the whole path, but only the values are changing. Consequently, we choose the best estimate of the parameter λ. We follow one standard error rule and we estimate the model for $\alpha = 0.1$, i.e. the elastic net, and for $\alpha = 1$, i.e. the lasso, hence we want to compare the mixed penalization to the strict ℓ_1-penalization. The value of α is opted for its performance in section 5.3.1.

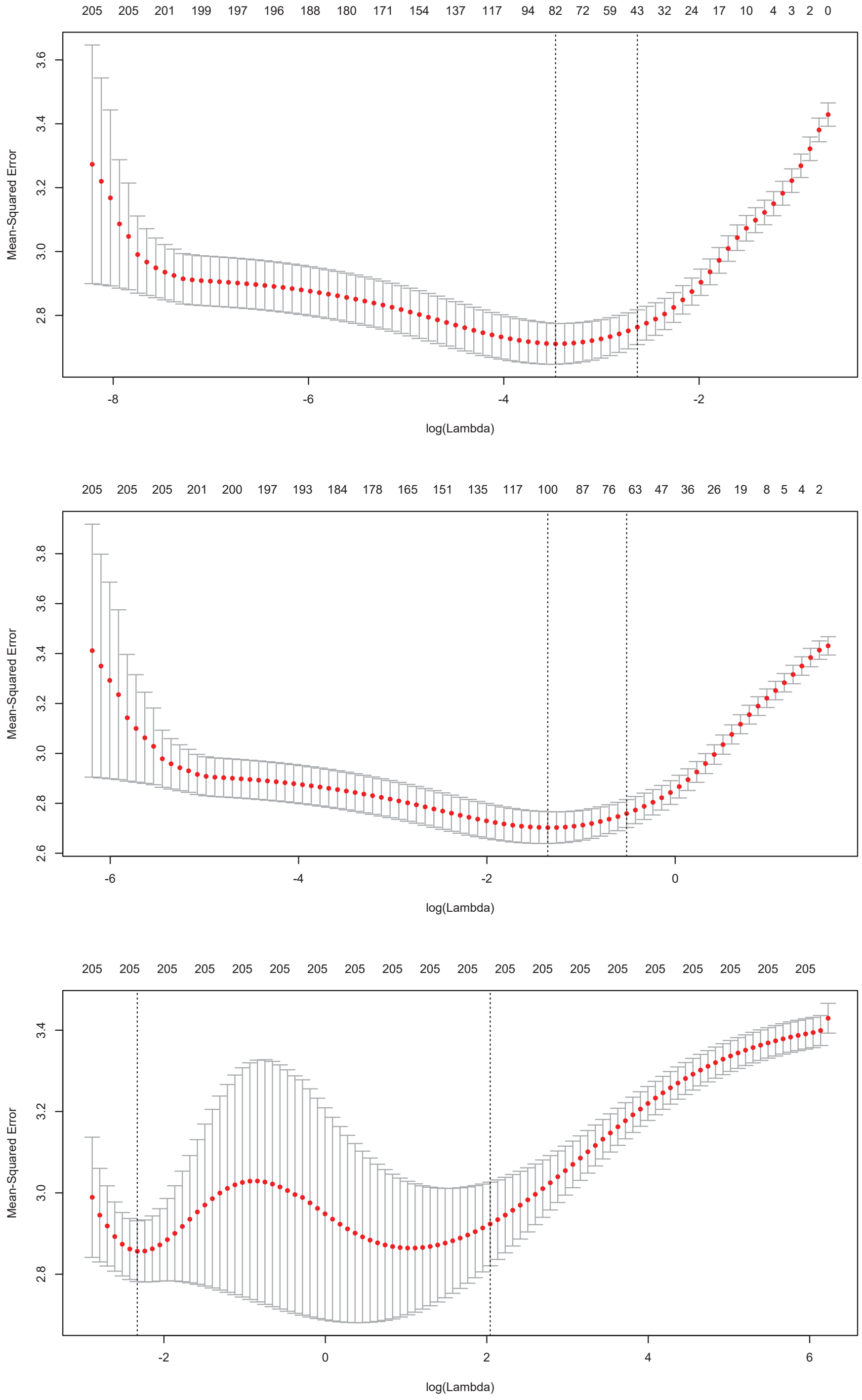


Figure 4: The cross-validation curves (lasso, elastic net, ridge regression) with the marked value of λ for the minimum value of CV curve (left) and the value of λ chosen by one standard error rule (right). The values of λ are marked by vertical lines.

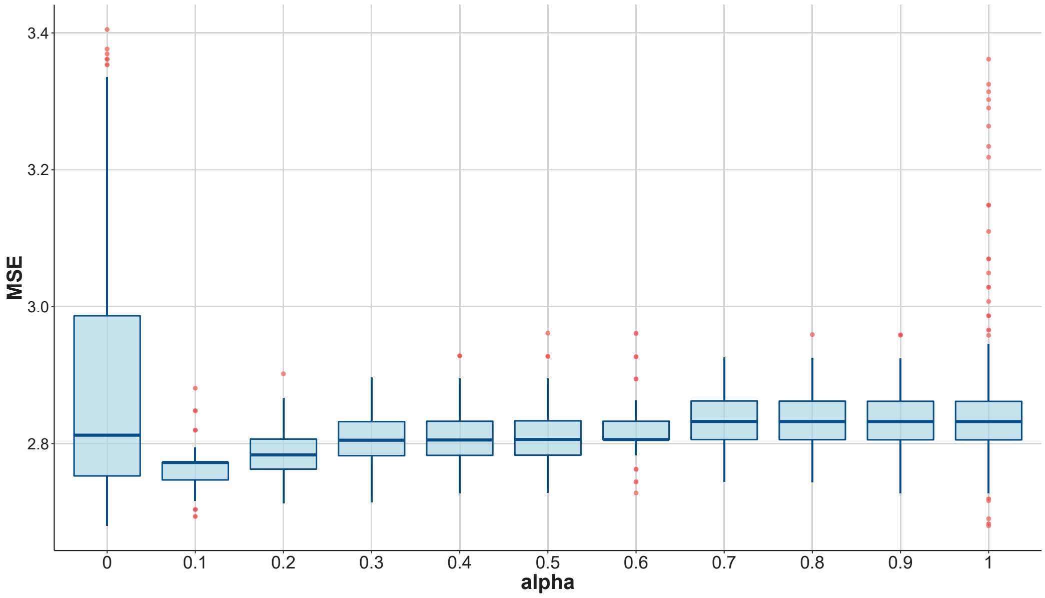


Figure 5: Comparison of MSE for the elastic net

The lasso allows 43 coefficients to be non-zero despite the elastic net is more gentle as it selects 67 relevant variables. These results can be expected due to the fact the elastic net (for $\alpha = 0.1$) is closer to the ridge regression.

We proceed with a brief model interpretation. We include Table 2 (Appendix) of selected variables and provide a comment of the results. Both the lasso and the elastic net proved to be able to determine the sex of student. The girls have higher expectation about their future education. But it is also dependent on parents‘ education, the more educated parents, the higher are goals of their child. The parents also ensure the condition at home. Children living in the house with an internet connection and a huge library are more likely to expect to attain a higher education level than children, who lack those possibilities. Notice that the elastic net selects more variables related to home possession, e.g. having a study desk, a calculator or a computer. This can be connected with social level of the families and parent‘s job; however, this information was not collected. Consequently, the lasso selects 6 variables describing students background at school. Reading more books

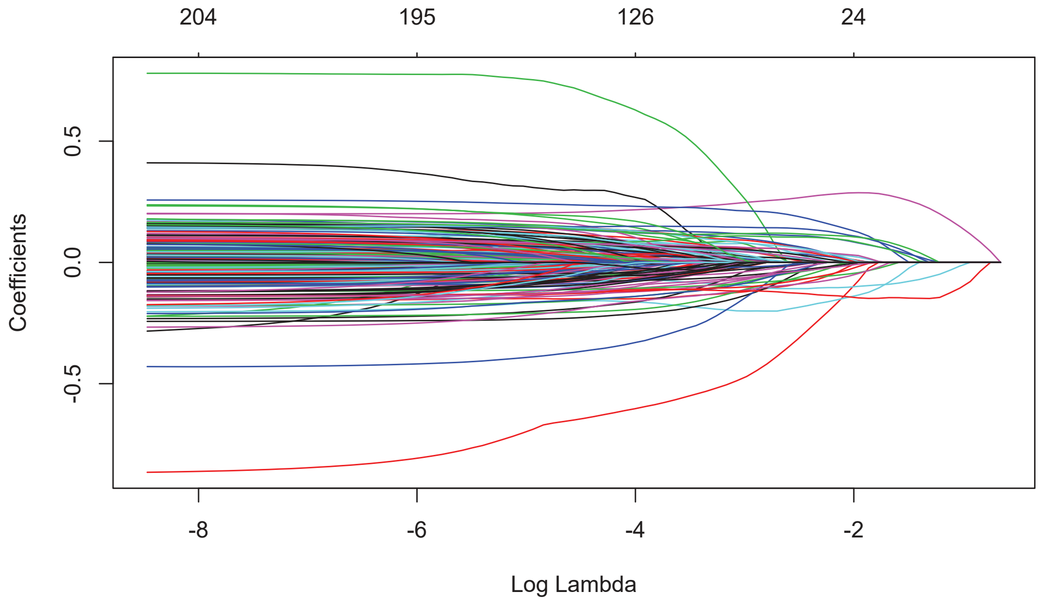


Figure 6: The lasso coefficient path

and enjoying being at school, has a positive effect on student‘s expectations, on the other hand the students with the higher expectation are more likely to be left out of actives by other students.

When focusing on the specific subjects, we can observe that both the lasso and the elastic net selects the most variables related to mathematics, that is 10 by the lasso and 13 by the elastic net. The optimal choice for another subject (biology, earth, chemistry, physics) lies between 7 to 10 variables for the elastic net, and between 4 to 6 for the lasso. Thus, the elastic net includes more variables concerning the other subjects than just mathematics. Generally, students, who disagree with the statement that learning a particular subject is difficult for them and they perform well, are those who are expect to continue in their studies. These students are also those that think the science is related to daily life. From mathematical variables there are additional relevant variables, that describes the needs of mathematics in other subjects, in university, or in getting a dream job.

While the first two variables have a positive effect on student‘s studies, the mathematics ceases to be important when the students speculate about their job. On the other hand, the students are 14 or 15 years old, so they might not be completely sure about their future dream job. However, o the best of our knowledge, there was no study that would analyze this data for the Czech Republic so far. Therefore, we do not have relevant reference to compare the results with.

5.3 Final discussion

To summarize, we have demonstrated that the penalized OLS estimators prevail over the OLS estimator when MSE is used as the crucial criterion. The penalized OLS method also makes our model simpler and more transparent than in the case of the OLS method. However, it is important to say that ideally we would like to use the best subset selection to estimate the best sparse model. As it in our case consists of computation of 2^{205} sub-models, computationally a very demanding procedure, we were unable to examine this method. Therefore we are unable to confirm, whether the lasso or the elastic net provide near-optimal solutions or whether their performance can be further improved.

Conclusion

The main interest of the thesis was the introduction of the econometric method of penalized least squares. We reduced the theory only for the linear model. Namely, we presented the best subset selection, the ridge regression and the lasso. In the final chapter, we have proposed that the elastic net should outperform both - the lasso and the ridge regression. This expectation was supported by the application on the real data.

The intention of the thesis was to illuminate the performance of this approach analyzing the real economic data. We analyzed the behavioral microeconomic data collected by TIMSS. The student‘s academic expectation was selected as the dependent variable and the full model consists of more than two hundred variables.

This thesis mainly focus on comparison of quality of the prediction in application several methods. The second goal was to show that some of the penalization techniques are able to find more transparent model that includes only the relevant variables. We have decided to choose the mean squared error as the suitable statistical criterion for comparison. The optimal value of the tuning parameter for the penalized OLS method was estimated by the cross-validation. We have provided repetitively computations to ensure the results are accurate. Finally, we have estimated the models based on the results obtained by the cross-validation. The interpretation was not demonstrated in depth. However, the table and short comment to the results is included. Moreover, we proposed the complex model that could lead to better results than our simple linear model. In the case of the estimation this suggested model, the interpretation would deserve more attention.

We have concluded that the penalized OLS estimators outperforms the OLS estimator. Especially, with the elastic net ($\alpha = 0.1$) providing the best result. Therefore, our recommendation is to use this technique in economic research. Moreover, an application of this method creates sparse and more transparent models, and this leads to the advantage of the nicely interpretable model.

For further research we propose an application of penalized methods on generalized linear models. In economics, we can meet with logistic regression or Poisson regression. We propose that the application of methods providing sparse models can be beneficial in the economics and finance, mainly in the case where the amount of available data is still increasing.

References

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives 28*(2), 29–50.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Caner, M. and H. H. Zhang (2013). Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics,(just-accepted)*.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Epprecht, C., D. Guegan, and Á. Veiga (2013). Comparing variable selection techniques for linear regression: Lasso and autometrics. *Documents de travail du Centre d'Economie de la Sorbonne*.

Fan, J., J. Lv, and L. Qi (2011). Sparse high-dimensional models in economics. *Annu. Rev. Econ. 3*(1), 291–317.

Foy, P. and J. F. Olson (2009). *TIMSS 2007 international database and user guide*. TIMSS & PIRLS International Study Center, Boston College.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics Springer, Berlin.

Friedman, J., T. Hastie, and R. Tibshirani (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version 1*(4).

Gillen, B. J., M. Shum, and H. R. Moon (2014). Demand estimation with high-dimensional product characteristics. In *Bayesian Model Comparison*, pp. 301–323. Emerald Group Publishing Limited.

Hansen, B. E. (2016). The risk of james–stein and lasso shrinkage. *Econometric Reviews 35*(8-10), 1456–1470.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Lee, S., M. H. Seo, and Y. Shin (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(1), 193–210.

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

Nardi, Y. and A. Rinaldo (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis 102*(3), 528–549.

Nasekin, S. (2013). High-dimensional lasso quantile regression applied to hedge funds' portfolio. Master's thesis, Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.

Persons, W. M. (1910). The correlation of economic statistics. *Publications of the American Statistical Association 12*(92), 287–322.

Schneider, U. and M. Wagner (2012). Catching growth determinants with the adaptive lasso. *German Economic Review 13*(1), 71–85.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R. J. et al. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics 7*, 1456–1490.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives 28*(2), 3–27.

Wang, H., G. Li, and C.-L. Tsai (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(1), 63–78.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zhang, Y., R. Li, and C.-L. Tsai (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association 105*(489), 312–323.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Zvára, K. (2008). *Regrese.* MATFYZPRESS, Praha.

Appendix

Data description

The final data set includes 206 variables. The dependent variable is the student‘s academic expectation (BS4GHFSG). The set of independent variables can be divided to the several groups. The set contains questions concerning students‘ circumstances at home and school, information about parents‘ education and nationality, classroom experience, out school activities, and questions related to several subjects: mathematics, biology, earth, chemistry and physics. We have 30 variables for each subject, the questions are repeating for a particular subject. The topics of the questions are following: self-perceptions of aptitude for and attitudes toward the subjects, homework, importance of subject, computer use, etc.

The variables describing general background are denoted by “G” on the fourth position, e.g. BS4GBOOK - a number of books at home. The mathematics is denoted by “M”, e.g. BS4MAUNI - the importance of mathematics to get to the university. A similar notation is used for remaining subjects (biology - “B”, earth - “E”, chemistry - “C”, physics - “P”). The variables are often factors with 4 or 6 possible answers. The most frequent types of questions are following. Firstly, some opinion is stated and students answer on scale from *totally agree with this opinion* to *totally disagree with this opinion.* Secondly, the students answer how often they do a particular activity. The special variable TOTWGT matches the weight to the observations.

This short description can be helpful for understanding the table of results, which is on the next page. Table 2 summarizes selected variables, the elasic net estimates and the lasso estimates.

Table 2: Summary of the results obtained in Section 5.2.

Variable	Elastic net	Lasso	Variable	Elastic net	Lasso
ITSEX	-0.139633307	-0.1773876	BS4BASTR	0.037369539	0.01831002
BS4GBRTM	0.002146847	-	BS4BAQKY	-0.038706487	-0.05233415
BS4GBOOK	0.214609006	0.2668449	BS4BHOBS	0.039525279	0.03493403
BS4GTH01	0.062181152	-	BS4BHDEI	0.010716221	-
BS4GTH02	-0.054662507	-	BS4BHWGO	0.011048925	-
BS4GTH03	0.034054119	-	BS4BHTEX	0.016562872	-
BS4GTH05	-0.187740963	-0.2153801	BS4BHWPO	0.04968980	0.05449062
BS4GMFED	0.060803009	0.07275888	BS4EACLM	0.035713538	0.01727839
BS4GFMED	0.021279988	0.01378366	BS4EASTR	0.078630669	0.09165192
BS4MAWEL	-0.010889604	-	BS4EAHDL	-0.004365147	-
BS4MACLM	0.061615930	0.06030820	BS4EHDEI	0.003182101	-
BS4MASTR	0.096844575	0.1412106	BS4EHPEI	0.007273947	-
BS4MAQKY	-0.112465283	-0.1545227	BS4EHWG	0.003545824	-
BS4MALIK	-0.022773416	-	BS4EHTEX	0.067176894	0.07581345
BS4MAOSS	-0.070501442	-0.07449366	BS4EHLAW	0.058188969	0.06109046
BS4MAUNI	-0.033840706	-0.02198885	BS4EHMDL	-0.025428013	-0.02163604
BS4MAGET	0.026657259	0.01418538	BS4EHWPO	0.016206786	0.001409634
BS4MHASM	-0.006244335	-	BS4CAMOR	0.025201786	0.01660375
BS4MHWFD	-0.018002199	-0.00292859	BS4CASTR	0.000886579	-
BS4MHGSA	0.010837172	-	BS4CAGET	0.015656270	0.00400680
BS4MHEXP	-0.037839375	-0.03281718	BS4CHWGO	0.001059031	-
BS4MHCOM	0.210067756	0.2574218	BS4CHMDL	-0.042059561	-0.03788214
BS4MCSWM	0.019502517	0.00869262	BS4CHHQT	0.059584835	0.06776088
BS4GALBS	-0.032466400	-0.02015659	BS4CHCOM	0.015412517	-
BS4GATTB	0.083255263	0.08529804	BS4PACLM	0.072697486	0.09657467
BS4GATSB	0.001247536	-	BS4PASTR	0.044403011	0.02766119
BS4GHURT	0.046402455	-	BS4PABOR	-0.008689371	-
BS4GMFUN	0.045536047	0.04378090	BS4PHTEX	0.003127259	-
BS4GPLFD	-0.041611274	-0.04260702	BS4PHLAW	-0.044464600	-0.03839998
BS4GWKPJ	-0.055679491	-0.05997411	BS4PHMDL	-0.005824456	-
BS4GREBO	0.062301330	0.05641128	BS4PHHQT	0.020660363	0.00480962
BS4GUSIN	0.000841060	-	BSDMSCM	-0.072178297	-0.03278479
BS4BAWEL	-0.006745357	-	BSDPSCS	-0.032497767	-0.01848127
BS4BACLM	0.013494964	0.00003164			