University of Duisburg-Essen
Faculty of Business Administration and Economics
Chair of Econometrics

Bayes Seminar

Advanced R for Econometricians

Seminar Paper

Submitted to the Faculty of
Ökonometrie
at the
University of Duisburg-Essen

from:

Jens Klenke

Reviewer:	Christoph Hanck
Deadline:	Jan. 17th 2020

Name:	Jens Klenke
Matriculation Number:	3071594
E-Mail:	jens.klenke@stud.uni-due.de
Study Path:	M.Sc. Economics
Semester:	5$^{\text{th}}$
Graduation (est.):	Winter Term 2020

Contents

List of Figures **II**

List of Tables **II**

List of Abbreviations **II**

1 Introduction **1**

2 Theory of Bayesian inference **2**

3 Data description **3**

4 Methodic Procedure **5**

4.1 Linear Model 5

4.2 Least Absolute Shrinkage and Selection Operator (LASSO) . 5

4.3 Bayesian Lasso 6

4.3.1 Gibbs Sampler 6

5 Estimation of the Bayesian Lasso **6**

6 Posteriod-based Estimation and prediction **6**

7 Residuals and Sensitive Analysis **6**

8 Discussion and further research **6**

9 References **6**

List of Figures

1 Histograms of player values and log player values 4

List of Tables

1 Summary of some important variables 3

List of Abbreviations

LASSO	Least Absolute Shrinkage and Selection Operator	I
RMSE	Root Mean Squared Error	3
MCMC	Markov chain Monte Carlo	2

1 Introduction

In recent years, the LASSO method of TIBSHIRANI has emerged as an alternative to ordinary least squares estimation. The success of the method is mainly due to its ability to perform both variable selection and estimation. As already Tibshirani pointed out in his original paper the standard LASSO model can be interpreted as a linear regression with a Laplace prior. PARK and CASELLA where the first to implement the Bayesian lLASSO »using a conditional Laplace prior specification«.

Our goal is to compare the result of the Bayesian LASSO with normal LASSO method and an ordinary least square estimation. The focus is particularly on the number of non-significant parameters in the linear model or, in case of the LASSOs the parameters equal to zero.

2 Theory of Bayesian inference

The Bayesian (inference) statistics based on the Bayes' theorem for events.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

For Bayesian statistics the event theorem gets (2.1) rewritten to apply it to densities. Where $\pi(\theta)$ is the prior distribution - which could be gained from prior research or knowledge, $f(y|\theta)$ is the likelihood function, and $\pi(\theta|y)$ is the posterior distribution, we then get the following.

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)} \tag{2.2}$$

From (2.2) the advantages and disadvantages of Bayesian statistics compared to frequentist statistics can directly be retrieved. One major adavantage is that the Bayesian approach can account for prior knowledge and points out a philosophical difference to the frequentist approach - that the obtained data stands not alone. Another, key difference and advantage is that in the Bayesian world the computation are made with distributions and this leads to a better information level than just the computation of the first and second moment. The computation of distributions are also the greatest disadvantages or more neutral the biggest problem of the Bayesian approach because in high dimensional problems the computation takes a lot of times or is sometimes even not possible. A solution to that is that with newer and better computers it is possible to simulate the integrals with a Markov chain Monte Carlo (MCMC) method. GHOSH

3 Data description

We collected the data from the online database platform *kaggel.* ZITATE The dataset included 6 years of data for all players which where included in the soccer simulation game *FIFA* from *EA Sports*. We dicided to just keep the data for 2019 and 2020. The Data for 2019 will be used for the estimation of the differen models whereas the 2020 data will be used to compare the quality of the models with an out of sample Root Mean Squared Error (RMSE).

A fundamental problem of the data set was that goalkeepers are systematically rated differently than field players. Therefore, in the subcategories of *overall* all field player categories were assigned to NAs. Conversely, all field players have NAs in all goalkeeper categories. Because the algorithm of LASSO in R cannot handle NAs, they are set to zero for all models.

Table 1: Summary of some important variables

	year	N	mean	sd
value_eur	2019	17538	2473043.68	5674963.22
	2020	18028	2518484.58	5616359.21
wage_eur	2019	17538	10085.87	22448.70
	2020	18028	9584.81	21470.29
overall	2019	17538	66.23	7.01
	2020	18028	66.21	6.95
age	2019	17538	25.17	4.64
	2020	18028	25.23	4.63
potential	2019	17538	71.40	6.15
	2020	18028	71.56	6.14

As one can see in Table 1 the differences between the versions for the most important variables are considerable small. From 2019 to 2020 the mean player *value* (response variable) increased by 4.54409×10^4 which is about 1.8 per cent or 0.01 standard deviations. Simular results are observabel for the probly most importanten righthand variables *wage* and *overall* with a difference in the means of -0.02 and -0.003 standard deviations between 2019 and 2020.

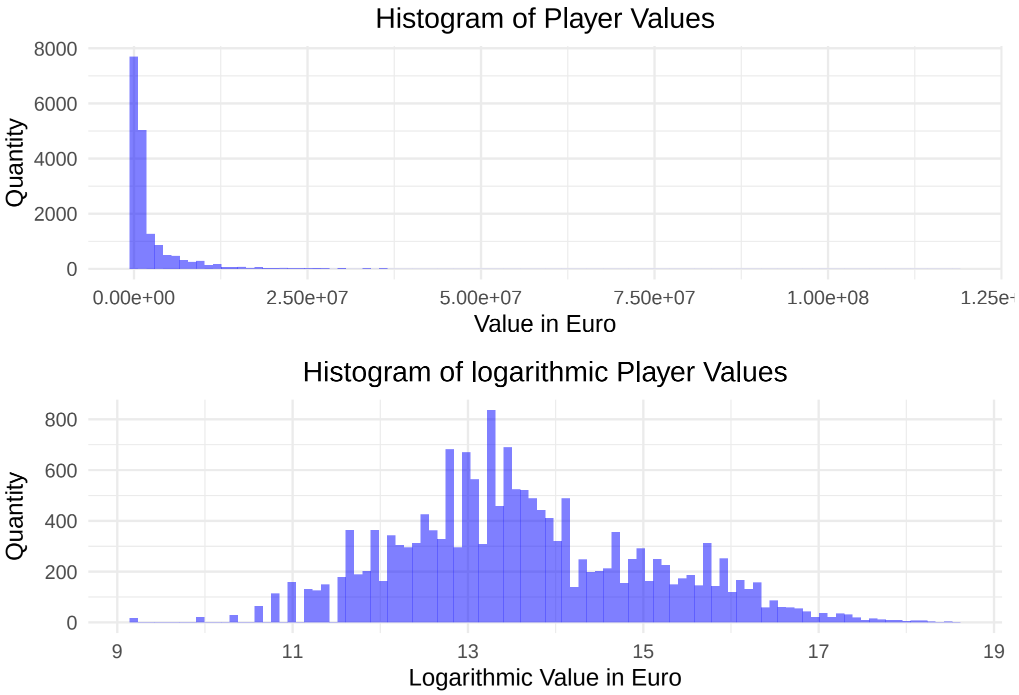

Figure 1: Histograms of player values and log player values

As you can see in Figure 1 Multikollinarity

4 Methodic Procedure

To compare the Bayesian LASSO we will analyse the data also with a linear multivariate model, and the frequentist LASSO. We wil start with the linear model and will modifie the model equations step by step forward the bayesian version.

4.1 Linear Model

The frequentist multivariate regression model has the follwing model equation.

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{4.1}$$

Where the coefficient will be estimated by the ordinary least square method, which means that $\boldsymbol{\beta}$ should be chosen so that the *Euclidean norm* ($||\mathbf{y} - \mathbf{X}\beta||_2$) is minimal. This yields in the conditon for the estimation of coefficients:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{\beta_0} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{\beta_0} - \boldsymbol{X\beta}) \tag{4.2}$$

4.2 Least Absolute Shrinkage and Selection Operator (LASSO)

In the LASSO method the model equation is the same as the equation for the multivariate but the condition for the optimization of the estimators has an additional punishment term.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^T \left(\boldsymbol{y} - \boldsymbol{X\beta}\right) + \lambda \sum_{i=1}^{p} |\beta_j| \tag{4.3}$$

4.3 Bayesian Lasso

4.3.1 Gibbs Sampler

5 Estimation of the Bayesian Lasso

6 Posteriod-based Estimation and prediction

7 Residuals and Sensitive Analysis

8 Discussion and further research

9 References

–

Eidesstattliche Versicherung

Ich versichere an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Essen, den ____________________ ______________________________

Jens Klenke