

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<sup>th</sup>

Graduation (est.): Winter Term 2020

# Contents

List of Figures	II
List of Tables	II
List of Abbreviations	II
<b>1 Introduction</b>	<b>1</b>
<b>2 Theory of Bayesian inference</b>	<b>2</b>
<b>3 Data description</b>	<b>3</b>
<b>4 Models used</b>	<b>5</b>
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
<b>5 Estimation of the Bayesian Lasso</b>	<b>6</b>
<b>6 Posterior-based Estimation and prediction</b>	<b>6</b>
<b>7 Residuals and Sensitive Analysis</b>	<b>6</b>
<b>8 Discussion and further research</b>	<b>6</b>

## List of Figures

1	Histograms of player values and log player values . . . . .	4
---	---	---

## List of Tables

1	Summary of some important variables for the 2019 FIFA edition	3
---	---	---

## List of Abbreviations

<b>LASSO</b>	Least Absolute Shrinkage and Selection Operator . . . . .	I
<b>RMSE</b>	Root Mean Squared Error . . . . .	3
<b>MCMC</b>	Markov chain Monte Carlo . . . . .	2
<b>i.i.d.</b>	independent and identically distributed . . . . .	5

# 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 were 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.

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

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.

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

From (2.2) the advantages and disadvantages of Bayesian statistics compared to frequentist statistics can directly be retrieved. One major advantage 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 *kaggle*. (Leone, 2020) ZITATE The dataset included 6 years of data for all players which were included in the soccer simulation game *FIFA* from *EA Sports*. We decided to just keep the data for 2019 and 2020. The Data for 2019 contains 17538 datapoints will be used for the estimation of the differen models whereas the 2020 data with 18028 will be used to compare the quality of the models with an out of sample Root Mean Squared Error (RMSE).

Leone (2020)

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.

It is not very realistic that a fielder has no values in the goalkeeper categories and vice versa. However, it can be argued, at least for outfield players, that goalkeeper attributes play no role in determining market values. This argumentation does not seem to hold for goalkeepers, at least passing can be assumed to be an influential variable for the market value. Nevertheless, due to the lack of alternatives, all NAs have been replaced by zero.

Table 1: Summary of some important variables for the 2019 FIFA edition

	year	N	mean	sd
value_eur	2019	17 538	2 473 043.68	5 674 963.22
	2020	18 028	2 518 484.58	5 616 359.21
wage_eur	2019	17 538	10 085.87	22 448.70
	2020	18 028	9 584.81	21 470.29
overall	2019	17 538	66.23	7.01
	2020	18 028	66.21	6.95
age	2019	17 538	25.17	4.64
	2020	18 028	25.23	4.63
potential	2019	17 538	71.40	6.15
	2020	18 028	71.56	6.14

As one can see in Table 1 the differences between the editions for the most important variables are considerable small. For example, from 2019 to 2020 the mean player *value* (response variable) increased by  $4.54e+04$  which is about 1.8 per cent or 0.01 standard deviations. Simular results are observabel

for the probably most important right-hand variables *wage* and *overall* with a difference in the means of -0.02 and -0.003 standard deviations between 2019 and 2020.

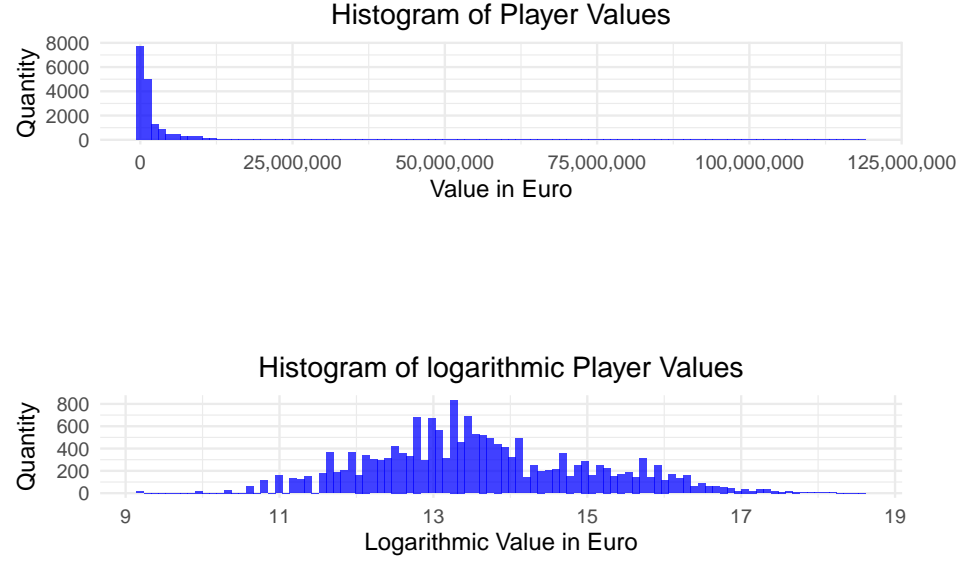


Figure 1: Histograms of player values and log player values

As can be seen for the variable value in Figure 1, this relatively strong right-skew is distributed, a similar pattern can be observed for the variable wage. Since we also estimate a linear model and this often leads to non-normally distributed residuals, these were logarithmized.

## 4 Models used

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.

$$\mathbf{Y} = \beta_0 + \mathbf{X}\beta + \epsilon \quad (4.1)$$

Where  $\mathbf{y}$  is the  $n \times 1$  response vector,  $\mathbf{X}$  is the  $n \times p$  matrix of regressors and,  $\epsilon$  is the  $n \times 1$  vecotr of independent and identically distributed (i.i.d.) errors with mean 0 and unknown variance  $\sigma^2$ . The coefficient will be estimated by the ordinary least square method, which means that  $\beta$  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:

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

### 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 in equation (4.2) has an additional punishment term. Which leads to the optimazation of:

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

for some  $\lambda \geq 0$ . This method is also often referred to as  $L_1$  -penalized least squares estimation.



Already in his original paper Tibshirani (1996) has pointed out the possibility that his methods can also be interpreted Bayesian. The LASSO estimates can be considered as posterior mode estimates with a double-exponential Laplace prior .

### 4.3 Bayesian Lasso

Park and Casella (2008) considered a fully Bayesian approach using a conditional Laplace prior of the form

$$\pi\left(\beta|\sigma^2\right)=\prod_{j=1}^p\frac{\lambda}{2\sqrt{\sigma^e}}e^{\frac{-\lambda|\beta_j|}{\sqrt{\sigma^2}}} \quad (4.4)$$

(Park & Casella, 2008)

#### 4.3.1 Gibbs Sampler

## 5 Estimation of the Bayesian Lasso

## 6 Posterior-based Estimation and prediction

## 7 Residuals and Sensitive Analysis

## 8 Discussion and further research

## References

- Leone, S.** (2020). FIFA 20 complete player dataset. Retrieved January 7, 2020, from <https://kaggle.com/stefanoleone992/fifa-20-complete-player-dataset>
- Park, T., & Casella, G.** (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681–686. doi:10.1198/016214508000000337
- Tibshirani, R.** (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. Retrieved January 6, 2020, from <https://www.jstor.org/stable/2346178>

### **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