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: | $5^{th}$ |
| Graduation (est.): | Winter Term 2020 |

# Contents

# List of Figures



# List of Tables



# List of Abbreviations

# 1 Introduction

In recent years, the LASSO method of Tibshirani (1996) 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 (2008) 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 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 computations 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 is 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 et al., 2006, p. 100) PAGE NUMBER!!

# 3 Data description

We collected the data from the online database platform *kaggel.* The dataset includes 6 years of data for all players who were included in the soccer simulation game *FIFA* from *EA Sports.* We decided to keep the data for 2019 and 2020, only. The Data for 2019 contains 17538 datapoints which will be used for the estimation of the different models whereas the 2020 data with 18028 will be used to compare the quality of the models with an out of sample RMSE. Both datasets consist of 104 variables which will not all be included in the estimations. Some Variables are just an ID or different length of names and URLs. (Leone, 2020)

A fundamental problem of the dataset consists as goalkeepers are systematically rated differently than field players. Therefore, in the subcategories of the variable *overall* all field player categories were assigned NAs for goalkeepers. Conversely, all field players have NAs in all goalkeeper categories. Because the algorithm of LASSO in R cannot handle NAs they have been 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, because is an essential asset for the passing game if the goalkeeper has possession of the ball. 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 considerably 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. Similar results are observable for the probably most important righthand variables *wage* and *overall* with a difference in the means of -0.02 and -0.003 standard deviations between 2019 and 2020.

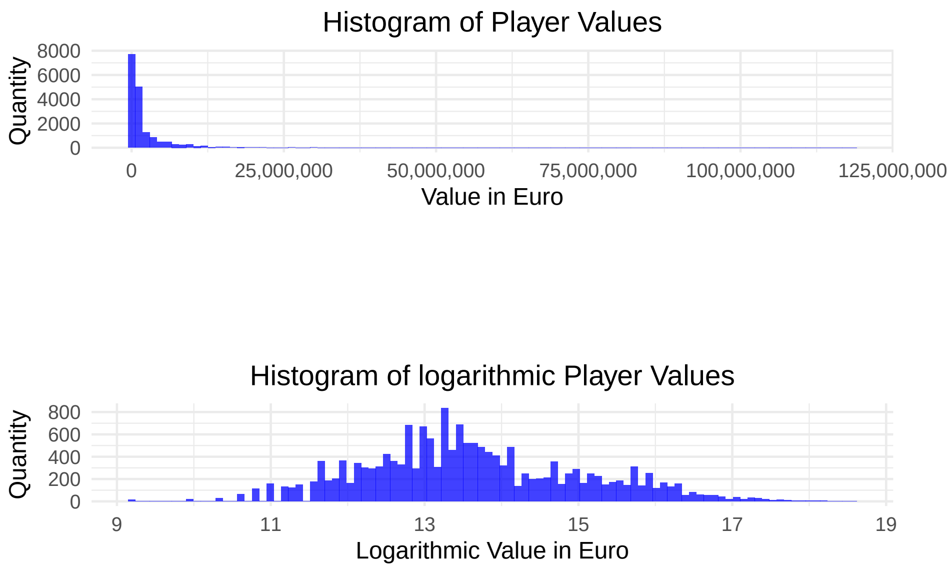


Figure 1: Histograms of player values and log player values

As 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 Used Models

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 forwards 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 $\boldsymbol{y}$ is the $n \times 1$ response vector, $\boldsymbol{X}$ is the $n \times p$ matrix of regressors and, $\boldsymbol{\epsilon}$ 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 $\boldsymbol{\beta}$ 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 in equation (4.2) has an additional punishment term. Which leads to the optimazation of:

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

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

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

(Park & Casella, 2008)

### 4.3.1 Gibbs Sampler and the full Model specification

The Gibbs Sampler is a special case of an MCMC algorithm, which is useful to approximate the combianed distribution of two or more regressors in a multidemsinoal problem.

The algortithm tries to find the approximate joint distribution and therefore the algortihm runs through the subvectors $\beta$ and draws ach subset conditional on all other values. (Gelman, 2004)

Bayesian LASSO the Gibbs sampler in the **monomvn** package in **R** (Gramacy & Turlach, 2019) samples from the following representation of the Laplace distribution

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{2\sqrt{\sigma^2}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds, \qquad a > 0 \tag{4.5}$$

Andrews and Mallows (1974)

The full model has the following hierarchical representation

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \\
\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim N_p(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau) \\
\boldsymbol{D}_\tau &= diag(\tau_1^2, \ldots, \tau_p^2) \qquad (4.6) \\
\sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim \pi\left(\sigma^2\right) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2 \\
\sigma^2, \tau_1^2, \ldots, \tau_p^2 &> 0
\end{aligned}
\tag{4.7}
$$

If $\tau_1^2, \ldots, \tau_p^2$ gets integreted out of the conditional prior on $\boldsymbol{\beta}$ , we get the form of (4.4). For $\sigma^2$ the inverse-gamma function of the form $\pi(\sigma^2) = \frac{1}{\sigma^2}$ was implemented in the **monomvn** package.

# 5 Estimation of the Models

To compare the performances of the models all three models got, obviously, estimated with the same regressors. We included as righthand variables: *log_wage* , *age*, *height_cm*, *weight_kg*, *overall*, *potential*, *shooting*, *contract_valid_until*, *pace*, *passing*, *dribbling*, and *defending* to predict the response variable *log_value*.

## 5.1 linear Model

Table 2: Summary of the linear model

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -8.5970 | 2.9222 | -2.9420 | 0.0033 |
| log_wage | 0.0679 | 0.0025 | 26.8466 | 0.0000 |
| age | -0.1004 | 0.0008 | -119.6665 | 0.0000 |
| height_cm | 0.0012 | 0.0004 | 2.7140 | 0.0067 |
| weight_kg | 0.0001 | 0.0004 | 0.3175 | 0.7508 |
| overall | 0.2098 | 0.0008 | 266.5231 | 0.0000 |
| potential | -0.0059 | 0.0007 | -8.1962 | 0.0000 |
| shooting | 0.0049 | 0.0003 | 17.7691 | 0.0000 |
| contract_valid_until | 0.0051 | 0.0014 | 3.5164 | 0.0004 |
| pace | 0.0008 | 0.0002 | 3.6118 | 0.0003 |
| passing | 0.0019 | 0.0004 | 4.5131 | 0.0000 |
| dribbling | -0.0016 | 0.0005 | -3.4678 | 0.0005 |
| defending | -0.0017 | 0.0002 | -10.6851 | 0.0000 |

In Table 2 one can see that only 1 parameter is not significant to a 5 per cent level. The variable *overall* has, naturally, the biggest (positive) impact on the *log_value (value)*, whereas *age* has the biggest negative effect.

Table 2 also shows that some coefficients are relatively small but still significant. However, a general problem with (ordinary least squares (OLS) estimation is that with increasing sample size, many "coefficients" become significant. This is because the standard errors become smaller with increasing N, the t-statistic becomes larger, and the p-value smaller. These coefficients (e.g.: *pace*, or *passing* ) could be zero in the LASSO estimation because of the punishment term.

## 5.2 Least Absolute Shrinkage and Selection Operator (LASSO)

For the frequentists LASSO $\lambda$-parameter we used the **cv.glmnet** cross-validation function from the **glmnet** package and we got a $\lambda$ of 0.01156 Hastle (2019)

## 5.3 Bayesian Lasso

### 5.3.1 Settings of the hyperparameters

In the **blasso** function of the **R** package **monomvn** it is possible to set the hyperparameters $\lambda$, for the penalty term, and $\alpha$ and $\beta$, which are the

Table 3: Summary of the LASSO

| | Estimate |
|---|---|
| (Intercept) | 1.72337 |
| log_wage | 0.066023 |
| age | -0.089616 |
| height_cm | - |
| weight_kg | - |
| overall | 0.200689 |
| potential | 0.001541 |
| shooting | 0.004659 |
| contract_valid_until | - |
| pace | - |
| passing | - |
| dribbling | - |
| defending | -0.000213 |

shape and rate parameter for the prior. The $\lambda$ is in our case an empirical parameter which will be approximate trhough an updating Gibbs sampler. The algorithm uses the parameter of the previous sample. So iteration $k$ uses the Gibbs sampler with hyperparmeter $k-1$. For the frequentists LASSO the $\lambda$-parameter was 0.01156, so we decided to set $\lambda = 10$, since the first 25% of the MCMC are not used for the estimation and the sampler convergence rather quickly. (Gramacy & Turlach, 2019)

$$\lambda^k = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}[\tau_j^2|\boldsymbol{y}]}}$$

The expectations are replaced with averages from the previous Gibbs sampler. As Park and Casella (2008) has shown any non extrem starting value for $\lambda$ can be used. In the first setting we did not pass any parameters for $\alpha$ and $\beta$.

As one can see in Tabel 4 the *median* for all regressors are unequal to zero, whereas for the frequentist LASSO we have 6 coefficient which are directly ecluded from the model, e.g. zero. However, it is unlikely that for multidimensional bayesian model the median for a parameter is zero, since the computation depends on a Gibbs sampler. If we instead look at the 95 % credible interval we finde that 6 of these intervall include the zero.

Table 4: Summary of the Bayesian LASSO

| | median | 2.5% | 97.5% |
|---|---|---|---|
| log_wage | 0.067554 | 0.063834 | 0.071904 |
| age | -0.100611 | -0.102215 | -0.098671 |
| height_cm | 0.001264 | 0.000000 | 0.002029 |
| weight_kg | 0.000000 | 0.000000 | 0.000000 |
| overall | 0.209992 | 0.208272 | 0.211500 |
| potential | -0.006020 | -0.007445 | -0.004492 |
| shooting | 0.004856 | 0.004186 | 0.005391 |
| contract_valid_until | 0.004777 | 0.000000 | 0.008346 |
| pace | 0.000714 | 0.000000 | 0.001135 |
| passing | 0.001771 | 0.000000 | 0.002559 |
| dribbling | -0.001549 | -0.002438 | 0.000000 |
| defending | -0.001596 | -0.001960 | -0.001072 |
| variance | 0.057742 | 0.056657 | 0.058963 |
| lambda.square | 0.000124 | 0.000037 | 0.000356 |

Table 5: Summary of the Bayessian LASSO with hyperpriors

| | median | 2.5% | 97.5% |
|---|---|---|---|
| log_wage | 0.067405 | 0.061796 | 0.072935 |
| age | -0.100695 | -0.102174 | -0.099112 |
| height_cm | 0.001274 | 0.000000 | 0.002045 |
| weight_kg | 0.000000 | 0.000000 | 0.000446 |
| overall | 0.209832 | 0.208392 | 0.211174 |
| potential | -0.005893 | -0.007389 | -0.004306 |
| shooting | 0.004792 | 0.004169 | 0.005301 |
| contract_valid_until | 0.004703 | 0.000000 | 0.007556 |
| pace | 0.000497 | 0.000000 | 0.001049 |
| passing | 0.001454 | 0.000000 | 0.002508 |
| dribbling | -0.000977 | -0.002296 | 0.000000 |
| defending | -0.001553 | -0.001958 | -0.001094 |
| variance | 0.057671 | 0.056589 | 0.058846 |
| lambda.square | 0.000087 | 0.000028 | 0.000340 |

Table 6: Summary of the Bayessian LASSO with hyperpriors

| | median | 2.5% | 97.5% |
|---|---|---|---|
| log_wage | 0.067405 | 0.061796 | 0.072935 |
| age | -0.100695 | -0.102174 | -0.099112 |
| height_cm | 0.001274 | 0.000000 | 0.002045 |
| weight_kg | 0.000000 | 0.000000 | 0.000446 |
| overall | 0.209832 | 0.208392 | 0.211174 |
| potential | -0.005893 | -0.007389 | -0.004306 |
| shooting | 0.004792 | 0.004169 | 0.005301 |
| contract_valid_until | 0.004703 | 0.000000 | 0.007556 |
| pace | 0.000497 | 0.000000 | 0.001049 |
| passing | 0.001454 | 0.000000 | 0.002508 |
| dribbling | -0.000977 | -0.002296 | 0.000000 |
| defending | -0.001553 | -0.001958 | -0.001094 |
| variance | 0.057671 | 0.056589 | 0.058846 |
| lambda.square | 0.000087 | 0.000028 | 0.000340 |

# 6 Root Mean Squared Error (RMSE) and "Sensitive Analysis"

# 7 Discussion and further research

LaSt Part

# References

**Andrews**, **D. F.**, & **Mallows**, **C. L.** (1974). Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(1), 99–102. Retrieved January 11, 2020, from https://www.jstor.org/stable/2984774

**Gelman**, **A.** (2004). *Bayesian data analysis* (2. ed.). Boca Raton [u.a.], Chapman & Hall/CRC.

**Ghosh**, **J. K.**, **Delampady**, **M.**, & **Samanta**, **T.** (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. New York, Springer-Verlag. https://doi.org/10.1007/978-0-387-35433-0

**Gramacy**, **R. B.**, & **Turlach**. (2019). Monomvn: Estimation for MVN and Student-t Data with Monotone Missingness. Retrieved January 12, 2020, from https://CRAN.R-project.org/package=monomvn

**Hastle**, **T.** (2019). Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. Retrieved January 12, 2020, from https://CRAN.R-project.org/package=monomvn

**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. https://doi.org/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