Manuscript

Christoph Völtzke

2023-01-26

1. Introduction

Playing to be the superstar or playing for Win-maximization

During my engagement as a handball coach I always try to emphasize the importance of defense for the success of the team. In my experience though, especially in amateur teams, the players who are successful offensively are viewed as the top players of the team, whereas the defensive leaders earn less credits for their performance. This leads to a situation where most players are mainly interested in performing well on offense and taking a rest in the defense. However, to score a goal on offense carries a nearly identical value as a goal denied on defense (Oliver 2004). Therefore, it would be expected that a win-maximizing team values defensive efforts equally to offensive efforts. To investigate whether there are differences between offensive and defensive efforts and which efforts are valued the most I will examine game statistics like points per game or defensive rebounds per game of NBA (National basketball association) players in the season 2020/2021. To evaluate the importance of certain statistics, they are fitted in a regression model to predict the players salary in the upcoming season (2021/2022). The NBA player data is used as it is similar to handball in terms of offensive and defensive efforts, the NBA franchises are mainly interested in win-maximization, which should lead to a more attenuated valuation of offensive and defensive efforts and the data regarding salary and game statistics are available open source.

The data was drawn from the website Basketball Reference (Reference of Website). The data contains information about every player who played in the NBA in the 2020/2021 season and receives a salary in the following season (2021/2022). The main variables of interest are the salary of the players (in US-dollar) as the dependent variable and two game statistics which refer to offensive efforts as well as two game statistics which refer to defensive efforts. The two offensive game statistics are average points per game and average assists per game. The two defensive game statistics are average defensive rebounds per game and average steals per game. These predictors are used as they are in american broadcasting among the main indicators to evaluate players performance and as they can be easily separated between offensive and defensive efforts.

The main question of interest is whether there are differences in the valuation of offensive and defensive game statistics and also which of the examined game statistics are valued the most. I expect, while controlling for all included predictors, points per game to have the highest predictive power as it is the most prestigious outcome to achieve in basketball (Reference). However, due to controlling for points we expect, the effect of assists to decrease and the effect of defensive rebounds per game to be stronger effect then assists per game. Last, we expect that assists per game have a stronger effect compared to steals per game. To check whether controlling for the included predictors has an influence on the effect of assists and rebounds, we expect that in a model with only these two predictors, assists per game to have a stronger effect than defensive rebounds per game. Of further interest is whether the proposed model with all four predictors provides an adequate description of the data, or whether models with less predictors might fit better. As I still expect the defensive predictors to have an effect on salary, I expect the model with all four predictors to have a better model fit.

H_1: β_{points} & $\beta_{assists} > 0$ & $\beta_{rebounds} > 0$ & $\beta_{steals} > 0$ H_2: $\beta_{points} + \beta_{rebounds} > \beta_{assists} > \beta_{steals}$ H_3: $\beta_{assists} > \beta_{rebounds}$ H_4: A model containing offensive and defensive predictors fits better then a model only containing offensive or defensive predictors

To answer the research questions, the data is analysed with Bayesian statistics and models will be compared by means of the Deviance information criterion (DIC), the Bayes Factor (BF) and 95%-Credibility Intervals (CI). Regression parameters are obtained using Markov Chain Monte Carlo (MCMC) sampling, that is, Gibbs sampling with a Metropolis-Hastings step. Convergence is assessed by means of trace plots, density plots, autocorrelations and acceptance ratios. To check model assumptions of a linear regression, a posterior predictive check is executed. In addition to the initial analyses, I am making use of the advantages of Bayesian analyses and will include historical data and use it as prior information.

2 Method

The summary statistics for the used data are presented below (Table 1). It can be seen that there are 364 NBA players in the data set. The outcome variable salary is highly left skewed and has a mean of 9.106.553 US-dollar. For this reason we decided to use the log transformed salary as the main outcome variable in the following analyses to get a more normal distribution and also to get lower numbers as it can lead to problems in the estimation procedure if the outcome values are large. The distributions for assists per game and points per game are slightly left skewed, but nothing too severe, so for now, any deviations from normality are not suspected.

Following the linear regression model which is mainly used to evaluate the hypothesis is presented:

$$log(Salary)_i = \beta_0 + \beta_1 * Points_i + \beta_2 * Assists_i + \beta_3 * DefensiveRebounds_i + \beta_4 * Steals_i + \varepsilon_i,$$

where $\varepsilon \sim \mathsf{N}(0, \sigma^2)$.

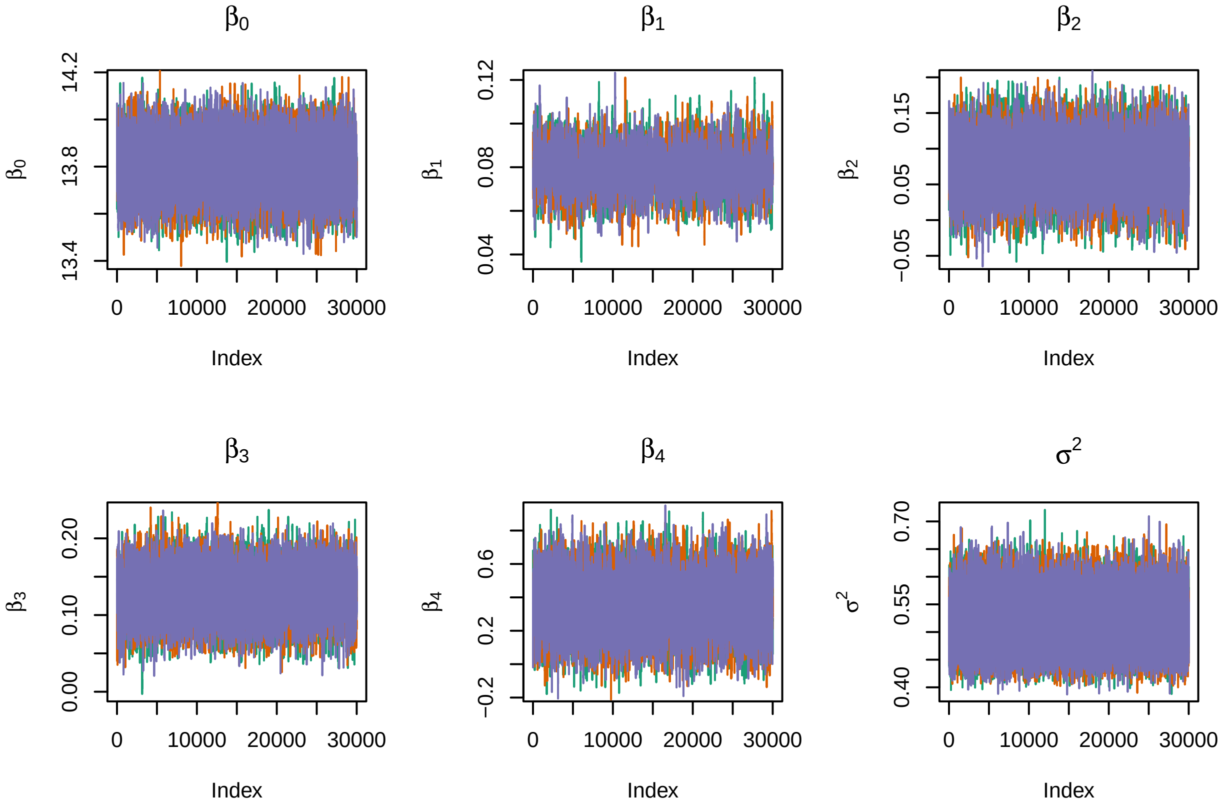
As stated before, the regression parameters are obtained by means of a Gibbs sampling procedure including a Metropolis-Hastings step, which is used to sample the estimate of β_1 . The two methods differ in their way of sampling. In summary, the Gibbs algorithm iteratively samples a parameter from a conditional posterior distribution given the sampling distribution and is used in this project to sample β_0, β_2, β_3, β_4 and σ^2. It can be seen as a special case of the MH algorithm. On the other hand, the MH algorithm samples from a proposal distribution that approximates the conditional posterior distribution and compares the conditional posterior's density at the point of the sampled value to its density of the previously sampled value. The proposal distribution used for the sampling of β_1 is normal with a mean equal to the maximum likelihood estimate of β_1 and its corresponding variance. When applying MCMC sampling it is needed to specify priors, which refer to our previous knowledge about the effect. As we didn't have any prior knowledge about the effects at the times of the analyses we use uninformative priors for the regression coefficients with a mean of 0 and a variance of = 1000. As our outcome variable is supposed to be normally distributed using normal priors for the regression coefficients yields a normal posterior distribution. Therefore, our prior are conjugate priors. The prior distribution for the variance is the inverse of a Gamma distribution with shape and rate parameters $\alpha_0 = .0001$ and $\beta_0 = .0001$, respectively. Using an Inverse-Gamma prior for the variance leads to a posterior Inverse-Gamma distribution, which also refers to conjugate priors. Another prerequisite for MCMC samplers is the use of initial values. We need initial values for all sampled parameters, which are randomly sampled from a uniform distribution in this project. To assure that the initital values don't influence the results we need to specify multiple chains, which each sample all estimates an equal amount of iterations with differing initial values. For this project we choose three chains. The initial values will be considered as non-influential if the chains converge, which will be checked in the next part.

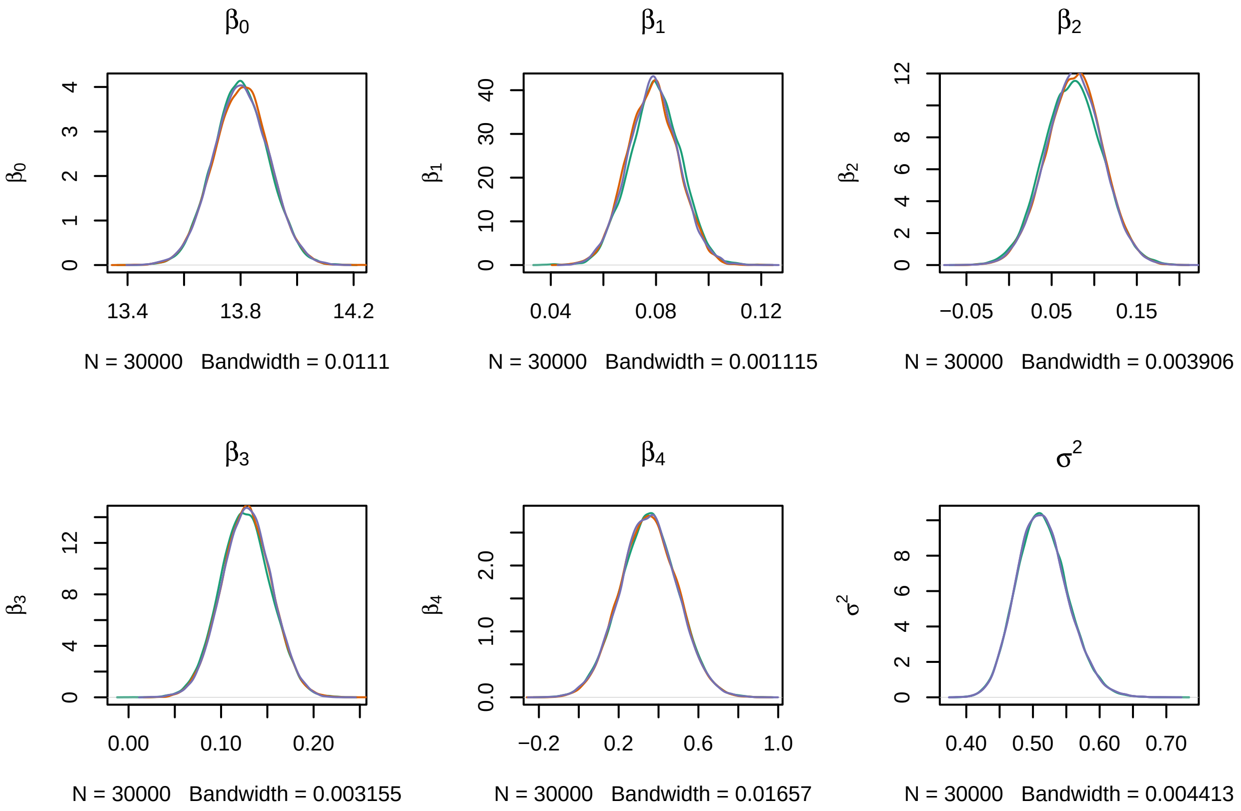
3 Results

3.1 Convergence

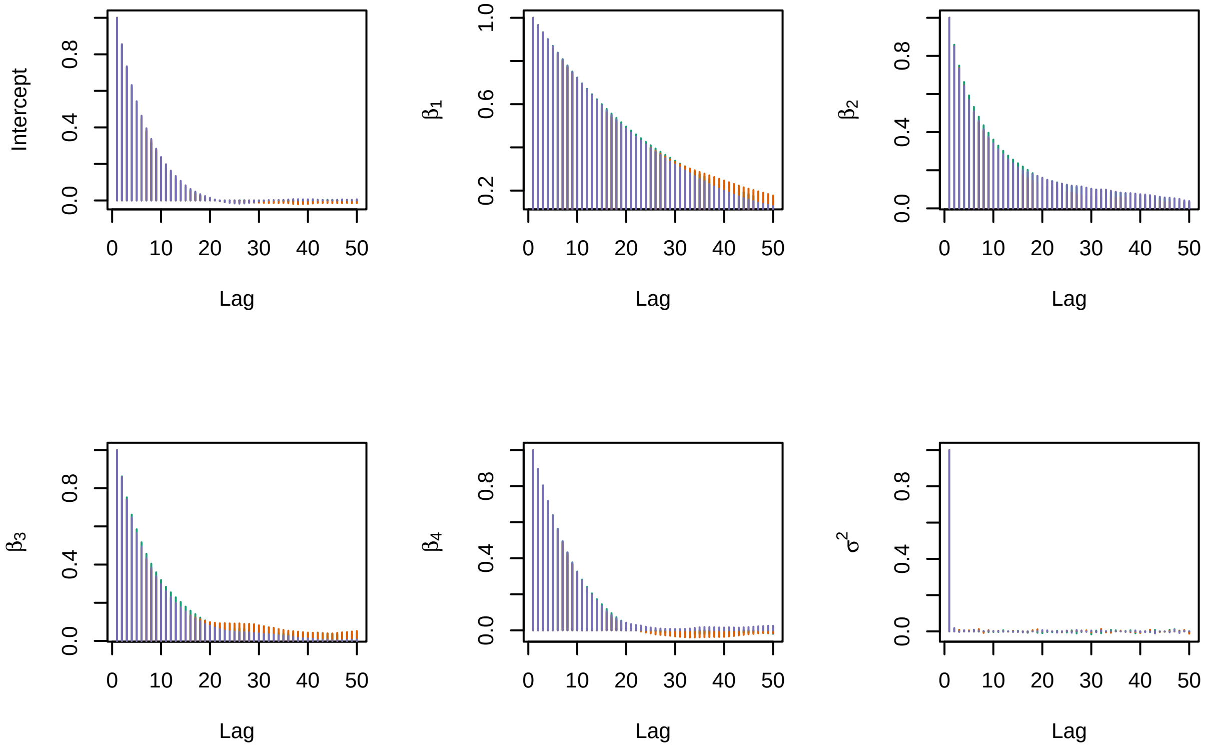
To check the convergence of the chains and to see if the estimates cover the whole posterior distribution trace plots will be used. Furthermore, density plots are assessed to check the normality of the posterior distributions. Additionally, the autocorrelations as well as the acceptance ratios are reported, which are mainly interesting for the MH step. Figure 1 depicts the trace plots of the parameters for each chain excluding the specified burn-in period. The visual inspection yields, that both the Gibbs sampler and the Metropolis-Hastings sampler move freely through the parameter space. All three chains are overlapping and seem to have an equal width, which indicates that all three chains sample throughout the same distribution, despite having different starting values. Exemplary, the density plot for β_0 and β_1 are included in Figure 2, showing a throughout normal distribution over all three chains. Figure 3 depicts the exemplary autocorrelations for β_0, β_1 and σ^2 to show that except for the residual error variance, the sampling procedure indicates a rather high autocorrelation and therefore requires a large number of iterations to cover the entire posterior distribution. The acceptance rate of 56.6% for the MH-step is however very promising. Regarding the autocorrelations we adjusted the number of iterations to be 30.000 per chain, which should be enough to cover the whole posterior distribution. Additionally, the Monte Carlo error (Naive SE) is sufficiently low for all variables. Overall, all indicators do not raise any concerns regarding non convergence.

```
convergencecheck(fourpred$sampled_values_chains,50,976)
```





β_0
β_0
0
1
2
3
4
13.4
13.8
14.2
N = 30000 Bandwidth = 0.0111
β_1
β_1
0
10
20
30
40
0.04
0.08
0.12
N = 30000 Bandwidth = 0.001115
β_2
β_2
0
2
4
6
8
12
−0.05
0.05
0.15
N = 30000 Bandwidth = 0.003906
β_3
β_3
0
4
8
12
0.00
0.10
0.20
N = 30000 Bandwidth = 0.003155
β_4
β_4
0.0
1.0
2.0
−0.2
0.2
0.6
1.0
N = 30000 Bandwidth = 0.01657
σ^2
σ^2
0
2
4
6
8
0.40
0.50
0.60
0.70
N = 30000 Bandwidth = 0.004413



```
## $traceplots
## NULL
##
## $densityplots
## NULL
##
## $autocorrelation
## NULL
##
## $acceptance_chains
## $acceptance_chains[[1]]
##       [,1]
## [1,] 1.000
## [2,] 0.566
## [3,] 1.000
## [4,] 1.000
## [5,] 1.000
## [6,] 1.000
##
## $acceptance_chains[[2]]
##       [,1]
## [1,] 1.000
## [2,] 0.564
## [3,] 1.000
## [4,] 1.000
## [5,] 1.000
## [6,] 1.000
##
## $acceptance_chains[[3]]
##       [,1]
## [1,] 1.000
## [2,] 0.568
## [3,] 1.000
## [4,] 1.000
## [5,] 1.000
## [6,] 1.000
```

3.2 Posterior predicitve check

To check whether our model was in line with two major assumptions of linear regression I decided to check the assumptions of homoscedasticity and linearity with two posterior predictive checks.

Homoscedasticity assumes that the variance of the outcome variable is independent of the predictors. This assumption is violated if the residuals and the fitted values are correlated. Therefore, the correlation between the observed residuals and the fitted values was compared with the correlation of the simulated residuals and the fitted value. This correlation is referred to as our discrepancy measure. Linearity assumes that the relationship between the predictor and the outcome is linear. Linearity is met if the residuals show no specific pattern and should lead to a mean close to zero over the whole distribution of the residuals. A deviation from 0 would indicate violations of the assumption of linearity. Therefore, I divided the distribution in three parts, took the mean of the residuals in each part and calculates the sum of the means. This sum is compared between the observed and the simulated residuals and is the second discrepancy measure.

The proportion of data sets of simulated values that have a larger correlation (larger sum) than the correlation (sum) of the corresponding observed data sets, can be seen as a measure of evidence against the null hypothesis of homoscedasticity (linearity) of the residuals. If the assumptions are not violated in the observed data set no clear trend about the proportion of the correlation (sum) should be observed. That would mean the observed data behaves as it would be expected regarding the proposed model and should be represented with a p-value around 0.5.

In this analyses 3000 data sets were simulated in an iterative procedure. Further the residuals for each set were obtained and the discrepancy measures for both mentioned posterior predictive checks were computed. The obtained value for the first check regarding the assumption of homoscedasticity has a p-value of 0.572 indicating that the observed data set has on average a slightly lower correlation between the residuals and the fitted values than the simulated residuals, which indicates no violation of homoscedasticity. The second test statistic regarding linearity has a p-value of 0.518, indicating that our observed data has a slightly lower sum of the means than the simulated data sets, which also indicates no violation of the second assumption:linearity.

Overall, both posterior predictive checks lead to the conclusion that our model fits to the data.

```
ppp(fourpred$sampled_values,y,x4,3000,976)
```

```
## $Homoscedactisty
## [1] 0.572
##
## $Linearity
## [1] 0.518
```

Table 1: Table 1: Parameter estimates for the four predictor model

	Mean	S.E.	MC-error	2.5%	Median	97.5%	Acceptance	Burn-in	Iterations
Intercept	13.80	0.10	0.00	13.61	13.80	13.99	1.00	30,000.00	90,000.00
Points per game	0.08	0.01	0.00	0.06	0.08	0.10	0.57	30,000.00	90,000.00
Assists per game	0.08	0.03	0.00	0.01	0.08	0.14	1.00	30,000.00	90,000.00
Defensive Rebounds per game	0.13	0.03	0.00	0.07	0.13	0.18	1.00	30,000.00	90,000.00
Steals per game	0.35	0.14	0.00	0.07	0.35	0.64	1.00	30,000.00	90,000.00
Variance	0.52	0.04	0.00	0.44	0.51	0.60	1.00	30,000.00	90,000.00

Note. The results are based on the MCMC sampling for the four predictor model

3.3 Parameter Estimates

```
apa_table(fourpred$estimates,
          align = c("l","c", "c", "c","c", "c", "c","c", "c"),
          note = "The results are based on the MCMC sampling for the four predictor model",
          caption="Table 1: Parameter estimates for the four predictor model"
          )
```

As no diagnostic criterion indicates violation of convergence, we can continue to interpret the parameter estimates obtained by the MCMC sampling procedure with 90000 iterations over the pooled chains and a combined burn-in period of 30000. As can be seen in Table 2, based on the estimates and the 95% credible intervals the points per game, assists per game, defensive rebounds per game and steals per game obtained in the 2020/2021 season are related to the earned log(salary) of an NBA player in the 2021/2022 season. The posterior mean of the intercept is 13.802 (95%-CCI = [13.612; 13.993]). Every additional point per game is related to a 0.079 increase in the log transformed salary. The credibility interval indicates that there is a 95% probability that the true parameter estimate is within the boundary values of 0.060 and 0.098. Further, every additional assist per game is related to a 0.076 (95%-CCI = [0.011; 0.141]) increase in the outcome variable. Next, every additional defensive rebound per game is related to a 0.127 (95%-CCI = [0.074; 0.181]) increase in the outcome variable. Last, every additional steal per game is related to a 0.352 (95%-CCI = [0.071; 0.635]) increase in the outcome variable. The mean of the residual error variance is 0.515 (95%-CCI = [0.445;0.596]).

3.4 Model comparison

The third hypothesis specified in this report assumes that a model containing offensive and defensive predictors fits better then a model only containing offensive or defensive predictors. Therefore, it is of interest to compare models which include different combinations of predictors compared to the main model including all four predictors (two offensive and two defensive). To evaluate model fit we test whether the proposed four predictor model has a better fit than three other competing models. The three models are first, a model with only two offensive statistics, second, a model with only two defensive statistics and third, a model with a combination of one offensive and one defensive statistic. The model fit will be evaluated by means of the DIC. The DIC balances the fit of the model with the complexity of the model, and further aims to provide the simplest best model. The proposed four predictor model has a DIC of 796.42. However, this value is only of importance when compared to other competing models. The first alternative model including two offensive predictors has a DIC of 822.64, the second alternative with two defensive predictors has a DIC of 899.71 and the third alternative model including the combination of one offensive and one defensive predictor has a DIC of 814.96. Therefore, considering all four DIC estimates, the DIC of the main model is the lowest indicating the best model fit compared to the other tested models. Therefore, showing that the combination of multiple offensive and defensive statistics are important to predict the salary, which indicates that different efforts within the categories of defensive and offensive efforts get all valued to a certain degree.

Of further interest is, that the model with only defensive statistics has the worst model fit. However, the model including a combination of one offensive and one defensive statistic has a better fit than only offensive statistics. This further shows, that offensive and defensive efforts seem to get valued separately and that several statistics of the same category don't lead to a big contribution.

Finally, we can accept the hypothesis that a model containing offensive and defensive predictors fits better then a model only containing offensive or defensive predictors.

3.5 Bayes Factor

To test the specified hypotheses the Bayes Factor (BF) will be used. Note, that all hypothesis were constructed a-priori and to test the informative hypothesis the data has to be standardized to get meaningful results. As there are no equality constraints in the hypotheses a sensitivity analyses is not necessary. The following two informative hypotheses will be evaluated based on the main four predictor model, thus, controlling for the effect of the other predictors in the analyses:

H_1: $\beta_{points} > 0$ & $\beta_{assists} > 0$ & $\beta_{rebounds} > 0$ & $\beta_{steals} > 0$ H_2: $\beta_{points} > \beta_{rebounds} > \beta_{assists} > \beta_{steals}$

The BF of the first hypothesis H_1 was equal to 598.61, indicating 598.61 times more support as compared to the unconstrained hypothesis. Therefore, strongly supporting the hypothesis that all four predictors have a positive influence on the salary of NBA players in the upcoming season. Additionally, the posterior probabilities including the fail safe hypothesis can be calculated. Fail safe hypothesis are added to the set of hypotheses, to prevent placing to much confidence in a hypothesis that is inappropriate. The posterior probability of H_1 is 0.998, showing that our hypothesis receives the most support as it can be interpreted as the relative support for the hypothesis on a 0-to-1 scale. The BF for H_2 is equal to 13.45, indicating 13.45 times more support as compared to the unconstrained hypothesis. This supports our hypothesis that the effect of points is highest, followed by the effect of rebounds, followed by the effect of assists and last the effect of steals, when controlling for the other predictors. Furthermore, the posterior probability including the fail safe hypothesis is 0.93, showing also a strong relative support for H_2.

To evaluate the hypothesis of assists per game to have a stronger effect than rebounds per game when not controlling for points and steals will be evaluated with a separate model. H_3: $\beta_{assists} > \beta_{rebounds}$

The BF for H_3 is equal to 1.63, indicating 1.63 times more support as compared to the unconstrained hypothesis. This only to a certain degree supports our hypothesis that the effect of assists is higher than the effect of rebounds as teh BF is lower than expected. Furthermore, the posterior probability including the fail safe hypothesis is 0.62, showing also a high uncertainty about the hypothesis. However, the trend still supports H_3, but follow up research or updating is needed.

3.6 Using historical data

One of the main advantages of the Bayesian approach is that it allows to incorporate historical data as prior information. For my data it is rather easy to incorporate historical data as the game statistics and the information regarding the salary of previous years is also accessible. Therefore, I decided to take the game statistics from the season 2016/2017 and the salary of the season 2017/2018 as a difference of four years should exclude major overlaps of the same players. Next, I fitted a linear regression with the same four game statistics as predictors and log(salary) as the outcome to get statistical point estimates and quantified uncertainties, to inform my currently used prior distribution. With this included prior information for each of the four predictors I obtained new parameter estimates using the Bayesian approach, which are depicted in Table 3. When comparing the estimates of Table 3 to Table 2 it can be seen that the credible intervals for all predictors and the intercept are smaller, the estimates slightly change as well. Especially, in case of the Steals per game the estimate changes from β_{steals}=0.352 to β_{steals}=0.402. Therefore, using prior knowledge seemed to help for obtaining more credible and more accurate results. Additionally, the DIC of the model

Table 2: Table 2: Parameter estimates for the four predictor model including priors

	Mean	S.E.	MC-error	2.5%	Median	97.5%	Acceptance	Burn-in	Iterations
Intercept	13.75	0.09	0.00	13.57	13.75	13.93	1.00	30,000.00	90,000.00
Points per game	0.08	0.01	0.00	0.06	0.08	0.09	0.56	30,000.00	90,000.00
Assists per game	0.07	0.03	0.00	0.01	0.07	0.14	1.00	30,000.00	90,000.00
Defensive Rebounds per game	0.14	0.03	0.00	0.09	0.14	0.19	1.00	30,000.00	90,000.00
Steals per game	0.40	0.14	0.00	0.14	0.40	0.67	1.00	30,000.00	90,000.00
Variance	0.52	0.04	0.00	0.44	0.51	0.60	1.00	30,000.00	90,000.00

Note. The results are based on the MCMC sampling for the four predictor model including priors

including the informative priors is slightly lower with DIC=795.84 compared to DIC=796.41 in the original model.

```
apa_table(fourpred_i$estimates,
          align = c("l","c", "c", "c","c", "c", "c","c", "c"),
          note = "The results are based on the MCMC sampling for the four predictor model including pri
          caption="Table 2: Parameter estimates for the four predictor model including priors"
          )
```

From a practical viewpoint,having the opportunity to update the model in order to get more credible estimates (increased uncertainty due to the priors) yields in long-term more accurate results. Additionally, directly adding information is more intuitive and easy than having to compare multiple studies indirectly as it is the normal procedure in frequentist meta analyses. Therefore, it seems easier to cross-validate and accumulate scientific evidence with Bayesian procedures, which should be one of the main goals of scientific practices.

4. Conclusion

Overall, it can be concluded that all examined game statistics seem to have a positive influence on the salary in the next season. Moreover, points as the main offensive game statistic are valued the most, while defensive statistics are of lower interest. Furthermore, secondary offensive statistics (assists) are valued less than the main defensive statistic (defensive rebounds) when controlling for points and steals, but seem to be valued more when other predictors are not examined. This allows to suggest that being the best in all offensive efforts, but bad in all defensive efforts should get valued less than being good in certain offensive and defensive efforts equally. This hypothesis should be further examined in future research. Moreover, these results might hold for NBA players and the examined statistics, but might not hold for the amateur level. Further research and the application within handball are needed to draw further conclusions. As a handball coach this still gives me some hope to back up the motivation of my players with some scientific findings, so they will give their best in defense and don't only focus on offense.

Regarding the performed analyses, using a frequentist approach would have yielded the same conclusion, but it would lack flexibility. This flexibility can be shown by means of the incorporated historical data, the use of the Bayes factor, obtaining a complete posterior distribution with credible intervals as well as the opportunity to create a test statistic, which allows to specifically test if the data matches the proposed model. In this report I am making use of all of these mainly Bayesian features and feel that my results are more credible compared to just performing a linear regression with point estimates in a matter of minutes.