Offense or Defense: How to become the superstar?

Christoph Völtzke - 9769870
c.voltzke@uu.nl
15/06/2022

In my experience as a handball coach of an amateur team, the players who are successful offensively are viewed as the top players of the team, whereas the defensive leaders earn less credits for their efforts. This leads to a situation where most players are mainly focused on performing well on offense, while taking a rest in the defense, which ultimately can lead to the loss of the game. 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 in the valuation of offensive and defensive efforts and to see which efforts are valued the most game statistics like points per game and rebounds per game of NBA (National basketball association) players in the season 2020/2021 are examined and fitted in a regression model to predict the players salary in the 2021/2022 season.

NBA player data is used as it is similar to handball in terms of offensive and defensive efforts, the NBA franchises are focused on win-maximization and the data is freely available.
The data was drawn from the website Basketball Reference (https://www.basketball-reference.com/). The data contains information about all 364 players who played in the NBA (2020/2021) and received a salary in the next season. The outcome of interest is the salary of the players (in US-dollar). The outcome variable salary is highly right skewed and has a mean of 9.106.553. For this reason I used the log transformed salary as the main dependent variable in the analyses to get a more normal distribution and also to deal with lower values, as high values might lead to problems in the sampling procedure. The predictors[1] are two offensive game statistics: average points per game and average assists per game and two defensive game statistics: average defensive rebounds per game and average steals per game. These predictors are among the main indicators to evaluate players performance in the NBA and 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 to have the highest predictive power. Due to controlling for points I expect, the effect of assists to be lower than the effect of defensive rebounds. I still expect that assists have a stronger effect than steals. To check whether controlling for the included predictors has an influence on the effect of assists and rebounds, another model with only these two predictors is fitted, where I expect assists to have a stronger effect than defensive rebounds. 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 the best fit of the compared models.

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: $\beta_{assists} > \beta_{rebounds}$

[1] The investigated game statistics are abbreviated by points, assists, (defensive) rebounds and steals in this report. All refer to the averages per game of each included NBA player.

2 Method

To answer the research questions, the data is analyzed with Bayesian statistics. Models will be evaluated by means of the Deviance information criterion (DIC), the Bayes Factor (BF) and 95%-Credibility Intervals (CI). Convergence is assessed by means of trace plots, autocorrelations and acceptance ratios. To check model assumptions of a linear regression, two posterior predictive checks are 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.

Regression parameters are obtained by means of a Gibbs sampling procedure including a Metropolis-Hastings (MH) step, which is used to sample the estimate of β_{points} . 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, $\beta_{assists}$, $\beta_{rebounds}$, β_{steals} and σ^2. 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 applied random walk MH sampler uses a normal proposal distribution for the sampling of β_{points} with a mean equal to the maximum likelihood estimate of β_{points} and its corresponding variance. The tuning parameter is set to 0.005. Over the pooled chains both sampling procedures are run 120.000 times with a burn in period of 30.000. When applying MCMC sampling it is needed to specify priors, which refer to as the previous knowledge about the effect. For now I use uninformative priors for the regression coefficients with a mean of μ_{0j} = 0 and a variance of $\sigma^2{}_{0j}$ = 1000. As the outcome variable is supposed to be normally distributed using normal priors for the regression coefficients yields a normal posterior distribution, which refers to as conjugate priors. The prior distribution for the variance is the inverse of a Gamma distribution with shape and rate parameters α_0 = .0001 and β_0 = .0001. This is also a conjugate prior and leads to a posterior Inverse-Gamma distribution as well. MCMC samplers also demand the use of initial values, which were randomly sampled from a uniform distribution for each parameter. To assure that the initial values don't influence the results multiple chains need to specified, which each sample all estimates an equal amount of iterations with differing initial values. For this project I 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, autocorrelations as well as the acceptance ratios are used. Figure 1 depicts the trace plots of all 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. Figure 2 depicts the exemplary autocorrelations for β_0, β_{points} 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 very promising as it indicates a fitting tuning parameter. Regarding the results of the autocorrelations the number of iterations was adjusted to be 40.000 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 (see Table 1). Overall, all indicators do not raise any concerns regarding non-convergence.

Figure 1: Trace plots for all obtained estimates

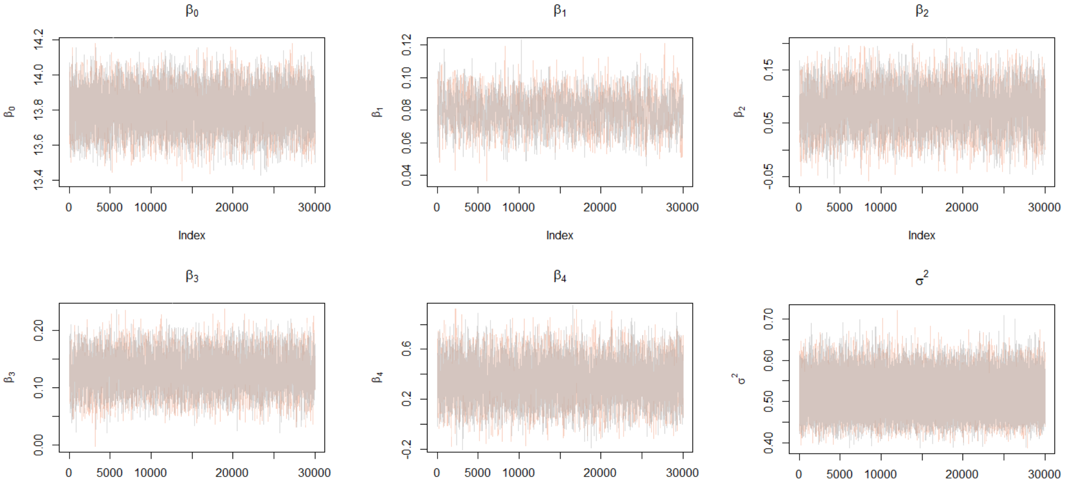
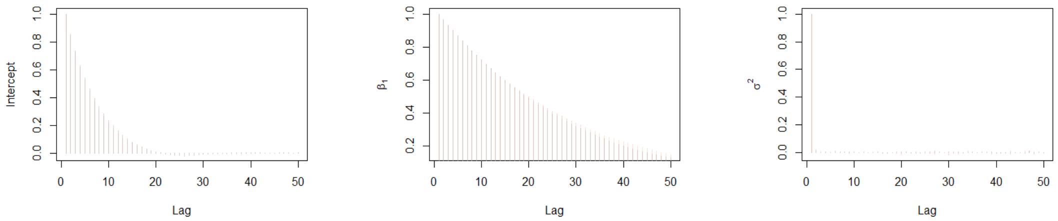


Figure 2: Autocorrelation plots with



3.2 Posterior predictive check

To check whether the model was in line with two major assumptions of linear regression I decided to check for homoscedasticity and linearity of the residuals with two posterior predictive checks. Homoscedasticity assumes that the variance of the outcome variable is independent of the predictors and 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 values. This correlation is the first discrepancy measure. Secondary, linearity holds if the residuals have no specific pattern and if they have a mean close to zero over the whole distribution of the residuals. To see how the residuals behaved over the whole distribution, I separated it in three equal parts and computed each mean. The sum of these three means is then 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 observed data set, can be seen as a measure of evidence against the null hypothesis of homoscedasticity (linearity). 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 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 the observed data has a slightly lower sum of the means than the simulated data sets, which also indicates no violation of the linearity. Therefore, no violations at all are expected.

3.3 Parameter Estimates

As there is no indication of violation of convergence or predictive checks, the parameter estimates can be interpreted. The MCMC sampling procedure included 90,000 iterations over the pooled chains and a combined burn-in period of 30,000. As can be seen in Table 1, based on the estimates and the 95% credible intervals the points, assists, defensive rebounds and steals are related to the log(salary) as none of them included zero in the credible intervals. The posterior mean of the intercept is 13.802 (95%-CCI = [13.612; 13.993]). Every additional point 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 is related to a 0.076 (95%-CCI = [0.011; 0.141]) increase in the outcome variable. Next, every additional defensive rebound is related to a 0.127 (95%-CCI = [0.074; 0.181]) increase in the outcome variable. Last, every additional steal 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]).

Table 1: Parameter estimates for the four predictor model

	Mean	S.E.	MC error	2.5%	Median	97.5%	Accepted	Burn-in	Iterations
Intercept	13.80	0.10	0.00	13.61	13.80	13.99	1.00	30,000	90,000
Points per game	0.08	0.01	0.00	0.06	0.08	0.10	0.57	30,000	90,000
Assists per game	0.08	0.03	0.00	0.01	0.08	0.14	1.00	30,000	90,000
Rebounds per game	0.13	0.03	0.00	0.07	0.13	0.18	1.00	30,000	90,000
Steals per game	0.35	0.14	0.00	0.07	0.35	0.64	1.00	30,000	90,000
Variance	0.52	0.04	0.00	0.44	0.51	0.60	1.00	30,000	90,000

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

3.4 Model comparison

Of secondary interest in this report is which combination and what number of predictors yields the best model fit. To evaluate model fit three competing models are proposed. The three models are, a model with only two offensive statistics, a model with only two defensive statistics and, a model with a combination of one offensive and 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 but 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 model (two offensive) has a DIC of 822.64, the second model (two defensive) has a DIC of 899.71 and the third alternative model including the combination of predictors (points & rebounds) 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 competing models. This suggests, that offensive statistics alone do not yield the best model fit, but the combination of multiple offensive and defensive statistics gets valued the most.

Of further interest is, that the model with only defensive statistics has the worst model fit, but the model including a combination of offensive and defensive statistics fits better than the model with only offensive statistics.

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:

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 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 the 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 the 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 informative hypothesis of assists having a stronger effect than defensive rebounds 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 partly supports the hypothesis that the effect of assists is higher than the effect of rebounds as the BF is lower than expected. Furthermore, the posterior probability including the fail-safe hypothesis is 0.62, showing also a high relative uncertainty about the hypothesis. As the trend still supports H_3 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 was rather easy to find historical data as the game statistics and the salary information of previous years is freely accessible. 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 used the statistical point estimates and quantified uncertainties of the same four predictors on the same outcome variable, 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 2. When comparing the estimates of Table 2 to Table 1 it can be seen that the credible intervals for all estimates have a smaller range, and the mean estimates slightly change as well. Especially, in case of β_{steals} the estimate changes from β_{steals}=0.352 to β_{steals}=0.402. Therefore, using prior knowledge seemed to help to obtain more credible and more accurate results. Additionally, the DIC of the model including the informative priors is slightly lower with DIC=795.84 compared to DIC=796.41 in the original model. On top, using informative priors shows an intuitive approach to accumulate scientific evidence, which is harder in frequentist procedures like meta analyses or cross validation.

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

	Mean	S.E.	MC error	2.5%	Median	97.5%	Accepted	Burn-in	Iterations
Intercept	13.75	0.09	0.00	13.57	13.75	13.93	1.00	30,000	90,000
Points per game	0.08	0.01	0.00	0.06	0.08	0.09	0.56	30,000	90,000
Assists per game	0.07	0.03	0.00	0.01	0.07	0.14	1.00	30,000	90,000
Rebounds per game	0.14	0.03	0.00	0.09	0.14	0.19	1.00	30,000	90,000
Steals per game	0.40	0.14	0.00	0.14	0.40	0.67	1.00	30,000	90,000
Variance	0.52	0.04	0.00	0.44	0.51	0.60	1.00	30,000	90,000

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

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 assume that being the best in all offensive efforts, but having only a small defensive effort gets 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, which I am mainly interested in. Additional 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 do not only focus on offense.

Regarding the performed analyses, using a frequentist approach would have yielded a similar conclusion, but it simply would lack flexibility. The flexibility of the Bayesian approach can be shown by means of the incorporated historical data, the use of the Bayes factor to evaluate informative hypothesis, 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.

5. References

Page, B., & NBA, 2. (2022). 2020-21 NBA Player Stats: Totals | Basketball-Reference.com. Basketball-Reference.com. Retrieved 9 June 2022, from https://www.basketball-reference.com/leagues/NBA_2021_totals.html.
Oliver, D., Daly, F., Martin, F. C., & McMurdo, M. E. (2004). Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review. Age and ageing, 33(2), 122-130.