Ola Runeson Rasmussen

Analyzing the Impact of COVID-19 on the Total Points Scored in NBA Basketball Matches:

A Comparative Study of Pre and Post-Pandemic Seasons

Bachelor's thesis in Mathematical Sciences
Supervisor: Jarle Tufto
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU

ABSTRACT

In this thesis I will analyze the impact that the Covid-19 pandemic had on the number of points scored in a NBA basketball match. I will find the model that best fits the data, and analyze the home court advantage and find out if the home court advantage changed during the Covid-19-season. I will also analyze the impact the number of fans present at a match had on the total amount of points obtained by both the home team and opposing team. As a bonus, I will find the strengths of each team throughout the years.

I denne oppgaven skal jeg analysere hvilken innvirkning Covid-19-pandemien hadde på antall poeng scoret i en NBA-basketballkamp. Jeg skal finne den modellen som passer best til dataene, og analysere hjemmebanefordelen og finne ut om hjemmebanefordelen endret seg i løpet av Covid-19-sesongen. Jeg vil også analysere hvilken innvirkning antallet tilstedeværende fans på en kamp hadde på det totale antallet poeng oppnådd av både hjemmelaget og motstanderlaget. Som en bonus vil jeg finne styrken til hvert lag gjennom årene.

PREFACE

I would like to thank my supervisor and my family for helping me though this process.

CONTENTS

Abstract **i**

Preface **ii**

Contents **iv**

1 Introduction **1**
1.1 Modelling Sports Results 1
1.2 NBA 2
1.3 Home Court Advantage 2

2 Theory **3**
2.1 Generalized linear Models 3
2.2 Generalized linear Mixed models 4
2.3 Poisson process 4
2.4 AIC 5
2.5 Likelihood Ratio Test (LRT) 6
2.6 AR(1)-process 6
2.7 OU-process 6

3 Method **7**
3.1 Collecting Data 7
3.2 Data management 7
3.3 Model Selection Method 8

4 Results and Discussion **9**
4.1 Model Selection 9
4.2 Best Model 10
4.3 Correlation between attack and defence strength 11
4.4 Strength of each team 12
4.5 Covid-19 impact and home court advantage 15
4.6 Number of fans 15
4.7 Extension of the best model 16
4.8 Future work 19

5 Conclusions **21**

References **23**

Appendices: **25**

A - Github repository **26**

CHAPTER

ONE

INTRODUCTION

1.1 Modelling Sports Results

Modelling sport results can have many different use cases. It can be used to predict future matches, player performance among other things. Coaches and analysts can use data to analyze strengths and weaknesses of players and opponents, and scouts can use models to find what players could have a potential future in that sport. The main objective of this thesis is to find a model that fit 8 seasons worth of data from the National Basketball Association (NBA), starting with the 2014-2015 season, skipping the 2019-2020 season, and ending with the 2022-2023 season. The reason why the 2019-2020 season is skipped is because this season got cut short because of the covid-19-pandemic. Something similar has been done in [1], were they modelled the scores in football matches to find the attack strength and defence strength of teams, but they also found the home-field advantage that a team has. In this thesis I will also find the home-court advantage but also find the differences between the 2020-2021 season and the other seasons. Due to covid restrictions, this season was played without any fans, so it will be interesting to see if the home-court advantage disappeared or got weaker. Further differences between what they have done and what I have done in this thesis will be looked at in Section 4.2.

1.2 NBA

The NBA, National Basketball Association, is the biggest basketball league in North-America. A regular season in the NBA, consists of 82 games for each team. There are 30 teams in the NBA, divided into 2 conferences, the eastern- and the western conference. Teams in the same conference plays 4 games against each other and 2 games against teams in the other conference. The 2020 to 2021 season, reffered to as the Covid season, only consisted of 72 games for each team. Also, under this season there where no fans allowed because of Covid-19 regulations.

There are 3 different scoring methods, or types, in a normal NBA game. You have the one-pointers, scoring method one, which are scored from free throws. Free throws are a form of penalties obtained when the other team commits a foul. Two-pointers, scoring method two, are obtained from getting the basketball into the opponents hoop while inside the three point line, while three-pointers, scoring method three, are obtained from scoring outside the three point line. Also, a team is awarded 2 free throws when fouled while attempting a two pointer, and 3 free throws when fouled while attempting a three pointer. But they are only awarded 1 free throw if they scored while being fouled.

1.3 Home Court Advantage

In [2], the author analyzed home court advantages for NCAA basketball statistics. NCAA, National Collegiate Athletic Association, is the college equivalent of the NBA. He came to the conclusion that when playing on their home court, a team received a boost in nearly all statistical categories. In this thesis, I will come to the same conclusion that teams scored more points while playing on their home court.

CHAPTER

TWO

THEORY

2.1 Generalized linear Models

All of this theory is collected from [3] if not stated otherwise. In a Generalized linear Model (GLM), where we have $i = 1, 2, \ldots, n$ observations, we want to find a linear relationship between the covariate vectors, the $\mathbf{x}_i$'s, and a transformation of the expected value, mean μ_i, of a distribution. That is,

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}. \tag{2.1}$$

where $g(\mu_i)$, the link function, is the function that transforms this into a linear relationship, and $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ is called the linear predictor. There are many different link functions for different distributions. One thing these distributions have in common is that they are a part of the Exponential Family of distributions. The probability mass function of a multivariate exponential family for the response variable y_i is defined by,

$$f(y_i|\theta_i) = exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i) \right\}, \tag{2.2}$$

where the parameter θ_i is called the natural or canonical parameter, the parameter ϕ is the dispersion parameter, and w_i, usually a weight, is a known value. For the function $b(\theta_i)$, it is required that $f(y_i|\theta_i)$ can be normalized and that $b'(\theta_i)$ and $b''(\theta_i)$ exist. The expected value and variance of this probability mass function is then defined by,

$$E[y_i] = \mu_i = b'(\theta_i), \quad Var[y_i] = \sigma_i^2 = \phi \frac{b''(\theta_i)}{w_i}.$$

2.2 Generalized linear Mixed models

Generalized linear Mixed models (GLMMs) are defined by adding a random effect $\boldsymbol{\gamma}_i$ to the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. The responses y_{ij}, where $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, with n_i being the measurements per individual or cluster. When adding a cluster specific random effect $\boldsymbol{\gamma}_i$ to the responses y_{ij}, the conditional mean $\mu_{ij} = E[y_{ij}|\gamma_i]$ is linked to the linear predictor,

$$\eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{u}_{ij}^\top \boldsymbol{\gamma}_i, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i, \tag{2.3}$$

through the link function $g(\mu_{ij}) = \eta_{ij}$. Here the random effects γ_i are independent and identically distributed $N(0, Q)$ where Q is the covariance matrix for the random effects [3] p. 391. To estimate these fixed and random effects, I will use the "glmmTMB" package in R. Simply said, this package will numerically approximate the estimated values.

2.3 Poisson process

One distribution that is a part of the exponential family of distributions, is the Poisson distribution. It is defined by the density,

$$f(y_i|\lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad y_i = 0, 1, \ldots, \tag{2.4}$$

where y_i is the total amount of times an event occurs in an interval, and $\lambda_i > 0$ is the rate at which the occurrences happen. In this thesis I will model the scores in a basketball match via a Poisson process. The mean and variance of the Poisson distribution is the same. That is,

$$E[y_i] = \mu_i = \lambda_i, \quad Var[y_i] = \lambda_i.$$

The natural link function to use for the Poisson distribution is the log-link function, which is given by,

$$g(\lambda_i) = \log(\lambda_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}. \tag{2.5}$$

This assures us that the rate $\lambda_i > 0$ for all $\beta \in \Re^p$. The Mixed Poisson model with the log-link function is called the mixed log-linear Poisson model. It is defined as seen below where $y_{ij}|\boldsymbol{\gamma}_i \sim \text{Poisson}(\lambda_{ij})$, where

$$\log(\lambda_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{u}_{ij}^\top \boldsymbol{\gamma}_i. \tag{2.6}$$

In this thesis, the random effects I will use is attack strength $\gamma_{A,i,k}$ and defence strength $\gamma_{D,i,k}$. The notation,

$$\log(\lambda_{ijk}) = \text{"fixed effects"}_k + \gamma_{A,i,k} + \gamma_{D,j,k}, \tag{2.7}$$

will be used where i is the attacking team, j is the defending team, and k is the scoring type.

A Poisson process is a continous time process, where events happen independently of one another with a certain rate (intensity) λ. The probability of n scores in a match is then,

$$P\big((\#\text{ scores of type } k) = n\big) = \frac{\lambda^n}{n!}e^{-\lambda}. \tag{2.8}$$

In [4], the authors comes to the conclusion that the scores in a basketball match can mostly be seen as a Poisson process, but in close games, the Poisson process assumption breaks. This will be discussed further in Section 4.8. In this thesis, I will assume that the scores follow a Poisson process.

2.4 AIC

Akaike Information Criterion (AIC) is an estimator for the quality of our model given our data. It penalizes complex models by taking the amount of estimated parameters into account. The formula for AIC is,

$$\text{AIC} = 2k - 2ln(\hat{L}), \tag{2.9}$$

where k is the number of estimated parameters and $\hat{L}$ is the maximized value of the likelihood function of our model. The AIC is an estimator for the Kullback Leibler distance between the actual and "true" model. The data will generate an unknown distribution, so the difference between the poisson distribution used and this unknown distribution is then the Kullback Leiber distance, showing how different these two distributions are [5]. I will use this criteria to find the best model for the Covid season and the other seasons.

2.5 Likelihood Ratio Test (LRT)

Together with the AIC of a model, the likelihood ratio Test is also used to find significant variables in a model. When testing the hypothesis $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, we can use the (log-)likelihood ratio statistic,

$$\text{lr} = 2\big\{l(\hat{\beta}) - l(\tilde{\beta})\big\} = -2\big\{l(\tilde{\beta}) - l(\hat{\beta})\big\}, \tag{2.10}$$

where $l(\tilde{\beta})$ is the log-likelihood for the restricted model under H_0 and $l(\hat{\beta})$ is the log-likelihood of the full model [3] p. 662-663

2.6 AR(1)-process

An AR(1), auto-regressive, process is a stochastic random process. The AR(1) process $\{X_t\}_{t\in Z}$ is defined as a causal stationary series satisfying the equation,

$$X_t = \varphi X_{t-1} + \epsilon_t, \quad t = 1, 2, \ldots, \tag{2.11}$$

where $|\varphi| < 1$ and ϵ_t is the white noise with zero mean and variance σ_ϵ^2. They are identically and independently distributed [6] p. 15.

2.7 OU-process

A stationary Gaussian Ornstein-Uhlenbeck process, Gaussian continuous-time AR(1) process, is considered as the continuous-time analogue of the discrete-time AR(1) process, [6] p.343. It can have irregular timepoints instead of the constant even time differences in the AR(1) process. The OU-process $\{U_t\}_{t\in R}$ is defined by the stochastic differential equation,

$$dU_t = -\theta U_t dt + \sigma dW_t, \tag{2.12}$$

where $\theta > 0$, $\sigma > 0$ are parameters and $U_t \sim N\big(u_o e^{-\theta t}, \frac{\sigma^2}{2\theta}\big\{1 - e^{-2\theta t}\big\}\big)$ [7]. W_t is the Wiener process, which is a two-sided Brownian motion [8]. An important property of 2.12 is that the autocovariancefunction decreases exponentially.

CHAPTER

THREE

METHOD

3.1 Collecting Data

All of the data used in this thesis is collected from the official NBA website and Basketball Reference seen below:

- `https://www.nba.com/stats/teams/boxscores-traditional`
- `https://www.basketball-reference.com/`.

3.2 Data management

All the data was changed into the format seen in Table 3.1. Attacking team i is the team attacking defending team j, y is the number of scores obtained, type is the scoring type, i.e. when type 3 means three pointer and so on. Home court is "yes" if attacking team i is playing on their home court. Date shows the number of days after the first game saved in the data. Win shows what team won the match. Covid indicate whether we are looking at the Covid season or a normal season, and Number of fans says how many fans were present.

Attacking Team i	Defending Team j	y	Type	Home Court	Date	Win	Covid	Number of Fans
LAL	HOU	3	3	yes	1	no	no	18997
LAL	HOU	25	2	yes	1	no	no	18997
LAL	HOU	31	1	yes	1	no	no	18997
HOU	LAL	12	3	no	1	yes	no	18997
HOU	LAL	19	2	no	1	yes	no	18997
HOU	LAL	34	1	no	1	yes	no	18997

Table 3.1: Six rows of the data frame containing data for one match between the Los Angeles Lakers and the Houston Rockets

3.3 Model Selection Method

To find the best model, I used the covariates Type, Home, and Covid. Type is the scoring type, Home is the home advantage, and Covid indicate the covid season. I made many different combinations of these covariates with different interactions and compared them using their AIC value. If two models had the same AIC, i used the Likelihood Ratio Test to find significant terms. These models can be seen in Table 4.1.

CHAPTER

FOUR

RESULTS AND DISCUSSION

4.1 Model Selection

Models:	Change from the best model:	ΔAIC from the best model
Mod1	$-\beta_{\mathrm{HC}}\ x_{\mathrm{HC,ijk}} + \beta_{\mathrm{HC,k}}\ x_{\mathrm{HC,ijk}}$	0
Mod2	$-\beta_{\mathrm{HC}}\ x_{\mathrm{HC,ijk}}$	3
Mod3	$-\beta_{\mathrm{C,k}}\ x_{\mathrm{C,ijk}} + \beta_{\mathrm{C}}\ x_{\mathrm{C,ijk}}$	958
Mod4	$-\beta_{\mathrm{H,k}}\ x_{\mathrm{H,ijk}} + \beta_{\mathrm{H}}\ x_{\mathrm{H,ijk}}$	14
Mod5	$-\beta_{\mathrm{C,k}}\ x_{\mathrm{C,ijk}} - \beta_{\mathrm{HC}}\ x_{\mathrm{HC,ijk}} + \beta_{\mathrm{C}}\ x_{\mathrm{C,ijk}}$	961
Mod6	$-\beta_{\mathrm{H,k}}\ x_{\mathrm{H,ijk}} - \beta_{\mathrm{C,k}}\ x_{\mathrm{C,ijk}} + \beta_{\mathrm{H}}\ x_{\mathrm{H,ijk}} + \beta_{\mathrm{C}}\ x_{\mathrm{C,ijk}}$	972
Mod7	$-\beta_{\mathrm{H,k}}\ x_{\mathrm{H,ijk}} - \beta_{\mathrm{HC}}\ x_{\mathrm{HC,ijk}} + \beta_{\mathrm{H}}\ x_{\mathrm{H,ijk}}$	16
Mod8	$-\beta_{\mathrm{H,k}}\ x_{\mathrm{H,ijk}} - \beta_{\mathrm{C,k}}\ x_{\mathrm{C,ijk}} - \beta_{\mathrm{HC}}\ x_{\mathrm{HC,ijk}} + \beta_{\mathrm{H}}\ x_{\mathrm{H,ijk}} + \beta_{\mathrm{C}}\ x_{\mathrm{C,ijk}}$	974

Table 4.1: Change in AIC from the best model

The best model can be seen in Equation 4.1, but below I will show some neighboring models to the best one that I compared it to in Table 4.1 along with the difference between the AIC for the best model and these neighboring models. In model 1, the interaction between the home court advantage and covid is replaced with a three way interaction between home court advantage, covid and each scoring type. In model 2, the interaction between the home court advantage and covid is removed. In model 3, the interaction between covid and each scoring type is replaced with a single covid term. In model 4, the interaction between home court advantage and each scoring type is replaced with a single home court term. Model 5 is almost the same as model 3 but the interaction between home court and covid is removed. Model 6 has both the interaction between home court and scoring type and the interaction between covid and scoring type replaced with just one

term for home court and one term for covid. Model 7 is similar to model 4, but the interaction between home court and covid is removed. Finally, model 8 is the same as model 6 but the interaction between home court and covid is removed.

4.2 Best Model

After a great deal of testing, I have found that the best model to use for this data is,

$$\log(\lambda_{i,j,\mathrm{k}}) = \beta_{\mathrm{k}} + \beta_{\mathrm{H,k}}\, x_{\mathrm{H,ijk}} + \beta_{\mathrm{C,k}}\, x_{\mathrm{C,ijk}} + \beta_{\mathrm{HC}}\, x_{\mathrm{HC,ijk}} + \gamma_{\mathrm{A},i,\mathrm{k}} + \gamma_{\mathrm{D},j,\mathrm{k}}. \tag{4.1}$$

The scoring type is denoted by index $\mathrm{k} = 1, 2, 3$, the attacking team is denoted by index i, and the defending team is denoted by index j. $x_{\mathrm{H,ijk}}$, $x_{\mathrm{C,ijk}}$, and $x_{\mathrm{HC,ijk}}$ are dummy, or indicator, variables. They indicate if we have a match where attacking team i has home court advantage, if we have a match played during the Covid-season, or if we have a match with home court advantage during the Covid-season, respectively. Table 4.2 shows the fixed effects of the model in Equation 4.1 together with their standard deviations, and Table 4.3 shows the percent the expected number of scores changes for each type when the different estimated values are included. This model is very similar to the one used in [1], but I have added a covid term and an interaction between home court and covid, and since there are three ways of scoring in basketball, i have added an interaction between each scoring types and home court term and the covid term.

The random effects in the model are $\gamma_{\mathrm{A},i,\mathrm{k}}$ and $\gamma_{\mathrm{D},j,\mathrm{k}}$. They are independent and identically distributed, with $\gamma_{\mathrm{A},i,\mathrm{k}} \sim N(0, \tau^2_{\mathrm{A,k}})$ and $\gamma_{\mathrm{D},i,\mathrm{k}} \sim N(0, \tau^2_{\mathrm{D,k}})$. These variances can be seen in Table 4.4. We assume that attack strength and defence strength of team i are independent of one another, i.e. $\gamma_{\mathrm{A},i,\mathrm{k}}$ and $\gamma_{\mathrm{D},i,\mathrm{k}}$ are independent for each scoring type, but I will discuss this assumption in Section 4.3. These random effects together with the fixed effects gives the conditional expected number of scores from scoring type k that attacking team i will score against defending team j by the equation,

$$E\big[y_{i,j,\mathrm{k}} | \gamma_{\mathrm{A},i,\mathrm{k}}, \gamma_{\mathrm{D},j,\mathrm{k}}\big] = \exp\big\{\beta_{\mathrm{k}} + \beta_{\mathrm{H,k}}\, x_{\mathrm{H,ijk}} + \beta_{\mathrm{C,k}}\, x_{\mathrm{C,ijk}} + \beta_{\mathrm{HC}}\, x_{\mathrm{HC,ijk}} + \gamma_{\mathrm{A},i,\mathrm{k}} + \gamma_{\mathrm{D},j,\mathrm{k}}\big\}. \tag{4.2}$$

k	$\hat{\beta}_{\mathrm{k}} \pm SD(\hat{\beta}_{\mathrm{k}})$	$\hat{\beta}_{\mathrm{H,k}} \pm SD(\hat{\beta}_{\mathrm{H,k}})$	$\hat{\beta}_{\mathrm{C,k}} \pm SD(\hat{\beta}_{\mathrm{C,k}})$	$\hat{\beta}_{\mathrm{HC}} \pm SD(\hat{\beta}_{\mathrm{HC}})$
1	2.8412 ± 0.0118	0.0343 ± 0.0035	-0.0229 ± 0.0063	-0.0124 ± 0.0060
2	3.3695 ± 0.0099	0.0166 ± 0.0027	-0.0226 ± 0.0052	-0.0124 ± 0.0060
3	2.3207 ± 0.0186	0.0277 ± 0.0045	$+0.2079 \pm 0.0072$	-0.0124 ± 0.0060

Table 4.2: Fixed Effects ± Standard Deviations

k	$e^{\hat{\beta}_{\mathrm{k}}}$	$e^{\hat{\beta}_{\mathrm{H,k}}}$	$e^{\hat{\beta}_{\mathrm{C,k}}}$	$e^{\hat{\beta}_{\mathrm{HC}}}$
1	≈ 17	$\approx 3.49\%$	$\approx -2.26\%$	$\approx -1.23\%$
2	≈ 29	$\approx 1.68\%$	$\approx -2.23\%$	$\approx -1.23\%$
3	≈ 10	$\approx 2.81\%$	$\approx +23\%$	$\approx -1.23\%$

Table 4.3: Exponential value of the Fixed Effects

4.3 Correlation between attack and defence strength

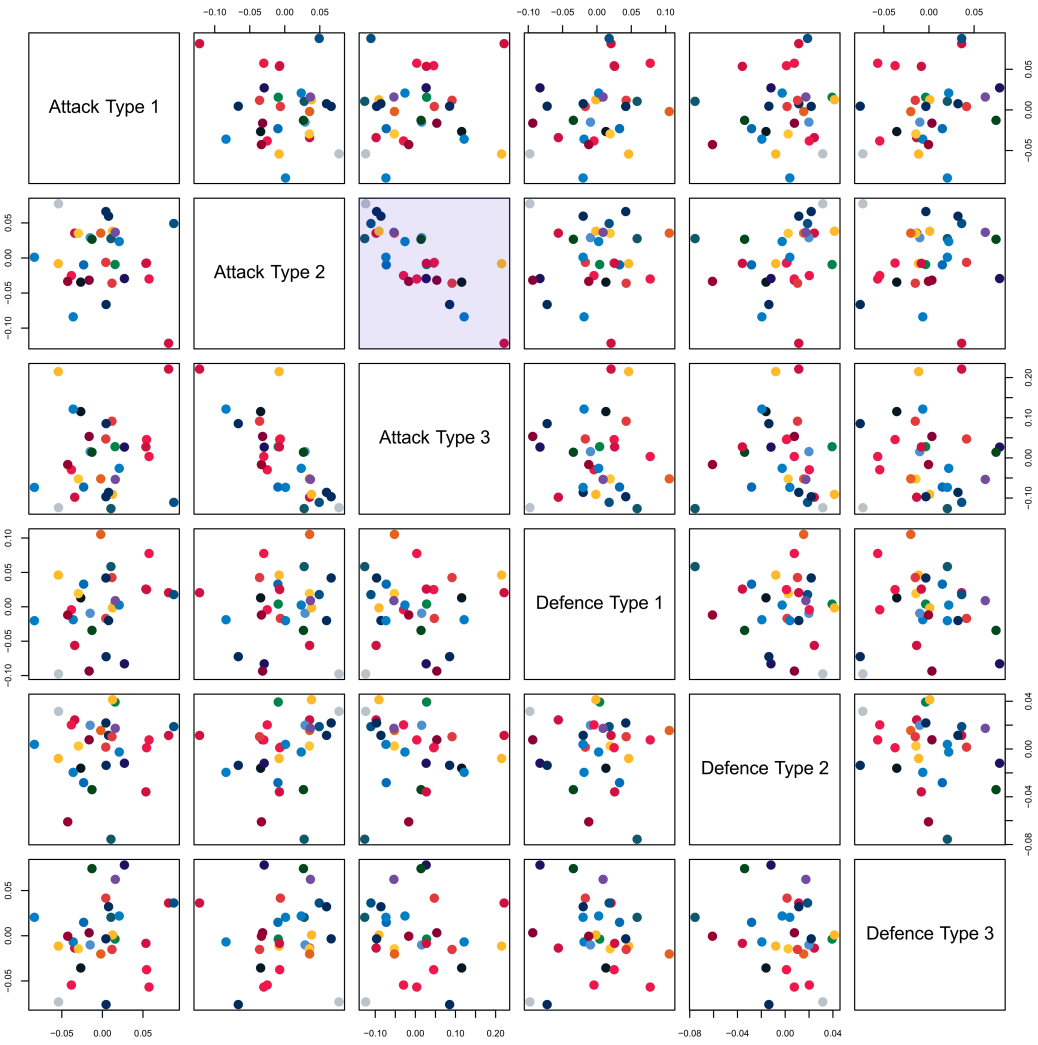

Figure 4.1: Correlation between the different attacking and defending strengths

The correlation between attacking and defending strength for each scoring type can be seen in Figure 4.1. From this figure we can see that there is a negative correlation between attacking strength of type 2 and 3 in the cell in row 2, column 3. This entails that teams who are very good at scoring with type 3 are bad at scoring or unwilling to score with type 2, and vice versa. The only team not following this norm is the Golden State Warriors, GSW. This makes sense because they have won 4 of the last 8 seasons excluding the 2019-2020 season, which means that they have been consistently good during these 8 seasons. There doesn't seem to be any correlation between attack and defence strengths, so the assumption that attack strength and defence strength are independent seem to mostly hold up looking at the correlation table.

4.4 Strength of each team

The attacking strengths for each team can be seen in Figure 4.2, 4.3, and 4.4, and the defending strengths can be seen in Figure 4.5, 4.6, and 4.7. Here we see which teams are the strongest at each scoring type. The expected values of the strength parameters are zero, so when a team has a strength of 0.10, it is $e^{0.10} \approx 11\%$ stronger that the mean strength. This means that they will score 11% more scores than a team with average strength.

k	$\tau^2_{A,k}$	$\tau^2_{D,k}$
1	0.0016857	0.0023295
2	0.0020435	0.0007715
3	0.0084381	0.0016362

Table 4.4: Variances of the Random Effects

The variances for the random effects can be seen in Table 4.4. From this table we can see that the three points attacking strength vary much more than the other attacking types and all defending types. This makes sense because the three point shot is the hardest scoring type, so all teams cannot be equally as good with that type. Also the two-pointers don't vary as much, which means that all teams score almost the same amount of two-pointers.

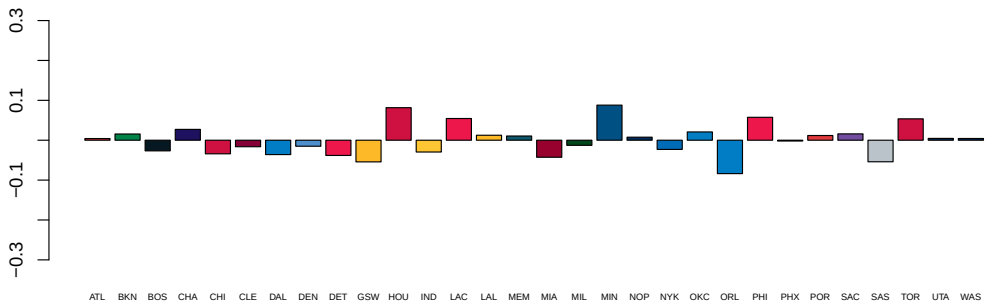

Figure 4.2: Strength of attacking type 1.

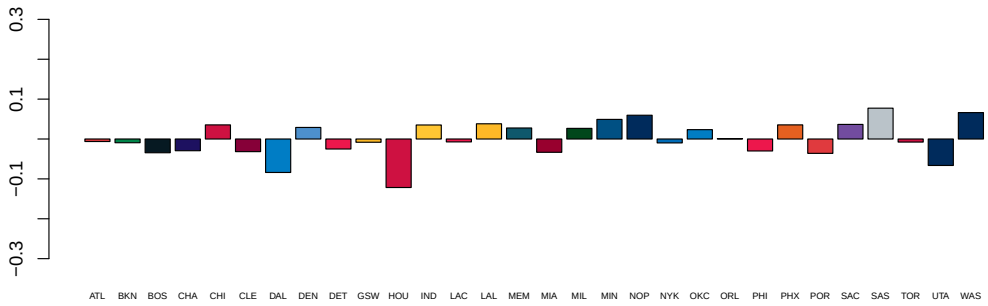

Figure 4.3: Strength of attacking type 2.

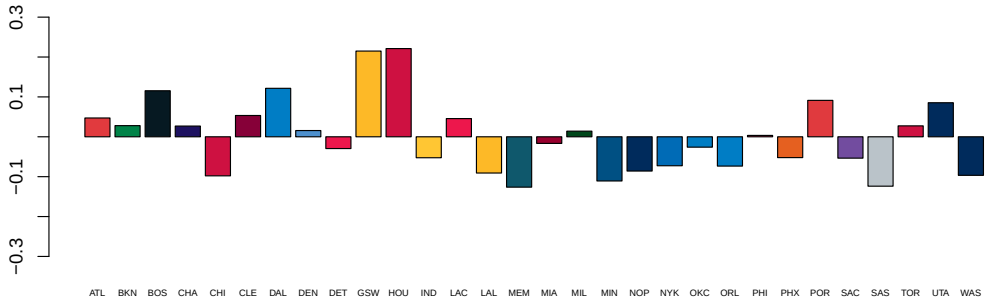

Figure 4.4: Strength of attacking type 3.

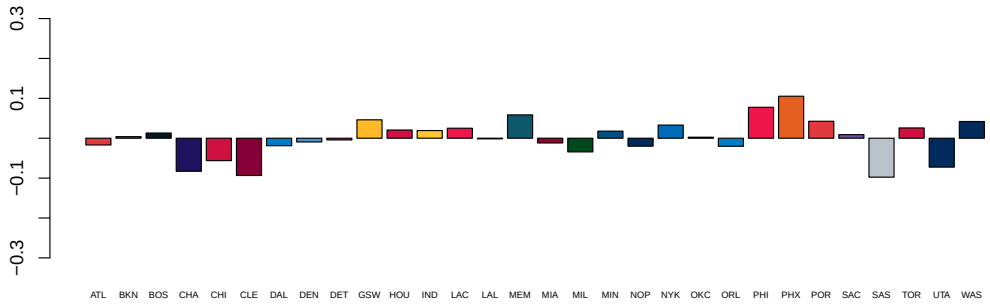

Figure 4.5: Strength of defending type 1.

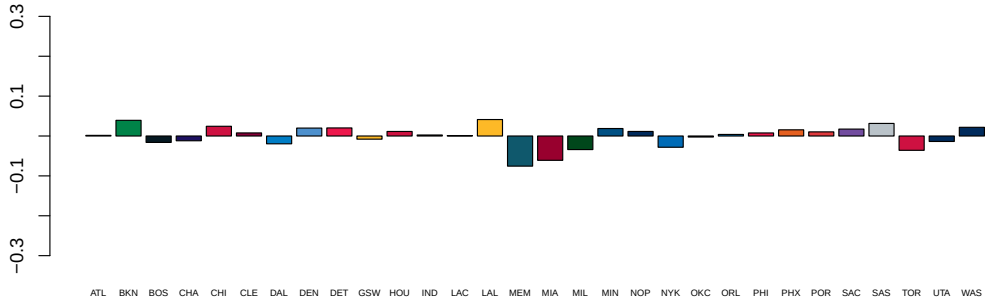

Figure 4.6: Strength of defending type 2.

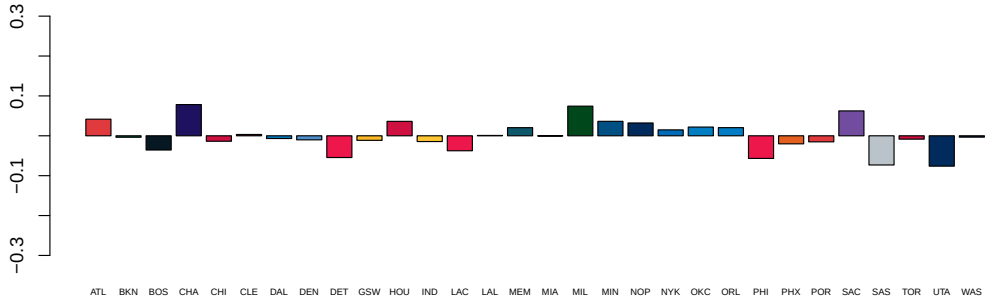

Figure 4.7: Strength of defending type 3.

4.5 Covid-19 impact and home court advantage

From Table 4.3 we can see the influence that the home court had and that Covid had on every scoring type. In the last column, we can also see that during covid, the homecourt advantage got 1.23% smaller during covid. There could be many reasons why the homecourt advantage got smaller, but one explanation of this could be that there were no fans spectating, and with less fans the home team didn't get the same boost that a full arena would give them.

4.6 Number of fans

Another interesting influence on the number of scores could be the number of fans spectating. The best model for this is,

$$\begin{aligned}\log(\lambda_{i,j,\mathrm{k}}) = \beta_{\mathrm{k}} + \beta_{\mathrm{H,k}}\, x_{\mathrm{H,ijk}} + \beta_{\mathrm{FANS,k}}\, x_{\mathrm{FANS,ijk}} \\ + \beta_{\mathrm{H.FANS}}\, x_{\mathrm{H.FANS,ijk}} + \gamma_{\mathrm{A},i,\mathrm{k}} + \gamma_{\mathrm{D},j,\mathrm{k}},\end{aligned} \tag{4.3}$$

which I found using the same method for testing as in Section 4.1. This model is almost identical to the model in Equation 4.1 but instead of Covid we are using the number of fans present at a match. x_{FANS} and $x_{\mathrm{H.FANS}}$ is then the number of fans present in a normal match and a normal match at home respectively. When the number of fans at a match is zero, this model is essentially the same as the model in Equation 4.1 during covid. Table 4.5 show the influence that a single fan has and the influence 20.000 fans has on the total amount of scores of each type. The last column show the influence added when having home court advantage. This table also show the amount of points during covid, because there were no fans present. It suggest that when the number of fans increased, the total amount of one-pointers and two-pointers increased but the total amount of three-pointers decreased. This is because during covid, the total amount of three-pointers scored in a match increased, as shown in Table 4.3.

k	$e^{\hat{\beta}_{\mathrm{k}}}$	$e^{\hat{\beta}_{\mathrm{FANS,k}}}$	$e^{\hat{\beta}_{\mathrm{FANS,k}}\cdot 20000}$	$\beta_{\mathrm{H.FANS}}$
1	≈ 17	$\approx 0.0001\%$	$\approx 2.1\%$	$\approx 0.0001\%$
2	≈ 28	$\approx 0.0001\%$	$\approx 2.5\%$	$\approx 0.0001\%$
3	≈ 13	$\approx -0.0012\%$	$\approx -20.7\%$	$\approx 0.0001\%$

Table 4.5: Exponential value of the Fixed Effects using the model with fans

4.7 Extension of the best model

This thesis has only used the strength of a team according to 8 seasons worth of data, but because some teams change a lot inbetween and during seasons, a more interesting approach could be to use the AR(1) process to find the strengths of each team where they changes during the seasons. This could tell us if some teams tend to be stronger at the start, middle or end of a season, and how their strengths have changed throughout the years. But since there are some breaks in a season, one could maybe use the OU process to find the strengths during a season. The OU process can account for irregularities of the times where values are recorded, and can therefore be more useful. This new OU model has the same formula as the best model in section 4.2, but the random effects are now modelled as a OU process. Tables 4.6 and 4.7 show the fixed effects and the exponential of the fixed effects respectively. These tables are pretty similar to tables 4.2 and 4.3, but the covid term now give a positive impact on one- and two-poiners. This may be because now the attacking strengths of each team are not constant, but changes over time, so their strengths during covid could be slightly lower to account for this. This can be seen in Figures 4.8, 4.9, and 4.10 where the highlighted areas are the covid season. From Figure 4.10 we can see that the attacking strengths have increased a lot throughout the years for every team. Figures 4.11, 4.12, 4.13, and 4.14 show the attacking strengths of the Golden State Warriors and Houston Rockets for both two-pointers and three-pointers. These two teams had the best three-points strengths, so it will be interesting to see how they compare throughout the years. We see that GSW has been very consistent with their strengths except for their two-pointer strength in the later seasons, but HOU have been very bad with two-pointers and inconsistent with three-pointers. This is due to them losing the players who scored three-pointers for them and GSW has kept most of theirs.

k	$\hat{\beta}_{\mathrm{k}} \pm SD(\hat{\beta}_{\mathrm{k}})$	$\hat{\beta}_{\mathrm{H,k}} \pm SD(\hat{\beta}_{\mathrm{H,k}})$	$\hat{\beta}_{\mathrm{C,k}} \pm SD(\hat{\beta}_{\mathrm{C,k}})$	$\hat{\beta}_{\mathrm{HC}} \pm SD(\hat{\beta}_{\mathrm{HC}})$
1	2.8298 ± 0.0148	0.0339 ± 0.0035	0.0387 ± 0.0230	-0.0126 ± 0.0061
2	3.3645 ± 0.0147	0.0167 ± 0.0028	0.0158 ± 0.0224	-0.0126 ± 0.0061
3	2.3134 ± 0.0150	0.0278 ± 0.0045	0.0926 ± 0.0240	-0.0126 ± 0.0061

Table 4.6: Fixed Effects ± Standard Deviations for the OU model

k	$e^{\hat{\beta}_{\mathrm{k}}}$	$e^{\hat{\beta}_{\mathrm{H,k}}}$	$e^{\hat{\beta}_{\mathrm{C,k}}}$	$e^{\hat{\beta}_{\mathrm{HC}}}$
1	≈ 17	$\approx 3.45\%$	$\approx 3.95\%$	$\approx -1.25\%$
2	≈ 29	$\approx 1.68\%$	$\approx 1.59\%$	$\approx -1.25\%$
3	≈ 10	$\approx 2.81\%$	$\approx 9.70\%$	$\approx -1.25\%$

Table 4.7: Exponential value of the Fixed Effects for the OU model

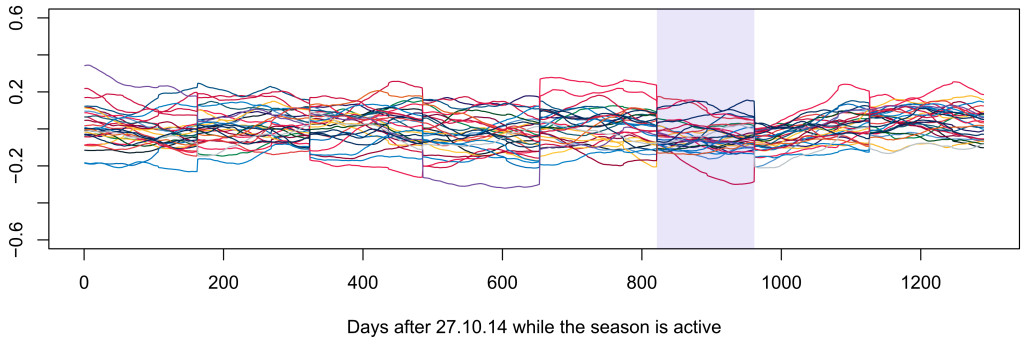

Figure 4.8: Strength of attacking type 1 using the OU model where the highlighted area is the covid season

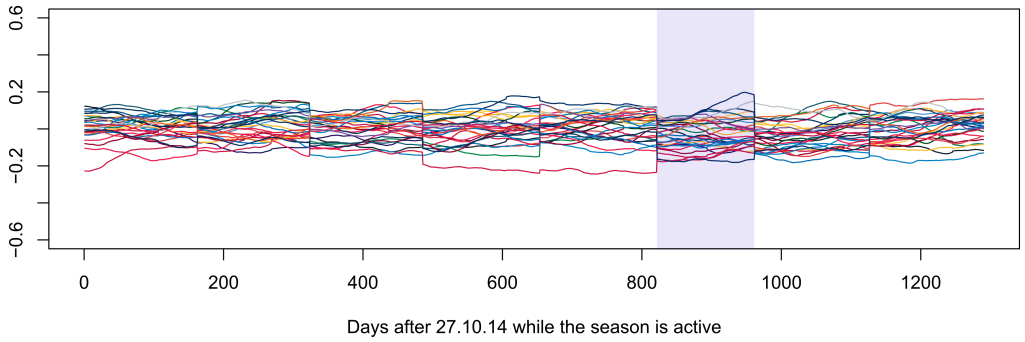

Figure 4.9: Strength of attacking type 2 using the OU model where the highlighted area is the covid season

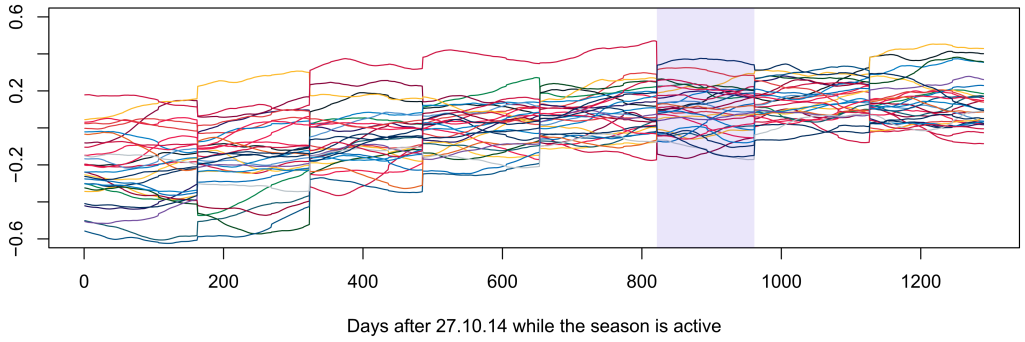

Figure 4.10: Strength of attacking type 3 using the OU model where the highlighted area is the covid season

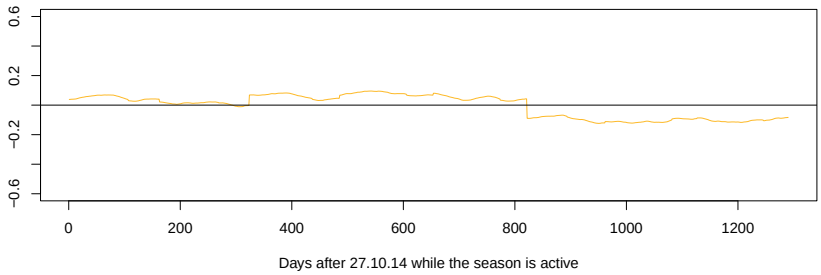

Figure 4.11: GSW Attack strength type 2

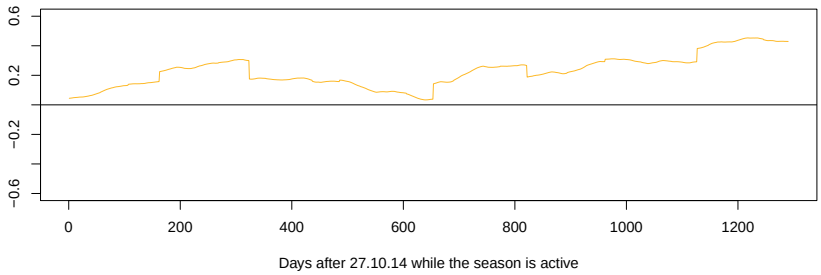

Figure 4.12: GSW Attack strength type 3

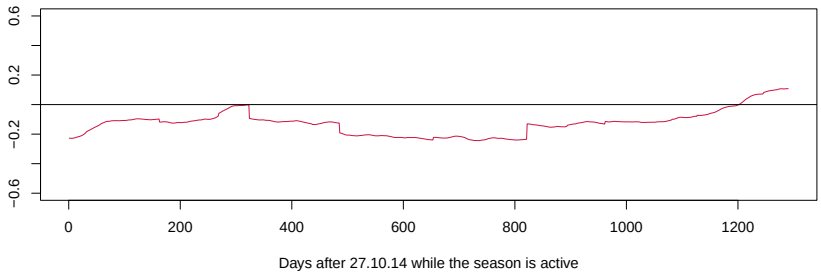

Figure 4.13: HOU Attack strength type 2

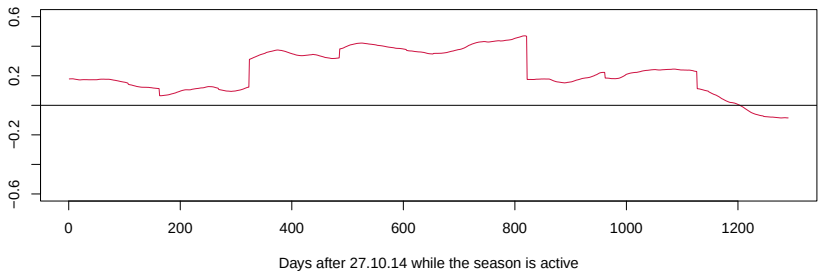

Figure 4.14: HOU Attack strength type 3

4.8 Future work

The variance of the number of fouls committed in a match is difficult to account for, because it is random, but in close matches there are an overabundance of fouls committed in an attempt to catch up to the opponent, which is closer looked at in [4]. To show that there is overdispersion for this in the data, I will first assume that the number of fouls committed, X in a match is Poisson distributed with rate λ, I will also assume that the number of scores, Y, conditional on the number of fouls is binomially distributed with parameters $2X$ and p, where we have an average of 2 attempts per foul. I.e.,

$$X \sim Poisson(\lambda), \tag{4.4}$$

$$Y|X = x \sim Bin(2X, p). \tag{4.5}$$

For finding the marginal expectation and variance of Y, I will use the law of total expectation and law of total variance.

$$\begin{aligned} E[Y] &= E[E[Y|X]] \\ &= E[2Xp] \\ &= 2\lambda p \end{aligned} \tag{4.6}$$

$$\begin{aligned} Var[Y] &= E[Var[Y|X]] + Var[E[Y|X]] \\ &= E[2Xp(1-p)] + Var[2Xp] \\ &= 2\lambda p(1-p) + 4p^2\lambda \\ &= E[y](1-p+2p) \\ &= E[y](1+p) \end{aligned} \tag{4.7}$$

Since $(1+p)$ is larger than one, this underlying data-generating process would generate overdispersion in the data.

CHAPTER

FIVE

CONCLUSIONS

In conclusion, the differences between the covid season and the other seasons analyzed is that the home-court advantage got smaller, but it did not disappear fully. This could be that because they still played near were they lived, so they still got a boost. Where that boost could come from could differ from player to player, but the main boost could come from better sleep or better relaxation before a match. This is because they were still in their home city. Also the number of fans at a match had a positive impact on all scoring types except three-pointers and they had a positive impact on the home-court advantage. The impact an individual fan had on the home-court advantage was positive in all three scoring types, but they had a negative impact on the number of three-pointers scored. This is because during covid, they scored more three-pointers, but they had a positive impact on the other scoring types. Also, we have seen that the attacking strengths of scoring type one and two have stayed pretty consistent, but the three-points attacking strengths of every team have increased a lot.

REFERENCES

[1] Gianluca Baio and Marta Blangiardo. "Bayesian hierarchical model for the prediction of football results". In: *Journal of Applied Statistics* 37 (Feb. 2010), pp. 253–264. DOI: 10.1080/02664760802684177.

[2] Matthew van Bommel. "Home Sweet Home: Quantifying Home Court Advantages for NCAA Basketball Statistics". In: (Jan. 2021). DOI: 10.3233/JSA-200450. URL: https://content.iospress.com/articles/journal-of-sports-analytics/jsa200450.

[3] L. Fahrmeir et al. *Regression. Models, Methods and Applications*. Springer, Jan. 2013. ISBN: 978-3-642-34332-2.

[4] Juan Manuel Martín-González et al. "The Poisson model limits in NBA basketball: Complexity in team sports". In: *Physica A: Statistical Mechanics and its Applications* 464 (2016), pp. 182–190. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2016.07.028. URL: https://www.sciencedirect.com/science/article/pii/S0378437116304599.

[5] Aparna Dhinakaran. *Understanding KL Divergence*. Feb. 2023. URL: https://towardsdatascience.com/understanding-kl-divergence-f3ddc8dff254. (accessed: 25.05.2023).

[6] J. Brockwell P. and A. Davis R. *Introduction to Time Series and Forecasting*. Springer, Apr. 2016. ISBN: 978-3-319-29852-8.

[7] Oliver C. Ibe. "9 - Brownian Motion". In: *Markov Processes for Stochastic Modeling (Second Edition)*. Ed. by Oliver C. Ibe. Second Edition. Oxford: Elsevier, 2013, pp. 263–293. ISBN: 978-0-12-407795-9. DOI: https://doi.org/10.1016/B978-0-12-407795-9.00009-8. URL: https://www.sciencedirect.com/science/article/pii/B9780124077959000098.

[8] Marko Voutilainen, Lauri Viitasaari, and Pauliina Ilmonen. "Note on AR(1)-characterisation of stationary processes and model fitting". In: *Modern Stochas-*

tics: Theory and Applications 6.2 (2019), pp. 195–207. ISSN: 2351-6046. DOI: 10.15559/19-VMSTA132.

APPENDICES

A - GITHUB REPOSITORY

All code and latex-files used in this document are included in the Github repository linked below.

Github repository link

- `https://github.com/Ola-R-R/Bachelor-Thesis`