Bachelor Thesis in Statistics and Data Analysis

A Bayesian approach to predict the number of soccer goals

Modeling with Bayesian Negative Binomial regression

Joakim Bäcklund
Nils Johdet

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköpings University

June 2018 | LIU-IDA/STAT-G–18/006–SE

Supervisor: Lecturer. Isak Hietala

Examiner: Lecturer. Ann-Charlotte Hallberg

Abstract

This thesis focuses on a well-known topic in sports betting, predicting the number of goals in soccer games. The data set used comes from the top English soccer league: Premier League, and consists of games played in the seasons 2015/16 to 2017/18. This thesis approaches the prediction with the auxiliary support of the odds from the betting exchange Betfair. The purpose is to find a model that can create an accurate goal distribution. The methods used are Bayesian Negative Binomial regression and Bayesian Poisson regression. The results conclude that the Poisson regression is the better model because of the presence of underdispersion. We argue that the methods can be used to compare different sportsbooks accuracies, and may help creating better models.

Acknowledgements

We would like to express our gratitude to our supervisor Lecturer. Isak Hietala for his perpetual guidance and assistance in keeping the progress on schedule. We would also like to extend our gratitude to Ph.D. Student Per Sidén for valuable insights and constructive suggestions. We would also like to thank Assistant professor Bertil Wegmann for ideas regarding the Bayesian modeling, his willingness to give his time so generously has been very much appreciated. Lastly, we wish to express our gratitude to our opponents Sjoerd Schelhaas and Hugo Hjalmarsson for providing much appreciated and useful feedback on the thesis.

Contents

List of Figures

List of tables

Keywords

Odds – A reflection of the likelihood of a possible event expressed numerically. In betting, the decimal odds is expressed as the ratio of payoff to the stake wagered.

Implied probability – A conversion of odds into a percentage, calculated by the inversion of the odds.

Sportsbook – An organization that accepts bets usually on sports. They handle the odds pricing, correction of the result and the payout of the winning.

Betting exchange – A service where the customers can choose to lay (give) odds, or place bets at other customers odds, also known as a prediction market, similar to a future exchange. The betting exchange provides the platform, leagues and games, correction of result and the payout of the winnings.

Total – A common bet in sports is whether the total number of goals scored by both teams is over or under a certain number, called the total-line.

Line – A number set by the the market or sportsbook before the event, where bets can be placed on over or under the given number.

1. Introduction

This chapter provides an introduction to sports betting and the betting market exchange Betfair. The second section presents previous studies in the field of goal predicting. The third section covers the purpose of this thesis, and the last section provides a reflection regarding the social and ethical aspects of this thesis.

1.1 Background

This section describes the history of sports betting, and a description of the betting market exchange Betfair.

1.1.1 Sports betting

Gambling in general dates back to before written history; while sports betting have allegedly existed for as long as sports has been around, there are records of gambling at sports events and outcomes of gladiator fights from the Roman empire. [1]

Before sports betting was legalized in Nevada in 1931; people in the U.S placed their wagers trough privately run enterprises referred to as "bookies". In United Kingdom, sports betting was not allowed until 1961. In 1994, Antigua and Barbuda was the first country to pass a law that allowed operators to apply for online gambling licences. However, sportsbooks did not get involved until 2001 when U.K territories Isle of Man and Gibraltar began to offer licenses. [2]

Thanks to the sports betting industry's online introduction, a great number of sportsbooks has emerged. The competition has driven their margins down, and put more pressure on the accuracy of their models, to still continue to generate profit. The lower margins has also given the market more incentive to try to beat the sportsbooks.

1.1.2 Soccer

The most popular sport is soccer, in sports betting this is not an exception. About 70 percent of the market share is estimated to come from soccer. The most popular type of betting is outcome betting, but the competition between sportsbooks has resulted in the appearances of other bet-types such as the total. [3]

1.1.3 Betfair

A well-known online gambling website is Betfair which was established in 2000. The company is particularly known for its betting exchange which is one of the largest in the world. The customers get to decide what events they are willing to place or lay bets on and to what odds. This results in a larger scale of possible wagers to be found compared to if the wagers were to be decided by the sportsbook itself. [4]

1.1.4 Odds

Odds for games are often available days in advance, before the actual start of the event. Information such as scoring average, player injuries, weather condition, and team line-ups can be expected to be reflected in the odds. This is due to a concept known as the wisdom of the crowds coined by James Surowiecki. He implies that the collective wisdom is often more accurate than a judgement from one single person. [5]

1.2 Previous studies

There have been plenty of previous studies regarding models that predict the expected number of goals in soccer. Most of them focus on using home and away scoring averages to predict the total number of goals in an upcoming game.

One thesis researches the betting markets risk management regarding closing odds; the authors use a time-independent Poisson model to predict the results and compare it to the odds. They state that this model is quite similar to the model that some sportsbooks already use. They conclude that the odds, in some cases, are beatable by the model, but it is often adjusted by the sportsbook next year. The method the authors used is foremost established from an article from Maher. [6] Maher remarks in the summary that “Previous authors have rejected the Poisson

model for association football scores in favour of the Negative Binomial." Mahers article investigates the Poisson model further by including parameters for attacking and defensive strengths of each team. The author draws the conclusion that an independent Poisson model gives a good description of the number of goals in soccer games, but improvements can be achieved by using a bivariate Poisson model instead of the independent Poisson model. [7]

In a recently published article in the International Journal of Forecasting, the authors look into a bivariate Weibull model to predict results and number of goals in soccer games. The authors rejected the Poisson distribution in favor of the Weibull count distribution, which provided better predictions of both results and the number of goals. [8]

Another thesis evaluates if the odds are beatable by trying three different models to predict the number of goals in a soccer game. These three models are the following:

- Gamblers assessment - using previous number of goals made by teams and calculating an average.
- Poisson Distribution Assessment - This model assumes that the number of goals follow the Poisson Distribution and uses the average of the previous goals and inserts it into the Poisson Distribution.
- Dixon-Coles Assessment - This model is based on previous number of goals scored, creating parameters for the team's offensive and defensive strength. It also takes into account whether the team plays at home or away.

The author concludes: "that with the approaches taken, it was not possible to create probability assessments which were better than those of the bookmakers. However, results show, that it is possible to almost match them." [9]

1.3 Purpose

The purpose of this thesis is to find a model which can predict the distribution of the total numbers of goals in soccer games, using total-lines odds set by the market. It would be convenient to use one or two total-lines to represent all the lines. Because sometimes all lines are not available, another reason is that a simpler model is often to be preferred. Also, if one or two total-lines would to represent all the total-lines, it would be an indication that the odds is useful when predicting the number of goals in soccer.

1.3.1 Research questions

- Can the odds be used to create a useful predictive goal distribution?
- Is negative binomial regression appropriate to model soccer goals in Premier League?

1.4 Social and ethical aspects

This thesis does not use any data that can be connected to a certain person or object which means that no cautions of data privacy policy has to be considered.

However, gambling is a controversial health issue and can have negative economical and social consequences. Gambling can become an addictive pursuit for people. The proliferation of internet, and as consequence profusion of sportsbooks, provides people with a perpetual appeal towards gambling. This thesis does not, in any way, recommend people to try gambling.

1.5 Delimitations

The data set in this thesis does not consist of all games played in Premier League during the seasons 2015/16-2017/18. The reason is that the data source Betexplorer did not have all the total lines available for every game. Therefore, only games with lines available between 0.5-4.5 has been selected.

2. Data

The data consist of the number of goals scored, total lines and odds from 687 soccer games played in the Premier League from season 2015/16 to 2017/18. Premier League is the highest ranked league of the English soccer league system and consist of 20 competing teams each season. Each team plays 38 games throughout the season. The data was scraped from the website Betexplorer. [10]

Betexplorer saves the results and the odds for each game from a number of sportsbooks and betting exchanges. The table below presents an example of the raw data.

Table 2.1: Example data of one observation from Betexplorer

game ID	sportsbook	total line	Over	Under
1	Betfair	1.5	1.5	2.8
1	Betfair	2.5	1.95	1.95
1	Betfair	3.5	2.9	1.4

Table 2.1 presents three total lines from a single game and their corresponding odds for over and under the total line. The over and under columns represent the closing odds on each outcome.

2.1 Data processing

The odds from table 2.1 are converted and stored as implied probabilities. The implied probabilities are an indication of how often a bet must win for it to break even in the long run. Suppose that someone wager 1 dollar on a bet with 2.0 odds and if the wager is won, 2 dollar will be handed out. In this case the implied probability is 50% which means he or she has to win half the time to break even in the long run.

The firs step in order to calculate the normalized implied probabilities given by the odds is to determine the factor used to normalize the implied probability (nmf) which is done by the following equation:

$$nmf = \sum_{i=1}^{n} \frac{1}{\text{odds of outcome } i}$$

Where n is the number of possible outcomes.

The next step is to calculate the breakeven odds (BE_{odds}). This is done by the following equation:

$$BE_{odds} = nmf \cdot \ odds \tag{2.1}$$

The implied probabilities are then calculated by inverting the BE_{odds} that was determined in equation 2.1.

When this procedure is done, the data of implied probabilities are stored into a database and extracted in the following format.

Table 2.2: Example of processed data with the implied probability for Over each line

obs.	0.5	1.5	2.5	3.5	4.5	5.5	6.5	Goals
1	0.917	0.746	0.520	0.296	0.139	0.059	0.022	4
2	0.878	0.642	0.373	0.192	0.080	0.031	0.010	0
3	0.944	0.800	0.572	0.358	0.185	0.086	0.035	1
4	0.928	0.747	0.490	0.284	0.129	0.056	0.021	1
5	0.978	0.901	0.752	0.565	0.361	0.208	0.099	2

Table 2.2 presents the implied probabilities for the outcomes over 0.5 to 6.5 goals from five different games and the number of goals scored in each game. This is the final data that was received after processing the original data.

2.2 Distribution of the number of goals

To get a visual understanding of how the number of goals is distributed, a bar graph is presented below.

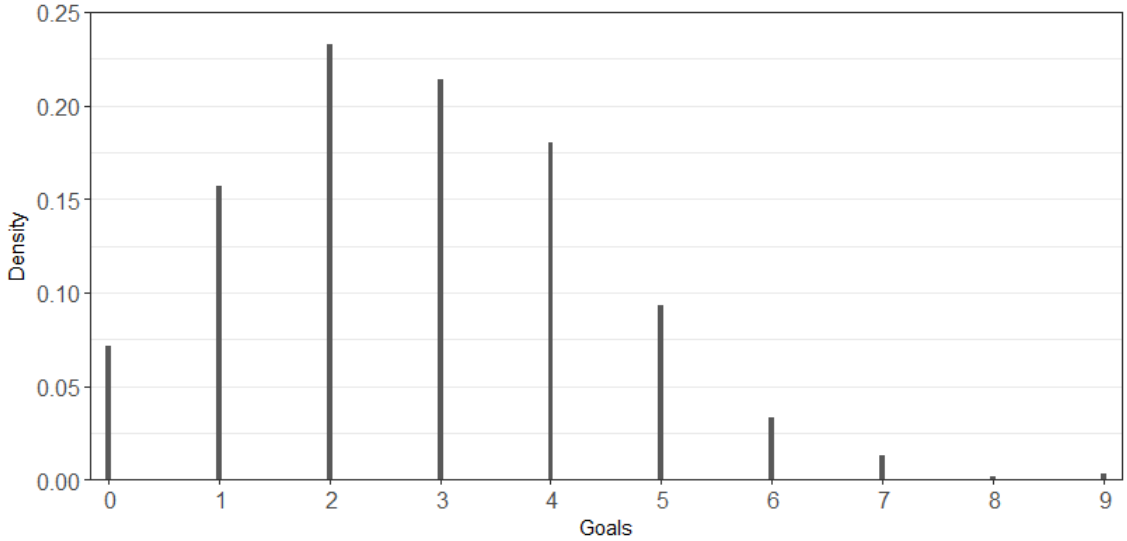

Figure 2.1: Bar graph of soccer goals distribution in the data set

Figure 2.1 illustrates a barplot for the goal distribution in the data set. The figure shows that most of the games end with two or three goals, and more than four goals are not as common.

3. Methods

In this chapter, the relevant distributions are first described; the second section covers Bayesian inference and modeling. The third section describes Markov Chain Monte Carlo algorithms and the diagnostic tools used. The fourth section is about model evaluation and comparison. In the last section, the packages used for Bayesian modeling in R are described.

3.1 Distributions

This section describes the distributions Poisson, Negative binomial and their relationship to each other. The section also covers the Gamma-Poisson distribution and how it is relevant for this thesis.

3.1.1 Poisson distribution

The Poisson distribution is often used for the counts of event that occur in a given interval of time. The assumptions of the Poisson distribution are the following

- The number of times an event occurs is denoted k and can take values as 0, 1, 2 ... n.
- Trials are made independently.
- The event has the same probability to occur throughout the whole time-interval.

The probability mass function has the following formula:

$$P(k \mid \lambda) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

where k is the actual number of events occurring from the Poisson experiment, λ is the average number of events occuring (mean) and is also equal to the variance. [11]

3.1.2 Negative binomial distribution

The negative binomial distribution is a discrete probability distribution of the number of failures from a sequence of independent Bernoulli-trials, until a specified (and fixed) number of successes occurs. The negative binomial distribution has two parameters r and p and has the probability mass function

$$f(k \mid r,p) = p(X = k) = \binom{r+k-1}{k} p^k (1-p)^r \quad \text{for } k = 0, 1, 2, ... \tag{3.1}$$

where k equals the number of **successes** that occur before the rth **failure** and p is the probability of success.

Another common formulation is

$$f(k \mid r,p) = p(X = k) = \binom{r+k-1}{k} p^r (1-p)^k \quad \text{for } k = 0, 1, 2, ...,$$

where k equals the number of **failures** that occur before the rth **success** and p is the probability of success. [11]

When counting the number of X failures before the r:th success, the expected number of failures is

$$E[X] = \frac{r(1-p)}{p} = \mu$$

and the variance

$$Var(X) = \frac{r(1-p)}{p^2} = \frac{r(1-p)p + r(p-1)^2}{p^2}$$

$$= \frac{r(1-p)}{p} + \frac{r(p-1)^2}{p^2} = \mu + \frac{\big(r(p-1)^2\big)r}{p^2} \cdot \frac{1}{r}$$

$$\Rightarrow Var(X) = \mu + \frac{\mu^2}{r}$$

as $r \to \infty$, Negative Binomial distribution converges to the Poisson distribution, And for small r, Negative Binomial gives a larger variance than Poisson. Therefore, r is referred to as the dispersion or shape parameter. Hence, the Negative Binomial distribution is an alternative to the Poisson distribution when the variance is greater than the mean (overdispersion). [12]

3.1.3 Gamma-Poisson mixture

Consider a Poisson distribution with parameter λ. The λ parameter follows a gamma distribution with shape parameter r, and the scale parameter $\theta = \frac{p}{1-p}$ which can be expressed as the rate parameter $\beta = \frac{1}{\theta} = \frac{1-p}{p}$. The mixture model can be expressed as

$$f(k \mid r,p) = \int_0^\infty f_{Poisson}\Big(k|\lambda\Big) \cdot f_{Gamma}\Big(\lambda|r,\tfrac{1-p}{p}\Big) d\lambda$$

$$= \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \cdot \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{(\frac{p}{1-p})^r \Gamma(r)} d\lambda$$

$$= \frac{(1-p)^r p^{-r}}{k!\Gamma(r)} \int_0^\infty \lambda^{r+k-1} e^{-\lambda/p} d\lambda$$

$$= \frac{(1-p)^r p^{-r}}{k!\Gamma(r)} p^{r+k} \Gamma(r+k)$$

$$\Rightarrow f(k \mid r,p) = p(X=k) = \frac{\Gamma(r+k)}{k!\Gamma(r)} p^k (1-p)^r \qquad \text{where } \{r \in \mathbb{R} \mid r > 0\}$$

$\frac{\Gamma(r+k)}{k!\Gamma(r)}$ corresponds to $\binom{r+k-1}{k}$ in the probability mass function of the negative binomial in Eq. 3.1. This means that r now has been extended to all positive real values which will prove to be an essential part in subsection 3.3.2. Hence, the Negative binomial distribution can be described as a Poisson distribution mixed with the Gamma distribution where λ is gamma distributed with shape parameter r and scale parameter $\theta = \frac{p}{1-p}$. [13]

3.2 Bayesian Inference and Modeling

This section describes how Bayesian inference works and how it can be applied to a regular multiple regression.

If no citation is specified in this section, it can be assumed that the source is R. McElreath, Statistical rethinking 2016. [14]

Bayesian inference gives the opportunity to keep previous information in the analysis. The prior information that the user would want to keep in the analysis can be information from experience, previous experiments or data. Suppose the user is a non-expert in the subject and has no previous information available. The user should then use a vague prior which means the prior distribution play a minimal role in the posterior distribution.

After deciding how vague the prior should be, the model is updated by actual data which educates the model further from the prior information, which results in a posterior. This update is performed by Bayes' theorem which is a fundamental part of Bayesian modeling. Bayes theorem is used to calculate conditional probabilities and is represented by the following equation

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)}$$

The equation below shows how Bayesian modeling is performed using Bayes theorem

$$p(\theta \mid \boldsymbol{data}) = \frac{p(\boldsymbol{data} \mid \theta) \cdot p(\theta)}{p(\boldsymbol{data})} \quad \propto \quad p(\boldsymbol{data} \mid \theta) \cdot p(\theta)$$

where θ is the set of parameters. Hence $p(\theta)$ is the prior, $p(\boldsymbol{data} \mid \theta)$ is the likelihood of data given the parameters, and $p(\theta \mid \boldsymbol{data})$ is the posterior probability distribution for the parameters θ given the observed data.

For every parameter intended to be estimated in a a Bayesian model, an initial sets of plausibilities has to be provided which are the priors.

3.2.1 Non-bayesian approach to regression

Consider a standard multiple linear regression, the equation is defined as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is n $\times$ 1 column vector of the response variable, $\mathbf{X}$ is a n $\times$ k matrix, $\boldsymbol{\beta}$ is a k $\times$ 1 vector of regression coefficients, and $\boldsymbol{\epsilon}$ is a n $\times$ 1 column vector of independent and identically normally distributed random variables.

This formula can be rewritten as the probabilistic model behind it as

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$$
$$y_i \sim \mathcal{N}(\mu_i, \sigma)$$

Where the elements y_i in $\mathbf{Y}$ follows the normal distribution with mean μ_i and standard deviation σ.

The parameters can be estimated by using ordinary least squares, done by minimizing the squared errors of fitted values, or by maximizing the likelihood function

$$\hat{\theta}_{\mathbf{MLE}} = \arg\max_{\boldsymbol{\beta},\sigma} \prod_{i=1}^{n} \mathcal{N}(y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma) \tag{3.2}$$

where θ is the set of parameters in the vector $\boldsymbol{\beta}$ and σ. $\mathcal{N}$ is a probability density function of the normal distribution, evaluated at y_i with mean $\boldsymbol{x}_i'\boldsymbol{\beta}$ and standard deviation σ. [15]

3.2.2 Bayesian approach to regression

In the Bayesian approach, instead of maximizing the likelihood function, each parameter is assigned a prior distribution, and then Bayes theorem is used

$$\underbrace{f(\boldsymbol{\beta}, \sigma \mid Y, X)}_{\text{posterior}} \propto \underbrace{\prod_{i=1}^{n} \mathcal{N}(y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma)}_{\text{likelihood}} \underbrace{f_{\boldsymbol{\beta}}(\boldsymbol{\beta})\, f_{\sigma}(\sigma)}_{\text{priors}}$$

When uniform priors are used, they correspond to $f_p(x) \propto 1$. Because the likelihood function is the same as in Eq. 3.2, maximum likelihood of the parameters will be the same as its Bayesian counterpart, maximum a posteriori probability (MAP) estimate with uniform priors.

If a posterior distribution belongs to the same distribution family as the prior probability distribution, it is called a conjugate prior. It is convenient because the parameters of the posterior distribution are then directly available.

3.2.3 Poisson regression

Poisson regression is a generalized linear model that assumes the response variable y follows a Poisson distribution. The model takes the form

$y_i \sim Poisson(\lambda_i)$

$log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$

$\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma)$

where $\boldsymbol{x}_i$ is a (1 + k) $\times$ 1 vector of k numbers of explanatory variables, and $\boldsymbol{\beta}$ is a (k+1) $\times$ 1 vector of regression coefficients.

σ is arbitrarily but some guidelines are:

- vague: $\sigma = 10^6$
- weak informative: $\sigma = 10$
- informative prior: $\sigma = 1$

A log link is applied to ensure that the parameter λ maps only to positive values.

3.2.4 The Negative Binomial case

A Negative binomial regression can be written as an extended Gamma-Poisson mixture generalized model. What makes it extended is the use of two linear functions, one for each parameter. It can be compared to fitting a regression model with normally distributed data with non-constant variance for σ. That situation would also require a linear function for each parameter. [16]

The gamma probability density function with shape parameter k and scale parameter θ has the following probability density function

$$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

with the mean $E[X] = k\theta$ and variance $Var(X) = k\theta^2$

By inserting the shape parameter from section 3.1.3 $k = r$ and the scale parameter $\theta = \frac{p}{1-p}$ It follows that μ is equal to $\frac{rp}{1-p}$.

$y_i \sim GamPois(\mu_i, \theta_i)$

$log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_\mu$
$\log(\theta_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_\theta$

Where $\boldsymbol{x}_i$ is a (1 + k) $\times$ 1 vector of k number of explanatory variables, $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_\theta$ are two (1 + k) $\times$ 1 vectors of the regression coefficients.

Independent priors for each coefficients in the $\boldsymbol{\beta}$ vectors:

$\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma)$

σ is arbitrary.

The parameters μ and θ must be positive in order to ensure that r is positive and p to map between 0 and 1; therefore, a log link is applied to each of the parameters.

3.3 Markov Chain Monte Carlo (MCMC)

If no citation is specified in this section, it can be assumed that the source is R. McElreath, Statistical rethinking 2016. [14]

Markov Chain Monte Carlo (MCMC) are a group of algorithms which purpose is to sample from the posterior by constructing a Markov Chain that uses the posterior distribution as the marginal distribution. The MCMC is commonly used when a conjugate prior cannot be used. The samples that are obtained by this process are used to approximate the posterior. The process requires no assumption of the shape of the posterior distribution, which makes it possible to sample directly from it. Generalized linear- and multilevel models, which produce non-Gaussian posterior distributions, has a great benefit of using this process. This is because the MCMC has the ability to directly estimate models, without assuming multivariate normality for instance. Besides these benefits that characterizes MCMC, the process is very time-consuming and some added monitoring of the process is also required to ensure the MCMC is performing well, which will be explained in subsection 3.3.3 .

3.3.1 Markov Chain

A Markov Chain is a stochastic mathematical process that generates transitions between different states of a variable. The process can be used on both discrete and categorical variables and are used to analyze how the outcome of the variable

changes within two consecutive time-periods. The set of all possible states for the variable is called state space which can include different types of states such as weather conditions, goals scored etc.

Information on the probability of transitioning from one state to another in the process at time t, is given by a transition matrix. A process that can describe a transition to n different states, $L_1, L_2, ..., L_n$. The probability for the process to be in a certain state at time t is presented in a vector such as

$$\boldsymbol{x}_t = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

Furthermore, the probability for the process to transit between one state to another can be presented in a transition matrix

$$P = \begin{bmatrix} 1 \to 1 & 2 \to 1 & ... & n \to 1 \\ 1 \to 2 & 2 \to 2 & ... & n \to 2 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 \to n & 2 \to n & ... & n \to n \end{bmatrix}$$

Since the elements in the transition matrix is probabilities, they range from 0 to 1 and each column in the matrix sums to 1.

With the aid of the two above-mentioned matrices, the probability vector for time $t + 1$ can now be calculated by the equation

$$\boldsymbol{x}_{t+1} = P\boldsymbol{x}_t$$

By repeating the equation, probabilities further into the future can be calculated by

$$\boldsymbol{x}_{t+s} = P^s\boldsymbol{x}_t$$

Where s is the number of steps into the future.

By letting the probability distribution of $\boldsymbol{x}_t$ be given by the $n \times 1$ vector $\boldsymbol{\pi}$, where n is the number of states. A Markov Chain has reached an equilibrium distribution $\boldsymbol{\pi}$ once it satisfies

$$P\boldsymbol{\pi} = \boldsymbol{\pi}$$

hence, $\boldsymbol{\pi}$ is an eigenvector with the eigenvalue 1. [17]

Regardless of which initial starting state that is chosen, the equilibrium probability distribution of states will be reached where no more changes will occur in the distribution. Different type of MCMC algorithms use this to construct stationary Markov Chains, so that the equilibrium probability distribution is the target distribution. If a stationary chain can be constructed, the chain can be iterated from an arbitrarily starting point many times. The draws generated would appear to be coming from the target distribution.

3.3.2 Hamiltonian Monte Carlo

In statistics, Monte Carlo refers to algorithms used to solve computationally heavy problems through simulations of random numbers and estimate the sample average i.e. of the draws from the Markov Chains with the help of the law of large numbers. The Hamiltonian Monte Carlo (HMC) is an algorithm to sample from an unknown posterior distribution through the MCMC process.

The HMC is an effective algorithm when models consist of hundreds or even thousands parameters. The HMC can be thought of as a algorithm which pretends that the vector of parameters determine the position of a particle that has no friction, comparable to a physics simulation. HMC builds upon knowing how the density is changing at the particle's current location. The surface for the frictionless particle to glide across is provided by the log-posterior. Depending on if the log-posterior is very flat or very steep, the particle can glide for a long period of time or a short period of time until it turns around. When the particle turns, it is because of the gradient changes direction.

The particle can glide for a long period of time until it changes direction when the log-posterior is very flat due to lack of information in the likelihood and flat priors. The particle does not glide for a long period of time until it turns around when the log-posterior is very steep due to very concentrated likelihood or priors. This process provides an understanding of how the parameter's distribution is scattered by learning from the gliding particle. The more time the particle spends at a location, the more dense the log-posterior and vice versa.

Only when the parameters are continuous, Hamiltonian Monte Carlo can be used since the particle cant glide through a discrete parameter's surface.

3.3.3 MCMC Diagnostic

In order to assess whether the convergence of the MCMC algorithm has occurred or not, a number of diagnostics can be used.

The most useful way for diagnosing a Markov Chain is too inspect a trace plot. A trace plot shows the samples in sequential order for each parameter. The first part of the plot is the warm-up phase (the gray regions in figure 3.1), where the chain is adapting for efficient sampling. The remaining region of the plot is the sampling used for inference. By inspecting the plot, a healthy Markov Chain can be identified by stationarity and good mixing. Stationarity means that the mean is stable through the plot, and a well-mixing chain means each sample is not highly correlated with the sample before it. A low or non-existent correlation between each sample provides a greater amount of information from a given number of draws from the posterior. Figure 3.1 illustrates two trace plots from an unhealthy (left) and a healthy Markov Chain (right).

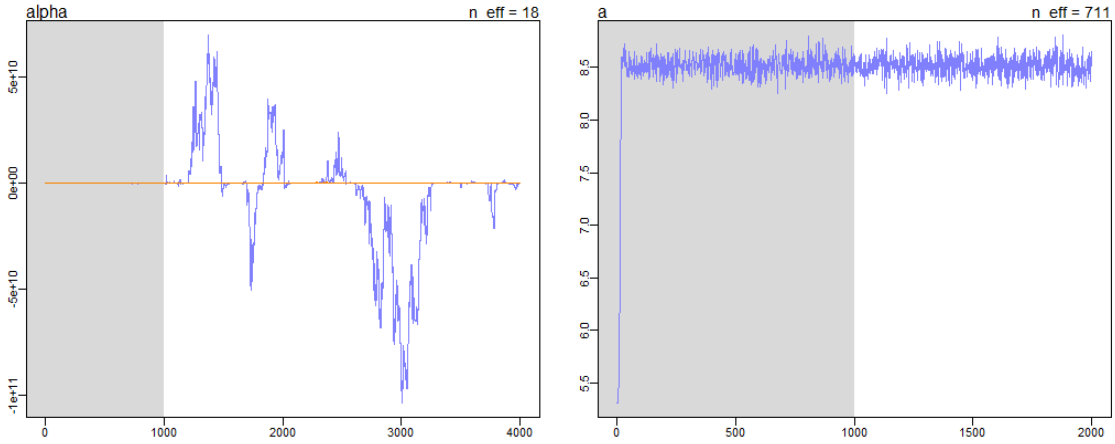

Figure 3.1: Trace plot comparison of an unhealthy and a healthy Markov Chain

One metric used for diagnostic is called $\hat{R}$. The metric gives an indication of whether a chain has converged to the equilibrium distribution. It is done by comparing its behavior to other randomly initialized chains. The $\hat{R}$ statistic measures the ratio of the average variance of samples, within each chain, to the variance of the pooled samples across all chains. If all the chains are at an equilibrium, these will be the same and $\hat{R}$ will be equal to one. If the chains have not converged to a common distribution, the $\hat{R}$ statistic will be greater than one. [18]

Another metric used is the effective sample size n_{eff}, it is an estimate of the number of independent draws from the posterior distribution where anything greater than 100 is considered adequate. Because the draws within a Markov Chain are not independent if there is autocorrelation, the effective sample size, n_{eff}, will be smaller than the total sample size, N. The larger the ratio of n_{eff} to N the better.

Plots for the accumulated posterior means and quantiles, for each parameter, can also be inspected for ensuring the convergence to a fixed value.

A Highest Posterior Density Interval (HPDI) is the smallest possible interval containing the probability mass specified. The interval is similar to a credibility interval, which is reminiscent of a confidence interval, but a HDPI can differ from a credibility interval when the posterior distribution is skewed or multimodal.

3.4 Model evaluation and comparison

If no citation is specified in this section, it can be assumed that the source is R. McElreath, Statistical rethinking 2016. [14]

There are many ways to evaluate and compare Bayesian models. Comparing the models can be useful even if the predictive accuracy is considered poor because it can help to decide where to go next in order to improve the model.

3.4.1 Kullback-Leibler divergence

One measurement used to provide a distance for how much a model deviates from a perfect model is called Information entropy and is is defined as

$$H(p) = -E[log(p_i)] = -\sum_{i}^{n} p_i(log(p_i))$$

Where there are n possible events, and each event i has the probability p_i

Information entropy needs to be quantified to be able to say how far a model is from the target distribution. The Kullback-Leibler divergence is the average difference between the true target distribution p and the predicted distribution q. The difference is measured in log probability and the formula is

$$D_{KL}(p \mid\mid q) = \sum_{i} p_i(log(p_i) - log(q_i)) = \sum_{i} p_i \; log\Big(\frac{p_i}{q_i}\Big) \tag{3.3}$$

In a scenario where $D_{KL} = 0$, q can be used to predict the true target distribution p which means the model that was used to predict the distribution q is the perfect model.

3.4.2 Deviance

While K-L Divergence provides a measure of distance, the target distribution p is still unknown. However, by comparing the divergences of two models q and m for example, it is known from Eq. 3.3 that $log(p_i)$ is used for both calculating q and m. Which means, when comparing them, they are subtracted from each other and shows that all p's can be canceled; it has no impact on how far the comparing models are apart. Hence, each model's average log-probability is sufficient for comparing models, and an approximation of these averages is obtained by summing the log-probabilities of each observed case for q and m.

This measurement is called deviance, and it is a relative model fit measurement and an approximation of K-L divergence. Deviance for model q is defined by

$$D(q) = -2\sum_{i} log(q_i)$$

where i denotes each case (observation) and each q_i is the likelihood of case i.

3.4.3 Widely Applicable Information Criterion (WAIC)

Unfortunately, deviance has the same flaw as the coefficient of determination; it always improves when the model gets more complex. Information criteria corrects that flaw by taking the complexity of the model into account and penalize complex models.

WAIC is the generalized version of the Akaike information criterion (AIC), it is an example of an information criterion for out-of-sample deviance. WAIC is calculated by taking averages of the loglikelihood over the posterior distribution. It does not require a multivariate normal posterior distribution, compared to other information criterion and is often more accurate.

The main feature of the WAIC is that it is pointwise. This is advantageous because the observations may have different uncertainties, and some can be harder to predict than others. The disadvantage is that it requires independent observations.

The WAIC consist of two parts, the first part is the log-pointwise-predictive-density:

$$lppd = \sum_{i=1}^{N} log[p(y_i)] \tag{3.4}$$

where $p(y_i)$ is the likelihood of observation y_i.

For each set of parameters sampled from the posterior distribution, the *lppd* computes the likelihood of observation y_i. It will then average the likelihoods of each observation i and sum over all observations. This is an analog of deviance averaged over the posterior distribution.

The second part is the effective number of parameters p_{WAIC}

$$p_{WAIC} = \sum_{i=1}^{N} V(y_i) \tag{3.5}$$

where $V(y_i)$ is the variance in log-likelihood for observation i.

Combining Eq. 3.4 and Eq. 3.5, WAIC is defined as

$$WAIC = -2(lppd - p_{WAIC})$$

and is an estimate of out-of-sample deviance.

3.4.4 Akaike weights

The Akaike weights is used to compare the models relative predictive accuracy. The weights are calculated by converting the expected deviance given by WAIC to a probability scale. In a set of m models, the weight for model i is given by

$$w_i = \frac{exp\Big[-\frac{1}{2}dWAIC_i\Big]}{\sum\limits_{j=1}^{m} exp\Big[-\frac{1}{2}dWAIC_j\Big]}$$

where $dWAIC_i$ is the difference between model i WAIC value and the model with the lowest WAIC.

The sum of all weights in the set of models will equal to 1 and each individual model will have a weight between 0 and 1. Each weight can be interpreted such as the probability that the model i has the best prediction ability, given the set of models.

3.5 Implementation in R

3.5.1 RStan Version 2.17.3

Stan is a state-of-the-art platform for statistical modeling and high-performance statistical computation. RStan is the R interface to Stan.

3.5.2 rethinking Version 1.59

The package rethinking essentially consists of two functions, map and map2stan. These functions force the user to build up a statistical function. In this thesis the function map2stan is used. The function builds a Stan model that is used to fit the model with Hamiltonian Monte Carlo sampling.

The functions dgampois and rgampois computes the density and produces a random sample from a Gamma-Poisson mixture probability distribution.The function parameters are the gamma parameters of the mean μ and the scale θ. Internally, the function uses dnbinom and rnbinom, which takes the parameters r and p. The parameters r and p are calculated by $r = \frac{\mu}{\theta}$ and $p = \frac{\theta}{1+\theta}$.

4. Results

In this chapter the model evaluation and comparisons are presented in order to determine which models that has the best prediction ability. The second section presents the MCMC diagnostics for the two best models. The last section illustrates the predictive abilities of the posterior distributions for these models.

4.1 Model comparison

In order to compare the different models predictive ability, a table is presented below with WAIC values and Akaike weights for each model.

Table 4.1: WAIC model comparisons

Model (expl. var.)	WAIC	pWAIC	dWAIC	Akaike Weight
Poisson (3.5)	2562.8	1.8	0.0	0.67
Poisson (2.5)	2564.3	1.7	1.5	0.32
Poisson (1.5)	2571.6	2.1	8.8	0.01
Negative Binomial (3.5)	2714.7	2.2	151.8	0.00
Negative Binomial (4.5)	2715.1	2.3	152.3	0.00
Negative Binomial (2.5)	2715.8	2.3	153.0	0.00
Negative Binomial (5.5)	2715.9	2.3	153.1	0.00
Negative Binomial (0.5, 1.5)	2717.0	3.8	154.1	0.00
Negative Binomial (0.5, 3.5)	2718.4	3.4	155.5	0.00
Negative Binomial (0.5, 4.5)	2718.8	3.5	156.0	0.00
Negative Binomial (0.5, 2.5)	2718.8	3.3	156.0	0.00
Negative Binomial (1.5)	2719.6	2.3	156.7	0.00
Negative Binomial (0.5)	2727.1	2.1	164.3	0.00

Table 4.1 presents the WAIC values and Akaike weights for each of the models. To each of the models, the same weak informative prior has been used for all parameters, normally distributed with mean 0 and standard deviation of 10.

The top three models that has the lowest WAIC values are all Poisson models. This value is an estimate of out-of-sample deviance and provides the conclusion that these three models has better predictive capability than the remaining models. Hence, the Poisson models has the best ability for predictions.

The Akaike weights confirms that the Poisson models are to prefer over the Nega-

tive Binomial models to predict new observations. The probability for the Poisson model with explanatory variable 3.5 to be chosen as the model with best predictive ability is 67 %. The table shows that there are no Negative Binomial Model that have a probability above zero which means that neither of them can be chosen as the model with best predictive ability. [1]

To summarize the model comparison of the thirteen models that is presented in table 4.1. the Poisson model with predictor variable 3.5 turns out to be the best model. The best Negative Binomial model is when the predictor variable 3.5 is used. These two models will be further analyzed in subsections 4.2.1 and 4.2.2.

4.2 MCMC Diagnostic

This section presents the MCMC diagnostics of the best Poisson model and the best Negative Binomial model that was selected from the model comparison. By observing MCMC diagnostics for each model, it is possible to determine whether they have fulfilled all the criteria for being reliable and useful models.

4.2.1 Poisson model with total line 3.5

A summary of the posterior result of the Poisson model with total line 3.5 is presented below.

Table 4.2: Parameter estimation and diagnostics, Poisson model (3.5)

	Mean	StdDev	lower 0.909	upper 0.909	n_{eff}	$\hat{R}$
β_{01}	0.58	0.10	0.42	0.75	156	1.00
β_1	1.45	0.29	0.98	2.01	158	1.00

Table 4.2 shows that the posterior mean for the intercept is 0.58 and 1.45 for the slope parameter. *0.909* and *upper 0.909* is a 90.9 percent highest posterior density interval. The reason why 90.9% was chosen as limit was because it provides a simple interpretation of the interval. The odds for the parameters posterior mean to be within the interval is $\frac{0.909}{1-0.909} = 10$, which means it is ten times more likely that the parameters posterior mean is within the interval than outside.

The intervals for β_{01} and β_1 indicate that both posterior distributions are reliably above zero. n_{eff} is greater than 100 and $\hat{R}$ is approximately 1, this indicates that the Markov chain has converged to an equilibrium distribution.

[1] Note that the probabilities are rounded

To investigate further whether the model has a well calibrated Markov Chain or not, a trace plot is presented below.

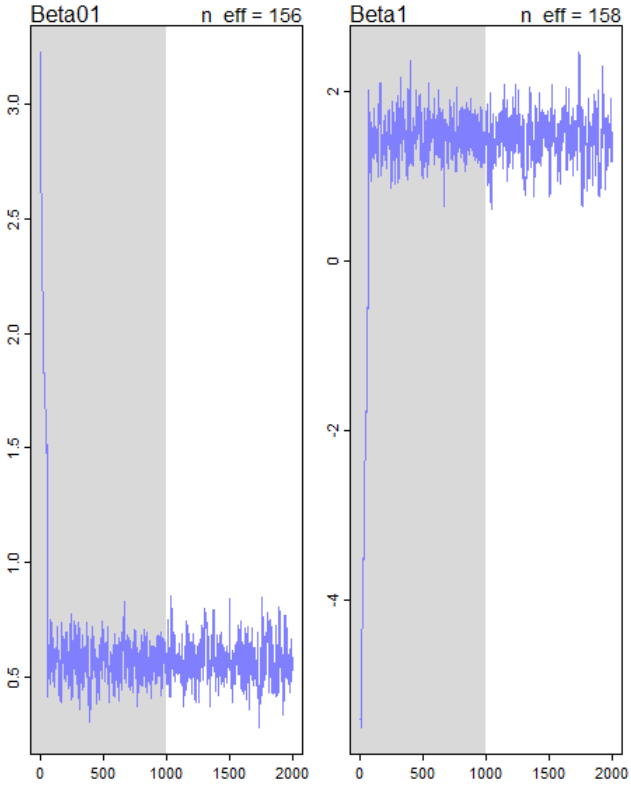

Figure 4.1: Trace plot for the Poisson model (3.5)

Figure 4.1 shows trace plots for the parameters in the Negative binomial model. The trace plot looks stationary, with a distinct zig-zag motion between the samples. The zig-zag motion is a sign that there is no correlation between the samples and the $_{eff}$ value above each plot is greater than 100 which indicates that the chain is healthy.

To determine how many iterations that is needed for the quantiles of the slope parameter to converge, the following plot is visualized below.

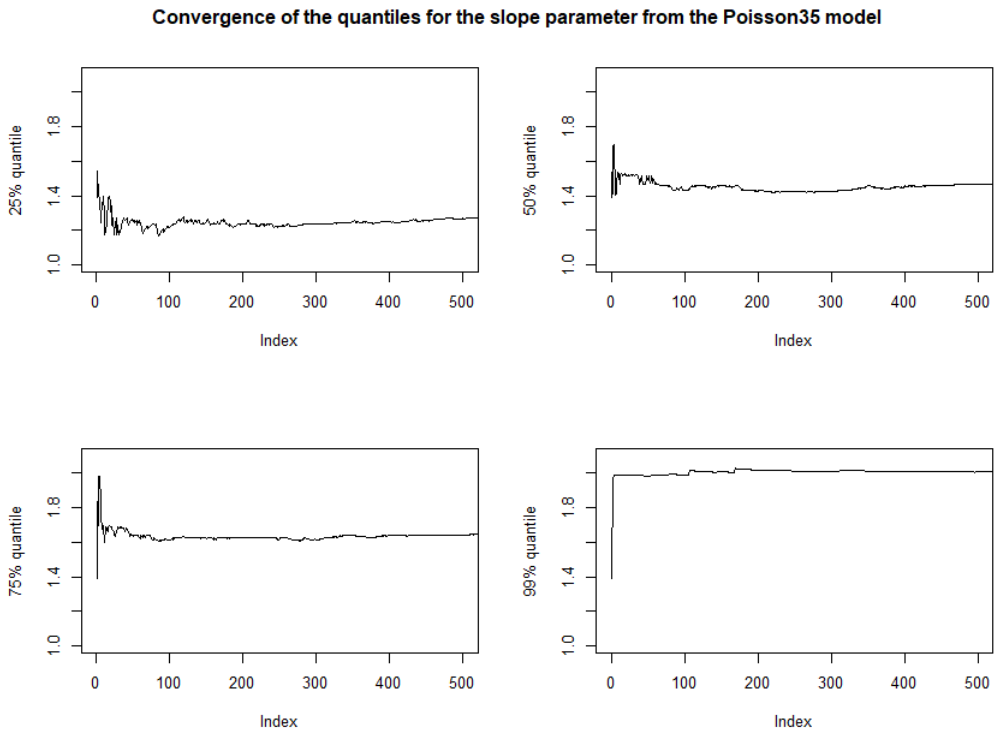

Figure 4.2: Accumulated posterior quantiles of β_1 from the Poisson model

Figure 4.2 illustrates a convergence plot for the accumulated posterior quantiles of the parameter β_1. The figure shows that after approximately 100 iterations, each of the quantiles have converged.

It is also of interest to determine if there exist correlation between each pair of parameters in the model. This is visualized in a pairs plot below.

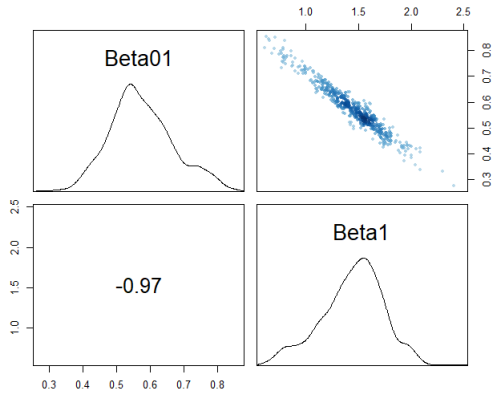

Figure 4.3: Pairs plot for Poisson model with total line 3.5

The diagonal of figure 4.3 presents the marginal estimate densities, which illustrates the distribution of the values of the parameter in the Markov chain. The lower left square shows the correlation coefficient between each pair of parameter. β_{01} is the intercept and β_1 is the slope parameter in the Poisson model. The strong correlation is not a big concern, since HMC can handle it well. [2]

4.2.2 Negative Binomial model with total line 3.5

A summary of the posterior result of the Negative Binomial model with total line 3.5 as predictor variable is presented below.

Table 4.3: Parameter estimation and diagnostics, Negative Binomial model (3.5)

	Mean	StdDev	lower 0.909	upper 0.909	n_{eff}	$\hat{R}$
β_{01}	0.54	0.17	0.22	0.81	501	1.00
β_{02}	6.61	3.26	1.30	12.05	171	1.01
β_1	1.56	0.55	0.65	2.50	505	1.00
β_2	7.96	8.64	-5.46	24.11	462	1.00

Table 4.3 shows that every parameter in the model has a n_{eff} value above 100 and a $\hat{R}$ value below 1.1. This is an indication that the models Markov Chain has converged to an equilibrium distribution. The 90.9 percent highest posterior density interval for the parameter β_2 is the only parameter that is not reliably above zero.

To investigate whether the model has a well adjusted Markov Chain or not, a trace plot is presented below.

[2] The correlation can be reduced by centering the explanatory variable, which has been tried, it gave a lower correlation but a slightly higher WAIC.

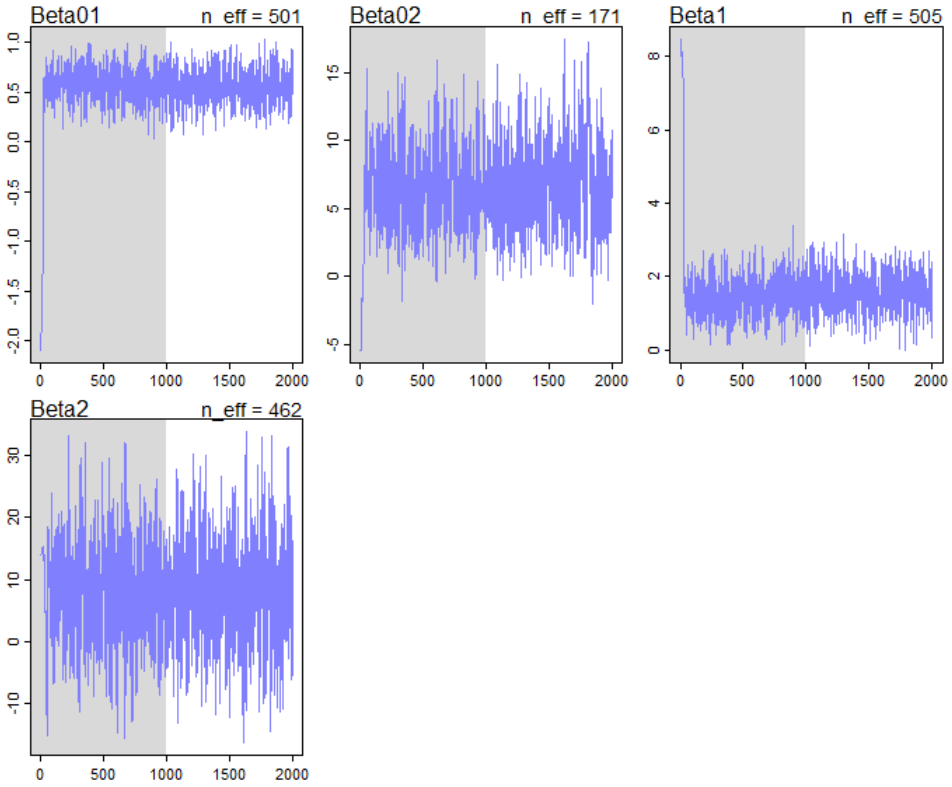

Figure 4.4: Trace plot for the Negative Binomial model

Figure 4.4 shows that the Markov chains appears to be stationary, with a distinct zig-zag motion between the samples. The Markov Chains looks to be healthy since the zig-zag motion is a sign that there are no correlation between the samples. The conclusion that was determined from table 4.3 that the parameter β_2 cannot be reliably over zero, can clearly be seen in figure 4.4 where the draws range from -11 up to 30.

To determine how many iterations that is needed for the quantiles of a the parameters to converge, a similar plot as in figure 4.2 was produced for both β_1 and β_2 which can be seen in the appendix (Figure 6.2). It appeared that each of the quantiles for the two parameters had converged. The correlation between each pair of parameter in the model is visualized in a pair plot which also can be seen in the appendix (Figure 6.3).

4.3 Predictive posterior distributions

This section aims to illustrate the predictive ability of the models from the last section. The models are also retrained on a sub-sample of the data set. These new models are tested on the sub-sample of data it has not seen before, in order to evaluate their predicting ability.

The predictive posterior distribution for the Poisson model and the Negative Binomial model is presented in the histogram below, together with data. The x-axis shows the numbers of goals and the y-axis shows the density.

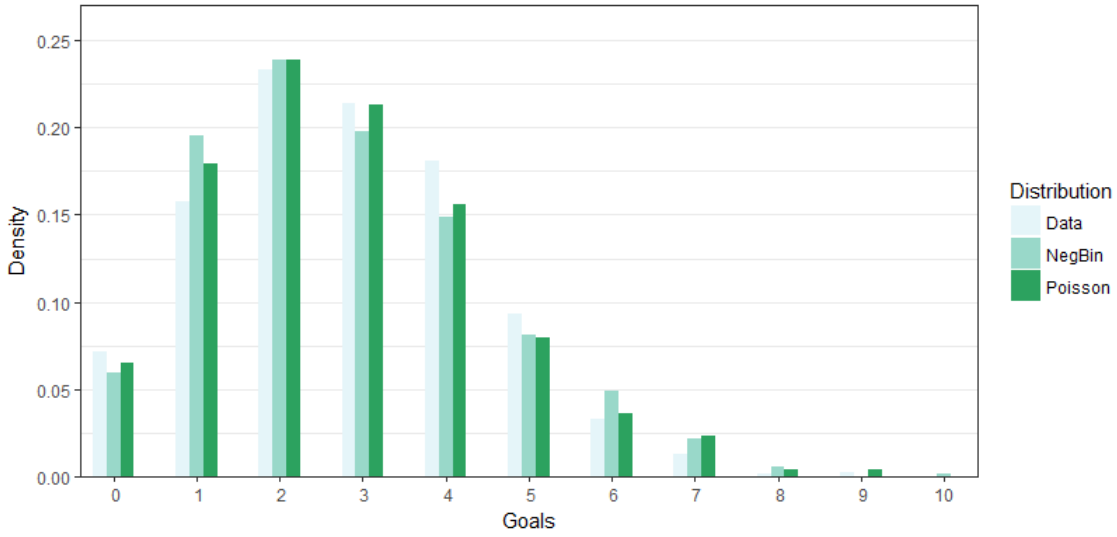

Figure 4.5: Predictive posterior distribution comparisons for models: Poisson35 And NegBin35

Figure 4.5 illustrates that both models posterior fits the data relatively well, but they both are underestimating the probability of 4 goals. Both models are accurate when predicting from 5 up to 9 goals and the predictions of 0 and 1 goals are not far away from data.

To see how well the two models predictive posterior distributions fits new data, the histogram below is presented.

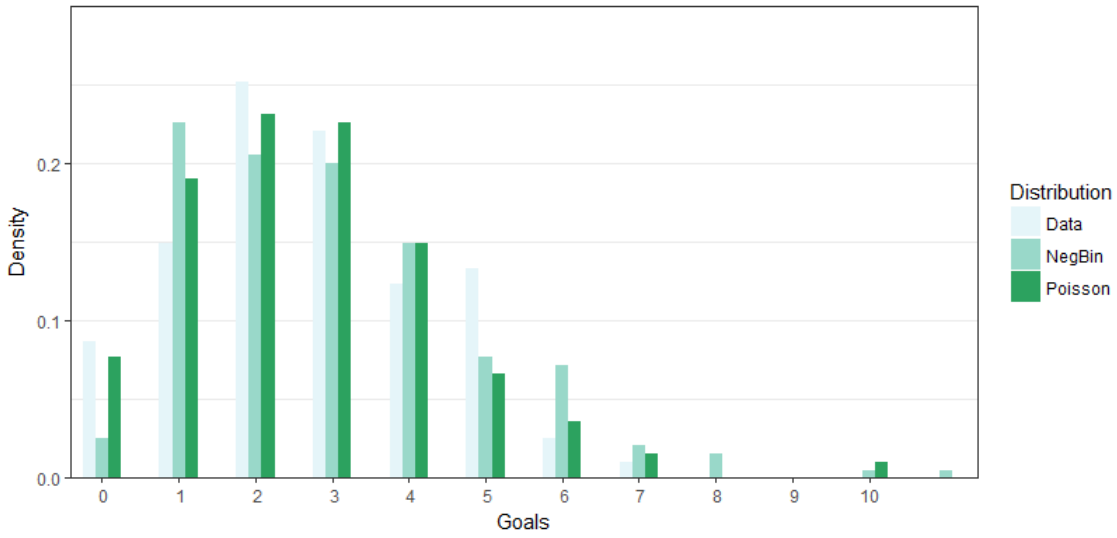

Figure 4.6: Predictive posterior distribution comparisons on new data between models: Poisson35 And NegBin35

Figure 4.6 illustrates the predictive posterior distributions for the Poisson and Negative Binomial Models. The models are trained on data which consists of odds and goals for the Premier League seasons 2015-2016 and 2016-2017. The data in this figure consists of the goal distribution of games played in Premier League season 2017-2018. The figure shows that the Poisson model predicts the new data better than the Negative Binomial model in the goal region of 0 to 3.

5. Discussion

In this chapter we discuss the results, their limitations, application areas, and future studies.

5.1 Limitations

One shortcoming of these results is that not every linear combination of explanatory variables could be tested. Instead, we have chosen to test the combinations we anticipated to be good predictors and not be too correlated. This means that the best model may have been evaded. Also, the data is time-dependent, even if the time dependency can be expected to be reflected in the odds. The thought of including future data in the training sample and test it on previous data does not seem reasonable.

When we produce the predictive posterior distributions, it is done by sampling from the predictive distributions. As a consequence, the results are only based on small samples.

This thesis only uses data from the Premier League which showed signs of underdispersion. The choice of the best predictive model might have been different if we instead included many different leagues since we do not know if Premier League tend to be more underdispersed than other leagues. The majority of leagues might have preferred the Negative Binomial over the Poisson when predicting number of goals.

5.2 Results

The predictive posterior distribution for the Poisson model shows to fit the data better than the Negative Binomial model. We believe that the cause for this is that the variance for the goals scored in the 2015-2018 seasons seems to be lower than the mean. This means that the Negative Binomial distribution should not be used since this distribution allows the variance to be higher than the mean but not lower. We had hoped to create a Negative Binomial model in order to beat the regular Poisson model when predicting the number of goals in soccer, since some of the previous studies indicates that the Poisson distribution should

be substituted by the Negative Binomial distribution. This is because goal-data tends to be overdispersed; however, our data was underdispersed. We believe that we might have been unlucky by choosing seasons where underdispersion existed, which should be reasonably rare. We believe that in general Negative Binomial might still be the best way to go compared to Poisson when predicting soccer goals since most seasons generally tend to be overdispersed.

5.3 Applications of method

We argue that by using odds as an explanatory variable, more variables that may not be incorporated into the odds can be added to the model. Therefore, if the models WAIC improves, the market might underestimate the variables added. We think that this method can also be used to compare different sportsbooks models, in order to find out which one is using the most accurate model for predicting the total number of goals.

5.4 Future work

For future work, we propose using predictors for estimating both home and away goals. And introducing a copula to account for correlation between home and away goals. Furthermore, we suggest using a method that can handle underdispersed data, such as Conway-Maxwell-Poisson regression. It would also be interesting to study more than one league.

6. Conclusion

This chapter summarizes the results in order to answer the research questions that were stated at the beginning of the thesis. The first research question was

- Can the odds be used to create a useful predictive goal distribution?

Answer: Useful predictive goal distributions can be obtained by using the odds. It is enough to know the odds for over 3.5 goals in a soccer game in Premier League to receive useful predictive goal distributions.

The secound reserach question was

- Is negative binomial regression appropriate to model soccer goals in Premier League?

Answer: It is not, negative binomial model has a considerable margin of error for point predicting the number of goals.

Bibliography

[1] *The history of Sports Betting*. URL: http://www.onlinegamblingsites.org/history/sports-betting/ (visited on 05/18/2018).

[2] *A Brief History of Gambling*. URL: https://medium.com/edgefund/a-brief-history-of-gambling-a7f46dbf4403 (visited on 05/18/2018).

[3] Frank Keogh and Gary Rose. *Football betting - the global gambling industry worth billions*. URL: https://www.bbc.com/sport/football/24354124 (visited on 05/18/2018).

[4] Betfair. *Betfair Exchange*. URL: https://www.betfair.com/ (visited on 03/04/2018).

[5] J. Surowiecki. *The Wisdom of Crowds*. Bantam Doubleday Dell Publishing Group Inc, 2011.

[6] Markus Ådahl och Kevin Musasa Kazadi Ehsan Fazlhashemi. "En studie i spelmarknadens riskhantering kring stängningsodds." "page 8". MA thesis. Umeå University, 2016. URL: http://www.diva-portal.org/smash/get/diva2:897973/FULLTEXT01.pdf.

[7] Maher M.J. "Association Football scores." In: (1982). "page 109-118".

[8] Georgi Boshnakov, Tarak Kharrat, and Ian Mchale. "A Bivariate Weibull Count Model for Forecasting Association Football Scores". In: *International Journal of Forecasting* 33.2 (2017), pp. 458–466. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2016.11.006.

[9] Rasmus B. Olesen. "Assessing the number of goals in soccer matches". "resume och page 8,9". MA thesis. Danmark: Ålborg universitet, 2008. URL: http://projekter.aau.dk/projekter/files/14466581/assessingthenumberofgoalsinsopdf.

[10] *Betexplorer*. URL: https://www.betexplorer.com (visited on 01/18/2018).

[11] Schervish M. DeGroot M. *Probability and Statistics*. Pearson Education, Inc, 2012.

[12] *Negative binomial regression*. URL: http://www.karlin.mff.cuni.cz/~pesta/NMFM404/NB.html (visited on 04/18/2018).

[13] Randall Reese. *Poisson versus Negative Binomial Regression*. URL: http://www.math.usu.edu/jrstevens/biostat/PoissonNB.pdf.

[14] R. McElreath. *Statistical rethinking: A bayesian course with examples in R and Stan*. Boca Raton: CRC Press., 2016.

[15] Andreas Svensson Fredrik Lindsten Thomas B. Schön and Niklas Wahlström. URL: http://www.it.uu.se/edu/course/homepage/sml/literature/probabilistic_modeling_compendium.pdf.

[16] *"Mail conversation with Assistant Professor Bertil Wegmann from Dept. of Computer and Information Science, Linköpings University"*.

[17] Sheldon M. Ross. *Introduction to Probability Models*. Academic press, 1997.

[18] *Stan: A Probabilistic Programming Language*. URL: https://www.jstatsoft.org/article/view/v076i01/v76i01.pdf (visited on 05/05/2018).

Appendix

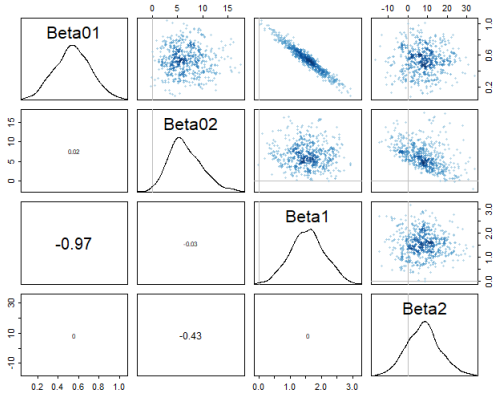

Figure 6.1: Pairs plot for Negative Binomial model with total line 3.5

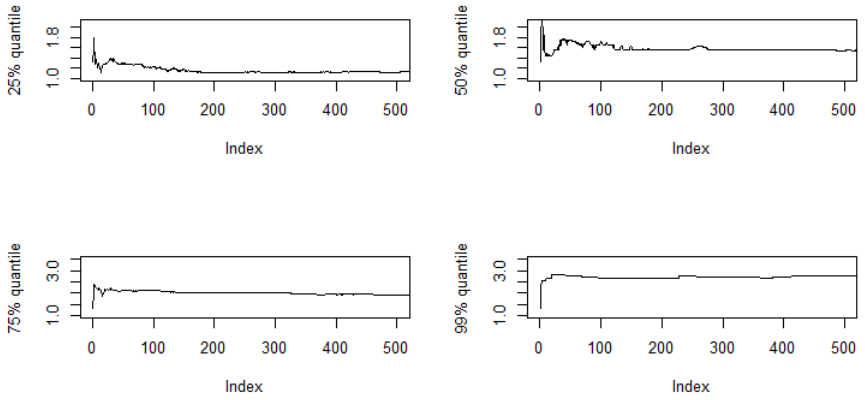

Figure 6.2: Accumulated posterior quantiles of β_1 from the Negative Binomial model

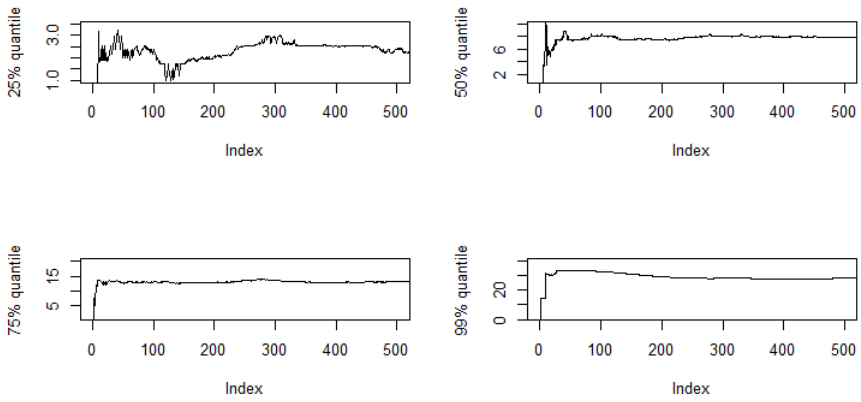

Figure 6.3: Accumulated posterior quantiles of β_2 from the Negative Binomial model