Sports Data Analysis by using Bivariate Poisson Models in the Bayesian Framework

Sabina Shahin

Department of Mathematical Sciences, Karakoram International University, Gilgit, Gilgit-Baltistan, Pakistan
*Corresponding author: sabina@kiu.edu.pk

Abstract

Bivariate distribution models are commonly used to analyze sports data and data from various fields. These models are used to analyze discrete count data with two dependent variables in the data. In this research article, we have used Bivariate Poisson and Diagonally Inflated Bivariate Poisson regression models. We have proposed an estimation procedure in the Bayesian framework in conjunction with the augmentation of data. For parameter estimation, we use Gaussian priors and beta priors for both models. To illustrate the fitting performances of our suggested models we have performed real data analysis on English Premier League soccer data.

Index Terms—Bayesian Analysis; Bivariate Poisson; Regression Models; Inflated Bivariate Poisson; Soccer data analysis

1. Introduction

THE univariate Poisson distribution has been applied as the simplest modeling technique to analyze the competition between two teams in sports. Researchers have found a low correlation when modeling with independent Poisson models, hence, they did not include or discuss such type of correlation while drawing inferences. Several researchers [1], and [2] found a relatively low correlation that can be ignored in modeling since it needs more sophisticated and advance techniques. In team matches the opposite teams interact, for example, in soccer or a football game the performances of the two teams are correlated, which affects and changes the speed of the game and opportunities for gaining more scores. Bivariate models that are extended or derived from the univariate Poisson model are more appropriate choices for dealing with team performances or the paired count data that exhibit correlation. These models are considered suitable options to analyze the data in various fields including medical research, labor mobility, sports, and accident data analysis. There is sparse literature about the applications and implementation of these models because these involve the complicated nature of computation, hence these are less in use by researchers. [3] discussed bivariate Poisson (BP) distribution but did not explain its uses, however, some other researchers discussed BP and its inflated forms in detail. [4] and [5] discussed zero-inflated BP models, whose computation regarding estimation is easy, but deals with excess (inflation) of zeros $(0, 0)$ only, in the data than that of other observations. [6] suggested a new diagonally inflated bivariate Poisson (DIBP) regression model, which is an extension of zero-inflated models. They have discussed the maximum likelihood estimation (MLE) of parameters through the expectation minimization algorithm (EM). [7] suggested bivariate Poisson regression to estimate forecasting models for scored goals and conceded. They also used ordered probit regression to compute forecasting models for in-game results. [8] model animal-vehicle collisions (AVC) data and carcass removal data together by using BP and DIBP models, which helps researchers to use this method for investigating AVCs from a different perspective for road safety. [9] analyzed the Zimbabwe Premier Soccer League and showed that there was not the only advantage of the home ground (HG) but the goals by any two teams'

ISSN: 2523-0379 (Online), ISSN: 1605-8607 (Print)
DOI: https://doi.org/10.52584/QRJ.2101.02

opponents' teams other were also correlated with a relatively low dependence coefficient. In recent years [10] used a Weibull inter-arrival-times-based count process and a copula for forecasting the total goals scored by home and away teams. [11] used BP regression models to predict winning probability and scoring potency. They also have shown that in team matches home and away performances and scores of teams are affected by corner profiles and shots on targets. Although the interpretation of the parameters of the differences between two Poisson and two BP variates is different but the function of the differences between the two BP variates under consideration is the same as the function of the difference between two independent Poisson variates. We have used BP and DIBP models with modifications to generate some simulation results by analyzing English Premier League (EPL) soccer data. We have constructed an efficient MCMC algorithm by using Gibbs sampler for BP and DIBP model for their effective implementations.

2. Models

2.1. Bivariate Poisson regression model

Consider three independent Poisson random variables T_1, T_2, and T_3 with means λ_1, λ_2, and λ_3, respectively. Now, define two new random variables Y_1 and Y_2 as $Y_1 = T_1 + T_3$ and $Y_2 = T_2 + T_3$. Then a random vector (Y_1, Y_2) follows a bivariate Poisson distribution (BP) and denoted by $(Y_1, Y_2) \sim BP(\lambda_1, \lambda_2, \lambda_3)$. The joint probability mass function (pmf) is:

$$p(y_1, y_2; \lambda_1, \lambda_2, \lambda_3) = \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\} \sum_{h=0}^{q} \frac{\lambda_1^{y_1-h} \lambda_2^{y_2-h} \lambda_3^{h}}{(y_1 - h)!(y_2 - h)!h!}, \quad (1)$$

where

$$q = \min(y_1, y_2)$$

with

$$E(Y_1) = \lambda_1 + \lambda_3, \qquad Var(Y_1) = \lambda_1 + \lambda_3,$$
$$E(Y_2) = \lambda_2 + \lambda_3, \qquad Var(Y_2) = \lambda_2 + \lambda_3,$$

and λ_3 is the measure of dependence between Y_1 and Y_2 and hence,

$$Cov(Y_1, Y_2) = \lambda_3.$$

The correlation coefficient ρ, between Y_1 and Y_2 is thereby

$$\rho = \frac{\lambda_3}{\sqrt{(\lambda_1 + \lambda_3)(\lambda_2 + \lambda_3)}}.$$

If $\lambda_3 = 0$, then the two random variables Y_1 and Y_2 become independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions, which is referred to as double Poisson distribution suggested by [12] in 1992.

2.2. Diagonally Inflated Bivariate Poisson regression model

Bivariate Poisson distribution can model data with positive correlation only, and the marginal distributions of BP are also Poisson therefore its marginals cannot handle to model under or over-dispersion in the data. Several BP mixture models either finite or infinite have been suggested by researchers to circumvent these problems but such models need complicated computations making them hard to implement for applications. [6] suggested an inflated BP regression model, that not only allows easy computation regarding estimation but also deals with under and over-dispersion as well as a negative correlation. In the case of diagonal inflated bivariate modeling, a draw or a trail is represented by diagonal terms, hence, adding an inflation term on the diagonal makes modeling more precise when there is a considerable amount of draws. The joint pmf of diagonally inflated bivariate Poisson distribution (DIBP) is

$$\begin{aligned} P_D(y_1, y_2; \lambda_1, \lambda_2, \lambda_3) = {} & I_{(y_1=y_2)}\Big\{(1-\omega)BP((y_1, y_2 \\ & ; \lambda_1, \lambda_2, \lambda_3)) + \omega D(y_1, \nu)\} \\ & + I_{(y_1 \neq y_2)}\{(1-\omega)BP((y_1, \\ & y_2; \lambda_1, \lambda_2, \lambda_3))\Big\}, \end{aligned} \quad (2)$$

or

$$\begin{aligned} P_D(y_1, y_2; \lambda_1, \lambda_2, \lambda_3) = {} & I_{(y_1=y_2)}\Big\{(1-\omega)\exp\{ \\ & -(\lambda_1 + \lambda_2 + \lambda_3)\} \sum_{h=0}^{q} \\ & \frac{\lambda_1^{y_1-h} \lambda_2^{y_2-h} \lambda_3^{h}}{(y_1 - h)!(y_2 - h)!h!} + \\ & \omega D(y_1, \nu)\Big\} + I_{(y_1 \neq y_2)} \\ & \Big\{(1-\omega)\exp\{-(\lambda_1 + \\ & \lambda_2 + \lambda_3)\} \sum_{h=0}^{q} \frac{\lambda_1^{y_1-h}}{(y_1 - h)!} \\ & \frac{\lambda_2^{y_2-h} \lambda_3^{h}}{(y_2 - h)!h!}\Big\}, \end{aligned} \quad (3)$$

where $D(y_1; \nu)$ is a discrete distribution with parameter vector ν that may be chosen from Poisson, Bernoulli, simple discrete distributions or the geometric distribution.

3. Methodology

Suppose we have a sample $(Y_{1i}, Y_{2i}) \overset{\text{indep}}{\sim} BP(\lambda_1, \lambda_2, \lambda_3)$ for $i = 1, \ldots, n$ in order to proceed for our analysis. We use the canonical link function by letting $\lambda_{ki} = \exp\{x'_{ki}\beta_k\}$, where β_k is a $p \times 1$ vector of unknown regression coefficients and x'_{ki} represents the ith row of $n \times p$ design matrix x_k with $p-1$ covariates when the intercept is included in the model for $k = 1, 2, 3$. Then the likelihood function for BP is

$$\begin{aligned} L(\beta_1, \beta_2, \beta_3) \;&=\; \prod_{i=1}^{n} \exp\Big\{-\Big(\exp\{x'_{1i}\beta_1\} + \\ &\exp\{x'_{2i}\beta_2\} + \exp\{x'_{3i}\beta_{3i}\}\Big)\Big\} \\ &\sum_{h=0}^{q_i}\Big(\frac{[\exp\{x'_{1i}\beta_1\}]^{y_{1i}-h}}{(y_{1i}-h)!(y_{2i}-h)!} \times \\ &\frac{[\exp\{x'_{2i}\beta_2\}]^{y_{2i}-h}[\exp\{x'_{3i}\beta_3\}]^{h}}{h!}\Big). \end{aligned} \tag{4}$$

Similarly, the likelihood for DIBP is,

$$\begin{aligned} L(\beta_1, \beta_2, \beta_3, \omega, D) \;&=\; \prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega)\Big(\exp \\ &\Big\{-\Big(\exp\{x'_{1i}\beta_1\} + \exp\{x'_{2i} \\ &\beta_2\} + \exp\{x'_{3i}\beta_3\}\Big)\Big\}\sum_{h=0}^{q_i} \\ &\Big(\frac{[\exp\{x'_{1i}\beta_1\}]^{y_{1i}-h}}{(y_{1i}-h)!} \times \\ &\frac{[\exp\{x'_{2i}\beta_2\}]^{y_{2i}-h}}{(y_{2i}-h)!} \times \\ &\frac{[\exp\{x'_{3i}\beta_3\}]^{h}}{h!}\Big) + \omega D(y_{1i}, \\ &\nu)\Big\} + I_{(y_{1i}\neq y_{2i})}\Big\{(1-\omega) \\ &\Big(\exp\Big\{-\Big(\exp\{x'_{1i}\beta_1\} + \\ &\exp\{x'_{2i}\beta_2\} + \exp\{x'_{3i}\beta_3\} \\ &\Big)\Big\}\sum_{h=0}^{q_i}\frac{[\exp\{x'_{1i}\beta_1\}]^{y_{1i}-h}}{(y_{1i}-h)!} \\ &\times\frac{[\exp\{x'_{2i}\beta_2\}]^{y_{2i}-h}}{(y_{2i}-h)} \times \\ &\frac{[\exp\{x'_{3i}\beta_3\}]^{h}}{!h!}\Big)\Big\}\Big], \end{aligned} \tag{5}$$

where $q_i = \min(y_{1i}, y_{2i})$ for $i = 1, \ldots, n$. Now log-likelihoods for BP and DIBP are

$$\begin{aligned} logL(\beta_1, \beta_2, \beta_3) \;&=\; -\sum_{i=1}^{n}\exp\{x'_{1i}\beta_1\} - \sum_{i=1}^{n}\exp\{x'_{2i} \\ &\beta_2\} - \sum_{i=1}^{n}\exp\{x'_{3i}\beta_3\} + \sum_{i=1}^{n} log \\ &\Big[\sum_{h=0}^{q}\frac{[\exp\{x'_{1i}\beta_1\}]^{y_1-h}}{(y_1-h)!} \times \\ &\frac{[\exp\{x'_{2i}\beta_2\}]^{y_2-h}\exp\{x'_{3i}\beta_3\}]^{h}}{(y_2-h)!h!}\Big], \end{aligned} \tag{6}$$

and

$$\begin{aligned} logL(\beta_1, \beta_2, \beta_3, \omega, D) \;&=\; \sum_{i=1}^{n} log\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega) \\ &\Big(\exp\Big\{-\Big(\exp\{x'_{1i}\beta_1\} + \\ &\exp\{x'_{2i}\beta_2\} + \exp\{x'_{3i}\beta_3\} \\ &\Big)\Big\}\sum_{h=0}^{q}\frac{[\exp\{x'_{1i}\beta_1\}]^{y_1-h}}{(y_1-h)!} \\ &\times\frac{[\exp\{x'_{2i}\beta_2\}]^{y_2-h}}{(y_2-h)!} \times \\ &\frac{[\exp\{x'_{3i}\beta_3\}]^{h}}{h!}\Big) + \omega D(y_1, \\ &\nu)\Big\} + I_{(y_{1i}\neq y_{2i})}\Big\{(1-\omega) \\ &\Big(\exp\Big\{-\Big(\exp\{x'_{1i}\beta_1\} + \\ &\exp\{x'_{2i}\beta_2\} + \exp\{x'_{3i}\beta_3\} \\ &\Big)\Big\}\sum_{h=0}^{q}\frac{[\exp\{x'_{1i}\beta_1\}]^{y_1-h}}{(y_1-h)!} \\ &\times\frac{[\exp\{x'_{2i}\beta_2\}]^{y_2-h}}{(y_2-h)!} \times \\ &\frac{[\exp\{x'_{3i}\beta_3\}]^{h}}{h!}\Big)\Big\}\Big]. \end{aligned} \tag{7}$$

3.1. Bayes Estimation

For the regression coefficients, we use vague or diffuse proper priors with large variance $\beta_k = (\beta_{k1}, \ldots, \beta_{kp})$ for $k = 1, 2, 3$. Specifically, these priors have been formed with an extreme kind of spread such as a Gaussian density but with an extraordinarily large variance. Diffuse or vague and weak informative priors have been excessively used in research articles [13] - [15]. Therefore, we also assume that the priors on β_{kl} are normally distributed with means μ_{1l}, μ_{2l}, and μ_{3l}, and variances σ^2_{1l}, σ^2_{2l} and σ^2_{3l} respectively, for $l = 1, \ldots, p$.

3.2. Estimation procedures

Estimation of the proposed models will be done in the Bayesian framework in conjunction with the augmentation of data. To estimate the parameters, the data augmentation algorithm can be used with a sampling-based Gibbs sampler. Gibbs sampler is proposed by [16] and is widely used as a common method in Bayesian computation [17] - [18]. We use Gaussian priors β_k's with mean μ_k and variance σ_k respectively for $k = 1, 2, 3$ for both BP and DIBP models. We use beta priors for ω with α and γ, for DIBP, to make theoretical results or derivations less complicated. For the same purpose of estimation of these parameters, we can also use some other conjugate priors like Inverse Gamma as well as non-informative priors. Here the posteriors for BP and DIBP are,

$$\begin{aligned}\pi(\beta_1,\beta_2,\beta_3) \;=\; & \frac{\exp\{-\frac{1}{2}(\frac{\beta_1-\mu_1}{\sigma_1})^2\}}{\sqrt{2\pi}\sigma_1}\frac{\exp\{-\frac{1}{2}(\frac{\beta_2-\mu_2}{\sigma_2})^2\}}{\sqrt{2\pi}\sigma_2} \\ & \frac{\exp\{-\frac{1}{2}(\frac{\beta_3-\mu_1}{\sigma_3})^2\}}{\sqrt{2\pi}\sigma_3}\prod_{i=1}^{n}\Big(\exp\Big\{-\Big(\exp \\ & \{x_i'\beta_1\}+\exp\{x_i'\beta_2\}+\exp\{x_i'\beta_3\}\Big)\Big\} \\ & \sum_{h=0}^{q}\frac{[\exp\{x_i'\beta_1\}]^{y_{1i}-h}[\exp\{x_i'\beta_2\}]^{y_{2i}-h}}{(y_{1i}-h)!(y_{2i}-h)!} \\ & \times\frac{[\exp\{x_i'\beta_3\}]^h}{h!}\Big), \end{aligned} \tag{8}$$

$$\begin{aligned}\pi(\beta_1,\beta_2,\beta_3,\omega,D) \;=\; & \omega^{\alpha-1}(1-\omega)^{\gamma-1}\times \\ & \frac{\exp\{-\frac{1}{2}(\frac{\beta_1-\mu_1}{\sigma_1})^2\}}{\sqrt{2\pi}\sigma_1}\times \\ & \frac{\exp\{-\frac{1}{2}(\frac{\beta_2-\mu_2}{\sigma_2})^2\}}{\sqrt{2\pi}\sigma_2}\times \\ & \frac{\exp\{-\frac{1}{2}(\frac{\beta_3-\mu_1}{\sigma_3})^2\}}{\sqrt{2\pi}\sigma_3}\times \\ & \prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega)\Big(\exp \\ & \Big\{-(\exp\{x_i'\beta_1\}+\exp\{x_i'\beta_2\}+ \\ & \exp\{x_i'\beta_3\})\Big\}\sum_{h=0}^{q}\frac{[\exp\{x_i'\beta_1\}]^{y_{1i}-h}}{(y_{1i}-h)!} \\ & \frac{[\exp\{x_i'\beta_2\}]^{y_{2i}-h}}{(y_{2i}-h)!}\frac{[\exp\{x_i'\beta_3\}]^{h}}{h!}\Big) \\ & +\omega D(y_{1i},\nu)\Big\}+I_{(y_{1i}\neq y_{2i})}\Big\{(1- \\ & \omega)\Big(\exp\Big\{-(\exp\{x_i'\beta_1\}+\exp\{x_i' \\ & \beta_2\}+\exp\{x_i'\beta_3\})\Big\}\sum_{h=0}^{q} \\ & \frac{[\exp\{x_i'\beta_1\}]^{y_{1i}-h}[\exp\{x_i'\beta_2\}]^{y_{2i}-h}}{(y_{1i}-h)!(y_{2i}-h)!} \\ & \times\frac{[\exp\{x_i'\beta_3\}]^{h}}{h!}\Big)\Big\}\Big]. \end{aligned} \tag{9}$$

The full conditionals for β_k in the case of the BP model are

$$\begin{aligned}\pi(\beta_{1l};\beta_{-1l},\beta_2,\beta_3) \;\propto\; & \exp\{-\frac{1}{2}(\frac{\beta_{1l}-\mu_1}{\sigma_k})^2\}\prod_{i=1}^{n} \\ & \exp\{-\exp\{x_i'\beta_{1l}\}\}[\exp\{x_i'\beta_{1l}\}]^{y_{1i}} \\ & \sum_{h=0}^{q}\Big(\frac{\exp\{x_i'\beta_3\}}{\exp\{x_i'\beta_{1l}\}\exp\{x_i'\beta_2\}}\Big)^h \\ & \times\frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}, \end{aligned} \tag{10}$$

and for β_1, β_2 and β_3 that are equivalent to the following,

$$\begin{aligned}\pi(\beta_{1l};\beta_{-1l},\beta_2,\beta_3) \;\propto\; & \exp\{-\frac{1}{2}(\frac{\beta_{1l}-\mu_1}{\sigma_1})^2\}\prod_{i=1}^{n}e^{-\lambda_{1i}} \\ & \lambda_{1i}^{y_{1i}}\sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}, \end{aligned} \tag{11}$$

$$\begin{aligned}\pi(\beta_{2l};\beta_{-2l},\beta_1,\beta_3) \quad \propto \quad & \exp\{-\frac{1}{2}(\frac{\beta_{2l}-\mu_2}{\sigma_2})^2\}\prod_{i=1}^{n} \\ & e^{-\lambda_{2i}}\lambda_{2i}^{y_{2i}}\sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}, \quad (12)\end{aligned}$$

$$\begin{aligned}\pi(\beta_{3l};\beta_{-3l},\beta_1,\beta_2) \quad \propto \quad & \exp\{-\frac{1}{2}(\frac{\beta_{3l}-\mu_3}{\sigma_3})^2\}\prod_{i=1}^{n} \\ & e^{-\lambda_{3i}}\sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}, \quad (13)\end{aligned}$$

and for λ_3 full conditional is

$$\begin{aligned}\pi(\lambda_3;\beta_1,\beta_2) \quad \propto \quad & \lambda_3^{\delta-1}\exp-\epsilon\lambda_3\prod_{i=1}^{n} \\ & e^{-\lambda_{3i}}\sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}. \quad (14)\end{aligned}$$

Now the full conditionals for β_k in the case DIBP model are

$$\begin{aligned}\pi(\beta_{1l};\beta_{-1l},\beta_2,\beta_3,\omega,D) \quad \propto \quad & \exp\Big\{-\frac{1}{2}(\frac{(\beta_{1l}-\mu_1)^2}{\sigma_1})^2\Big\} \\ & \prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega) \\ & \Big(\exp\Big\{-\Big(\exp\{x_i'\beta_1\}+ \\ & \exp\{x_i'\beta_2\}+\exp\{x_i'\beta_3\} \\ & \Big)\Big\}[\exp\{x_i'\beta_{1l}\}]^{y_{1i}}\times \\ & [\exp\{x_i'\beta_2\}]^{y_{2i}}\times \\ & \sum_{h=0}^{q}\Big(\frac{\exp\{x_i'\beta_3\}}{\exp\{x_i'\beta_{1l}\}\exp\{x_i'\beta_2\}}\Big)^h \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big)+ \\ & \omega D(y_{1i},\nu)\Big\}+I_{(y_{1i}\neq y_{2i})} \\ & \Big\{(1-\omega)\Big(\exp\Big\{-\Big(\exp \\ & \{x_i'\beta_1\}+\exp\{x_i'\beta_2\}+ \\ & \exp\{x_i'\beta_3\}\Big)\Big\}[\exp\{x_i'\beta_{1l}\}]^{y_{1i}} \\ & [\exp\{x_i'\beta_2\}]^{y_{2i}}\sum_{h=0}^{q} \\ & \Big(\frac{\exp\{x_i'\beta_3\}}{\exp\{x_i'\beta_{1l}\}\exp\{x_i'\beta_2\}}\Big)^h \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big)\Big\}\Big], \\ & (15)\end{aligned}$$

and for β_1, β_2, β_3 that are equivalent to the following,

$$\pi(\beta_{1l}; \beta_{-1l}, \beta_2, \beta_3, \omega, D) \propto \exp\Big\{-\frac{1}{2}(\frac{(\beta_{1l}-\mu_1}{\sigma_1})^2\Big\} \prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega) \Big(\exp\Big\{-(\lambda_{1i}+\lambda_{2i} +\lambda_{3i})\Big\}\times\lambda_{1i}^{y_{1i}}\times\lambda_{2i}^{y_{2i}} \sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big) +\omega D(y_{1i},\nu)\Big\}+I_{(y_{1i}\neq y_{2i})} \Big\{(1-\omega)\Big(\exp\Big\{-(\lambda_{1i}+ \lambda_{2i}+\lambda_{3i})\Big\}\lambda_{1i}^{y_{1i}}\times\lambda_{2i}^{y_{2i}} \sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big)\Big], \tag{16}$$

$$\pi(\beta_{2l}; \beta_{-2l}, \beta_1, \beta_3, \omega, D) \propto \exp\Big\{-\frac{1}{2}(\frac{(\beta_{2l}-\mu_2}{\sigma_2})^2\Big\} \prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega) \Big(\exp\Big\{-(\lambda_{1i}+\lambda_{2i}+ \lambda_{3i})\Big\}\lambda_{1i}^{y_{1i}}\lambda_{2i}^{y_{2i}}\sum_{h=0}^{q} \Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big) +\omega D(y_{1i},\nu)\Big\}+I_{(y_{1i}\neq y_{2i})} \Big\{(1-\omega)\Big(\exp\Big\{-(\lambda_{1i}+ \lambda_{2i}+\lambda_{3i})\Big\}\lambda_{1i}^{y_{1i}}\lambda_{2i}^{y_{2i}} \sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!} \Big)\Big\}\Big] \tag{17}$$

and

$$\pi(\beta_{3l}; \beta_{-3l}, \beta_1, \beta_2, \omega, D) \propto \exp\Big\{-\frac{1}{2}(\frac{(\beta_{3l}-\mu_3}{\sigma_3})^2\Big\} \prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega) \Big(\exp\Big\{-(\lambda_{1i}+\lambda_{2i}+ \lambda_{3i})\Big\}\lambda_{1i}^{y_{1i}}\lambda_{2i}^{y_{2i}}\sum_{h=0}^{q} \Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\frac{1}{(y_{1i}-h)!} \frac{1}{(y_{2i}-h)!h!}\Big)+\omega D(y_{1i}, \nu)\Big\}+I_{(y_{1i}\neq y_{2i})}\Big\{(1-\omega) \Big(\exp\Big\{-(\lambda_{1i}+\lambda_{2i}+ \lambda_{3i})\Big\}\lambda_{1i}^{y_{1i}}\lambda_{2i}^{y_{2i}}\sum_{h=0}^{q} \Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big)\Big\}\Big], \tag{18}$$

where $\beta_{\mathbf{-kl}} = \{\beta_{\mathbf{k1}}, \beta_{\mathbf{k2}}, ..., \beta_{\mathbf{kl-1}}, \beta_{\mathbf{kl+1}}, ..., \beta_{\mathbf{kp}}\}$ or β_{-kl} is a $(p-1)\times 1$ vector with l^{th} component being excluded from β_k. For ω, the full conditional is

$$\pi(\omega; \beta_1, \beta_2, \beta_3, D) \propto \omega^{\alpha-1}(1-\omega)^{\gamma-1} \prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})}\Big\{(1-\omega) \Big(\exp\Big\{-(\lambda_{1i}+\lambda_{2i} +\lambda_{3i})\Big\}\times\lambda_{1i}^{y_{1i}}\times\lambda_{2i}^{y_{2i}} \sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big)+ \omega D(y_{1i},\nu)\Big\}+I_{(y_{1i}\neq y_{2i})} \Big\{(1-\omega)\Big(\exp\Big\{-(\lambda_{1i}+\lambda_{2i} +\lambda_{3i})\Big\}\times\lambda_{1i}^{y_{1i}}\times\lambda_{2i}^{y_{2i}}\times \sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^h\times \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big)\Big\}\Big]. \tag{19}$$

Finally the full conditional for λ_3 is

$$\begin{aligned}\pi(\lambda_3;\beta_1,\beta_2,\omega,D) \quad \propto \quad & \lambda_3^{\delta-1}e^{-\epsilon\lambda_3}\prod_{i=1}^{n}\Big[I_{(y_{1i}=y_{2i})} \\ & \Big\{(1-\omega) \\ & \Big(\exp\Big\{-(\lambda_{1i}+\lambda_{2i}+\lambda_{3i})\Big\} \\ & \lambda_{1i}^{y_{1i}}\lambda_{2i}^{y_{2i}}\sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^{h}\times \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big) \\ & +\omega D(y_{1i},\nu)\Big\}+I_{(y_{1i}\neq y_{2i})} \\ & \Big\{(1-\omega)\Big(\exp\Big\{-(\lambda_{1i}+ \\ & \lambda_{2i}+\lambda_{3i})\Big\}\times\lambda_{1i}^{y_{1i}}\times\lambda_{2i}^{y_{2i}} \\ & \sum_{h=0}^{q}\Big(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\Big)^{h}\times \\ & \frac{1}{(y_{1i}-h)!(y_{2i}-h)!h!}\Big)\Big\}\Big]. \end{aligned} \quad (20)$$

4. Results and discussions

We conduct real data analysis by using BP and DIBP models to examine their performances regarding parameters. We analyze the data for English Premier League (EPL) soccer data for the season 2015-2016. We have taken the soccer data from www.espn.com. The data set consisted of two independent variables, effective shots, and fouls. Whereas the dependent variables are the number of goals at home y_1 and the number of goals away y_2. In this analysis, we ran the Gibbs sampler for 11,000 iterations and we compute the posterior means and 95% Highest Posterior Density (HPD) intervals for the parameters after the initial 1,000 iterates are discarded as a burn-in, the computed statistical values are reported in Table **??** and Table **??**. We noted that the signs of the estimates are congruent with what we would expect from the analysis perspectives of the performances of soccer teams. In Table **??** we can observe that, as the effective shots β_{12} and β_{22} increase the chances of goals across home y_1 and away y_2 also increases. There is a negative sign with the estimated value of parameters for fouls β_{13} and β_{23} i.e. natural because the increase in fouls has an inverse effect on the number of goals. However, if we compare the estimated values of the parameter between BP and DIBP the BP overestimates β_{22} for away home i.e. it gives $\beta_{12} < \beta_{22}$ but the expected values should show a reverse relation i.e. $\beta_{22} \leq \beta_{12}$ because teams performed well at home than away. The expected values for fouls for the home should be lower in the home than away but the estimated values of fouls for BP are $\beta_{23} \leq \beta_{13}$ that misleads interpretations. However, the estimated values in the case of DIBP for both parameters are $\beta_{12} > \beta_{22}$ and $\beta_{13} < \beta_{23}$ as per expectations. Hence on the basis of estimated values of DIBP, we can conclude that teams performed well in the first half of the season 2015-2016 EPL. Table **??** represents statistics for the second half of the season 2015-2016. We observe the same type of estimated values as we observed in Table**??** i.e. as the effective shots (β_{12} and β_{22}) increase so does the chances of goals across home y_1 and away y_2. But we can observe a negative sign with the estimated value of the parameter for fouls β_{13} in home goals case y_1 but a positive value observed for β_{23} that can be misleading to the interpretations. BP results again overestimate the relation by showing that performance in the case of away goals and effective shots is stronger than at home. However, the estimated values for both cases home and away DIBP models performed well by showing some natural outcomes, i.e. a negative value of fouls parameter showing that an increase in fouls has a negative effect on the number of goals, whereas positive values β_{12} and β_{22}. If we look at the estimated values of effective shots parameters β_{12} and β_{22} in the case of DIBP both parameters fulfill the expected condition, i.e. $\beta_{12} > \beta_{22}$ and $\beta_{13} < \beta_{23}$ as per expectations. Hence on the basis of estimated values of DIBP, we conclude that teams performed well in the second half of the season 2015-2016 too and teams performed well at home-ground than away.

5. Conclusion

In this research work, we have focused on the comparison of two bivariate Poisson regression models i.e. bivariate Poisson and Diagonally Inflated bivariate Poisson. We analyze both methods to compute efficient parameters by applying them to English Premier League football data for the season 2015-16. We observed statistics that both models performed well but DIBP is a better option to analyze inflated data. DIBP successfully deals with correlated data and we can avoid under or over-estimation of statistical values. We can apply these models in other fields e.g. medical and health, risk analysis research, geological surveys, investment analysis, etc to deal with the bivariate data.

TABLE 1: Parameter estimates based on EPL data

		BP		DIBP	
Dependent variable	Parameter	Post-mean	95% HPD	Post-mean	95% HPD
y_1	β_{11}	-0.2178	(-0.6877, 0.2924)	-0.0108	(-0.1978, 0.1939)
	β_{12}	0.1428	(0.1039, 0.1787)	0.0347	(-0.1529, 0.2312)
	β_{13}	-0.0104	(-0.0512, 0.0314)	-0.0536	(-0.2160, 0.0781)
	λ_3	0.0121	(0.0002, 0.0745)	0.0018	(0.0001, 0.0086)
	ω			0.1093	(0.0026, 0.3544)
y_2	β_{21}	-0.8140	(-1.4036, -0.2659)	-0.0108	(-0.1988, 0.1907)
	β_{22}	0.2421	(0.1877, 0.2973)	0.0344	(-0.1553, 0.2306)
	β_{23}	-0.0009	(-0.0410, 0.0435)	-0.0540	(-0.2173, 0.0783)
	λ_3	0.0944	(0.0105, 0.1975)	0.0010	(0.0000, 0.0033)
	ω			0.1108	(0.0031, 0.3685)

TABLE 2: Parameter estimates based on EPL data

		BP		DIBP	
Dependent variable	Parameter	Post-mean	95% HPD	Post-mean	95% HPD
y_1	β_{11}	-0.7589	(-1.5153,-0.1083)	-0.0286	(-0.3038,0.3182)
	β_{12}	0.2033	(0.1379,0.2607)	0.0125	(-0.2290,0.2901)
	β_{13}	-0.0071	(-0.0493,0.0393)	-0.0525	(-0.2574,0.1956)
	λ_3	0.1460	(0.0008,0.3450)	0.0010	(0.0000,0.0039)
	ω			0.1093	(0.0005,0.4824)
y_2	β_{21}	-1.2409	(-2.2226,-0.1445)	-0.0285	(-0.3073,0.3193)
	β_{22}	0.2208	(0.1449,0.3039)	0.0122	(-0.2289,0.2893)
	β_{23}	0.0185	(-0.0453,0.0721)	-0.0530	(-0.2592,0.1922)
	λ_3	0.1868	(0.0041,0.4006)	0.0023	(0.0001,0.0202)
	ω			0.1108	(0.0007,0.4815)

References

1. A. J. Lee, ”Modeling scores in the Premier League: is Manchester United really the best?,” Chance, vol. 10, no. 1, pp. 15-19, 1997.
2. D. Karlis and I. Ntzoufras, ”On modeling soccer data,” Student, vol. 3, no. 4, pp. 229-244, 2000.
3. M. J. Maher, ”Modelling association football scores,” Statistica Neerlandica, vol. 36, no. 3, pp. 109-118, 1982.
4. N. Gan, ”General zero-inflated models and their applications,” Ph.D. dissertation, North Carolina State University, 2000.
5. J. F. Walhin, ”Bivariate ZIP models,” Biometrical Journal, vol. 43, no. 2, pp. 147-160, 2001.
6. D. Karlis and I. Ntzoufras, ”Analysis of sports data by using bivariate Poisson models,” Journal of the Royal Statistical Society: Series D (The Statistician), vol. 52, no. 3, pp. 381-393, 2003.
7. J. Goddard, ”Regression models for forecasting goals and match results in association football,” International Journal of Forecasting, vol. 21, no. 2, pp. 331-340, 2005.
8. Y. Lao, G. Zhang, Y. Wu, and Y. Wang, ”Modeling animal–vehicle collisions considering animal–vehicle interactions,” Accident Analysis and Prevention, vol. 43, no. 6, pp. 1991-1998, 2011.
9. D. Mwembe, L. Sibanda, and N. C. Mupondo, ”Application of a bivariate Poisson model in devising a profitable betting strategy of the Zimbabwe premier soccer league match results,” American Journal of Theoretical and Applied Statistics, vol. 4, no. 3, pp. 99-111, 2015.
10. G. Boshnakov, T. Kharrat, and I. G. McHale, ”A bivariate Weibull count model for forecasting association football scores,” Int. J. Forecast., vol. 33, no. 2, pp. 458-466, 2017.
11. R. K. Ankomah, E. K. Amoah, and E. A. Obeng, ”Predictive Modeling of Association Football Scores Using Bivariate Poisson,” 2020.
12. S. Kocherlakota and K. Kocherlakota, Bivariate Discrete Distributions. CRC Press, 2017.
13. A. FM Smith and D. J. Spiegelhalter, ”Bayes factors and choice criteria for linear models,” J. R. Stat. Soc. B, vol. 42, no. 2, pp. 213-220, 1980.
14. D. J. Spiegelhalter and A. FM Smith, ”Bayes factors for linear and log-linear models with vague prior information,” J. R. Stat. Soc. B, vol. 44, no. 3, pp. 377-387, 1982.

15. V. E. Akman and A. E. Raftery, ”Bayes factors for non-homogeneous Poisson processes with vague prior information,” J. R. Stat. Soc. B, vol. 48, no. 3, pp. 322-329, 1986.
16. S. Geman and D. Geman, ”Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 6, pp. 721-741, 1984.
17. A. E. Gelfand and A. FM Smith, ”Sampling-based approaches to calculating marginal densities,” J. Amer. Stat. Assoc., vol. 85, no. 410, pp. 398-409, 1990.
18. S. L. Zeger and M. R. Karim, ”Generalized linear models with random effects: A Gibbs sampling approach,” J. Amer. Stat. Assoc., vol. 86, no. 413, pp. 79-86, 1991.