Modeling Football Match Scoring Outcomes using Multilevel Models

Liberato Camilleri, Naomi Farrugia
Department of Statistics and Operations Research
University of Malta
Msida (MSD 06) Malta
E-mail: liberato.camilleri@um.edu.mt

KEYWORDS
Hierarchical nested data, random coefficient model, intra class correlation, multilevel model.

ABSTRACT

Multilevel modelling technique recognizes the existence of hierarchal structures in the data by allowing for random effects at each level in the hierarchy, thus assessing the variation in the dependent variable at several hierarchical levels simultaneously. Multilevel modelling is becoming an increasingly popular technique for analysing nested data with such popularity accredited to the computational advances in the last two decades. In many sports, including football, the game fixtures are nested within seasons, which in turn are nested within country leagues invoking a multilevel structure in the data. Many gaming companies engage in sport data analysis in a bid to understand the dynamics and patterns of the game. This will assist the gaming company in developing fantasy sport games that will enhance gamer engagement and augment revenue to the company.

This paper presents a comprehensive description of two and three level models, which are applied to a real football data set accessed from an online free football betting portal. The aim is to examine the relationship between the number of goals scored during a football match and several game-related predictors. These multilevel models, which assume a Poisson distribution and a logarithmic function, are implemented using the facilities of GLLAMM (Generalized Linear Latent and Mixed Models), which is a subroutine of STATA.

1. Introduction

The concept of Generalized Linear Models (GLMs) was first introduced by Nelder and Wedderburn (1972) where several widely used distributions, including the Normal, Poisson, Binomial, Gamma, Geometric, Multinomial and Inverse Gaussian distribution were combined together as members of the exponential family. The iteratively reweighted least squares algorithm was used for maximum likelihood estimation. A fundamental assumption of GLMs is that the responses are independent making these models inappropriate for longitudinal data, repeated measures and multilevel data with a nesting structure. To overcome this limitation, Liang and Zeger (1986) developed the concept of Generalized Estimating Equations (GEE) by removing the independence assumption. This development gave rise to GEE models that accommodate highly correlated data by specifying a structure for the working correlation matrix. To accommodate nested hierarchical structured data, Bryk and Raudenbush and (1992) introduced the concept of multilevel models. In contrast with the GLM and the GEE, these models take into consideration the hierarchical nature of the nested data by accommodating the error term and random effects at each hierarchical level of nesting. The development of software packages and the introduction of supercomputers alleviated the implementation of multilevel models to large data sets, particularly when the hierarchical structure exceeds two levels of nesting and the number of random effects is considerable.

2. Theory

In a generalized linear model framework, the expected value of the response $\mathbb{E}(y_{ij}) = \mu_{ij}$ is related to linear predictor η_{ij} through a non-linear invertible link function $g(\cdot)$ given by:

$$\mathbb{E}(y_{ij}) = g(\eta_{ij})$$

In this generalized linear model, the response mechanism is fully described by the conditional probability density function of the response y_i given the linear predictor η_{ij}. The model is completed by specifying a distribution for the observed response $y_{ij}|\mu_{ij}$, which in the case of count data is the Poisson distribution with parameter μ_{ij}.

$$\mathbb{P}(y_{ij}|\mu_{ij}) = \frac{{\mu_{ij}}^{y_{ij}} e^{-\mu_{ij}}}{(y_{ij})!}$$

where

$$\mathbb{E}(y_{ij}) = Var(y_{ij}) = \mu_{ij}$$

The link function $g(\cdot)$ for count data is the logarithm link specified in the following way:

$$g^{-1}(\mu_{ij}) = \log \mu_{ij} = \eta_{ij}$$

A Poisson model assumes that the duration of the observation period is fixed in advance (constant exposure); however, this is not always the case. The model can be extended further by including a varying exposure rate m_{ij}. As a result the Poisson regression model can be written in the form:

$$\log \mu_{ij} = \log m_{ij} + \beta_{0j} + \beta_{1j} x_{1ij} + \cdots + \beta_{pj} x_{pij}$$

This implies that the relationship between μ_i and the linear predictor η_i is offset by the amount $\log m_{ij}$. This term is a fixed part offset and if required, it is centred on the mean so as to avoid numerical instabilities. Yet, we do not always require an offset, or where the offset is a constant. A two-level random intercept model with one explanatory variable x_{1ij} can be provided for count data and is given by:

$$\log \mu_{ij} = \log m_{ij} + \beta_{0j} + \beta_{1j} x_{1ij} + U_{0j}$$

An extended two-level random intercept model with several explanatory variables is given by:

$$\log \mu_{ij} = \log m_{ij} + \boldsymbol{x}'\boldsymbol{\beta} + U_{0j}$$

Similarly, the two-level random coefficient, 1-predictor model for count data is given by:

$$\log \mu_{ij} = \log m_{ij} + \beta_{0j} + \beta_{1j} x_{1ij} + U_{0j} + U_{1j} x_{1ij}$$

More generally, we have that:

$$\log \mu_{ij} = \log m_{ij} + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{z}'\boldsymbol{U}_j$$

$\boldsymbol{U}_j$ follows a multivariate normal distribution $\boldsymbol{U}_j \sim \mathcal{N}(\mathbf{0}, \mathbb{T}_j)$ respectively with:

$$\mathbb{T}_j = \begin{bmatrix} var(U_{0j}) & cov(U_{0j}, U_{1j}) \\ cov(U_{1j}, U_{0j}) & var(U_{1j}) \end{bmatrix} = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{bmatrix}$$

A three-level random intercept model with one explanatory variable x_{1ijk} can be provided for count data and is given by:

$$\log \mu_{ijk} = \log m_{ijk} + \beta_{0jk} + \beta_{1jk} x_{1ijk} + U_{0jk} + V_{00k}$$

An extended three-level random intercept model with several explanatory variables is given by:

$$\log \mu_{ijk} = \log m_{ijk} + \boldsymbol{x}'\boldsymbol{\beta} + U_{0jk} + V_{00k}$$

where $U_{0jk} \sim \mathcal{N}(0, \tau_0^2)$ and $V_{00k} \sim \mathcal{N}(0, \theta_0^2)$
Similarly, the three-level random coefficient, 1-predictor model for count data is given by:

$$\log \mu_{ijk} = \log m_{ijk} + \beta_{0jk} + \beta_{1jk} x_{1ijk} + U_{0jk} + U_{1jk} x_{1ijk} + V_{00k} + V_{10k} x_{1ijk}$$

where $U_{1jk} \sim \mathcal{N}(0, \tau_1^2)$ and $V_{10k} \sim \mathcal{N}(0, \theta_1^2)$

More generally, we have that:

$$\log \mu_{ijk} = \log m_{ijk} + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{z}'^{(2)}\boldsymbol{U}_j + \boldsymbol{z}'^{(3)}\boldsymbol{V}_k$$

3. Application

The data set is sourced from www.football-data.co.uk, a free football betting portal that provides historical results and odds. The dataset comprises information about 6,860 football matches, two professional European football leagues and ten football seasons dating from 2005/2006 to 2014/2015. One of the European leagues is the German football league, the Bundesliga, where in every season there are 306 match fixtures. The other football league is the Serie A, an Italian league with 380 match fixtures per season.
Every football match is nested in the season during which it was played and, each season is nested in either one of the two football leagues. This structure invokes the multilevel nature of this data set where the level-1units are the football matches, the seasons are the level-2 units and the football leagues are the level-3 units. The response variable is the number of goals scored per match, and the ultimate scope of the study is to investigate the variability in this response variable induced by observed and unobserved heterogeneity. The following table defines the explanatory variables used in the Poisson multilevel models.

Table 1: Description of predictors

Notation	Predictor
hthg	The total number of goals scored by the home team during the first half
htag	The total number of goals scored by the away team during the first half
sa	The shooting accuracy is the ratio of the total shots on target to the total number of shots
fouls	The total number of fouls committed during the match
cards	Total number of yellow and red cards received during the match
home1h	1 corresponds to a home team win after the first half and 0 corresponds to otherwise
away1h	1 corresponds to an away team win after the first half and 0 corresponds to otherwise
corners	The total number of corners awarded during the match
dhtg	The absolute difference between the total home and away goals after the first half

In the two-level random intercept Poisson model given by:

$$\log \mu_{ij} = \log m_{ij} + \boldsymbol{x}'\boldsymbol{\beta} + U_{0j}$$

$\boldsymbol{x}'$ is a row vector including the values of the explanatory variables, $\boldsymbol{\beta}$ is a column vector of regression parameters in the fixed component of the multilevel model and U_{0j} is the random intercept with distribution $U_{0j} \sim \mathcal{N}(0, \tau_0^2)$. In this section, a parsimonious two-level random intercept Poisson model is fitted using seven predictors $x_{1ij}, \dots, x_{7ij}$, where $\beta_1, \dots, \beta_7$ are the corresponding parameters and β_0 is the intercept parameter. The model is implemented using the facilities of GLLAMM.

The adaptive quadrature converged after two iterations and another five iterations were needed to update the parameters using the Newton-Raphson algorithm. The log-likelihood of the parsimonious two-level random intercept Poisson model is -11255.4. The explanatory variables *dhtg* and *corners* were not significant and so were removed from the model fit. The estimated parameters $\beta_0, \beta_1, \dots, \beta_7$ and estimated variance τ_0^2 are displayed in Table 2.

Since the mean and variance of the Poisson distribution are equal then the variance to mean ratio is 1. Thus a value of 1 is used for level 1 variance. The fractions of residual variability that are attributed to level 1 and level 2 are 0.975 and 0.025

respectively. This implies that 97.5% of the total variance is accounted for by level-1 variations between matches and 2.5% is accounted for by level 2 variations between seasons.

Table 2: Parameter estimates, standard errors and p-values

Parameter	Coef.	S.E.	Z	$P>\|z\|$
Constant	-2.285	0.050	-45.6	0.000
hthg	0.273	0.011	25.6	0.000
htag	0.288	0.012	24.3	0.000
sa	1.455	0.074	19.7	0.000
fouls	-0.006	0.001	-5.52	0.000
cards	0.008	0.004	2.09	0.037
home1h	0.100	0.021	4.68	0.000
away1h	0.116	0.023	4.98	0.000
offset	2.398			
Level-1 var.	1			
Level-2 var. (int.)	0.025	0.007		

Figure 1 displays the path diagram to present the structure of the implemented 2-level random intercept model.

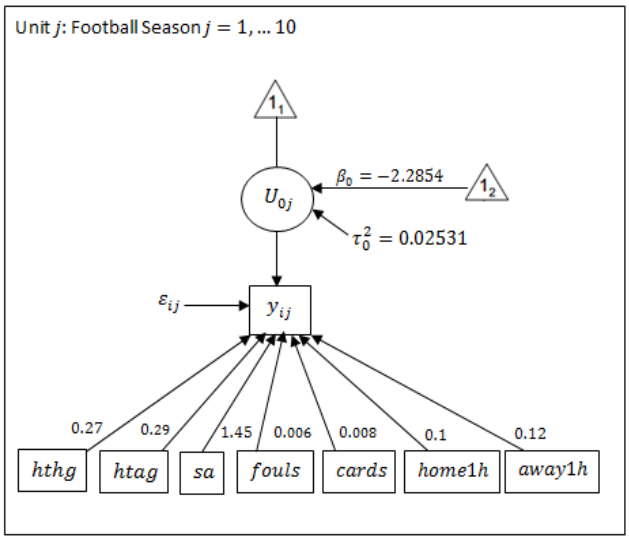

Figure 1: Path diagram for 2-level random intercept model

The gllapred poster directive is used to estimate the posterior means and posterior standard deviations using empirical Bayes prediction for random effects. For this 2-level random intercept model, different posterior means and posterior standard deviations are estimated for each of the 10 seasons. In order to predict the level-2 units specific regression lines with varying intercepts, the parameter estimates and the empirical Bayes estimates of the random intercept are plugged into the model. The gllapred predict, linpred directive is used to compute the linear predictor of the fixed component and adds it to the posterior mean.

The posterior standard deviations are the conditional standard deviations of the prediction errors given the observed responses and treating the parameters as known in a Bayesian context. Taking the square root of these standard deviations, one gets the conditional mean squared error of prediction conditional on the observed responses. The empirical Bayes estimates of the random intercept $U_{0j} \sim \mathcal{N}(0, 0.025)$ for each level-2 unit $j = 1,2, \ldots, 10$ along with the posterior standard deviations are provided in Table 3.

Table 3: Posterior means and posterior standard deviations

Season	Posterior Mean	Posterior St. Deviation
1	0.0322	0.0288
2	0.0286	0.0228
3	0.0741	0.0223
4	0.0399	0.0229
5	0.0474	0.0231
6	0.0617	0.0228
7	0.6658	0.0230
8	0.0990	0.0229
9	-0.0219	0.0231
10	-0.1704	0.0231

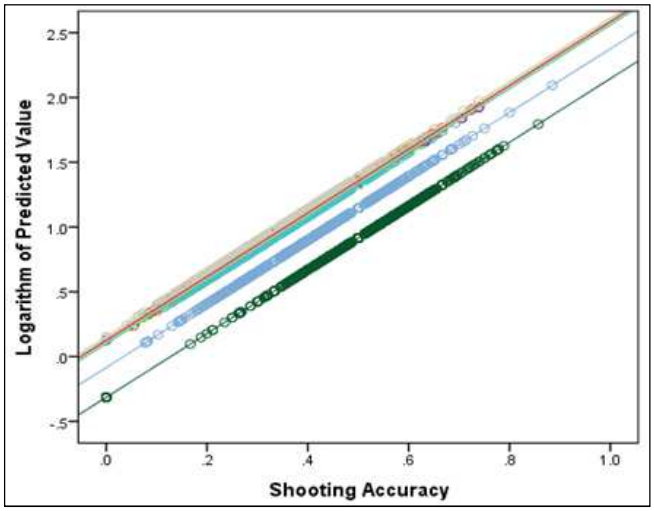

Figure 2: Log of predicted values against shooting accuracy

Figure 2 displays the logarithm of the predicted values against the shooting accuracy of the football match. The parameter estimate of *sa* is 1.4554 which implies that for every 1 unit increase in shooting accuracy the logarithm of the predicted value is expected to increase by 1.455, given that the other effects are kept fixed. The ten seasons trajectories displayed in Figure 2 have different intercepts but the same gradient as conditioned by two-level random intercept model. The ten trajectories have positive gradients implying that the number of goals per match increases with the shooting accuracy. It can be noted that the trajectories for the 2014/2015 and 2013/2014 seasons are below the other seasons which implies that in the last two football seasons the number of goals scored per match was less compared to the other seasons.

In the two-level random coefficient Poisson model given by:

$$\log \mu_{ij} = \log m_{ij} + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{z}'\boldsymbol{U}_j$$

$\boldsymbol{U}_j$ includes the random intercept U_{0j} and the random slope U_{1j}. The row vector $\boldsymbol{z}'$ holds a single explanatory variable, x_{3ij}, which is the shooting accuracy during the match. So the random slope U_{1j} allows the linear relationship between the logarithm of predicted values and shooting accuracy to have a

different slope for each of the ten seasons. $\boldsymbol{x}'$ includes the values of the predictors and $\boldsymbol{\beta}$ holds the regression parameters. The log-likelihood of the parsimonious two-level random coefficient model is -11253.9. The parameter estimates $\beta_0, \beta_1, \ldots, \beta_7$ and the estimated variances τ_0^2, τ_1^2 and τ_{10} are displayed in Table 4. Figure 3 displays the path diagram to present the structure of the implemented 2-level random coefficient model.

Table 4: Parameter estimates, standard errors and p-values

Parameter	Coef.	S.E.	Z	$P>\|z\|$
Constant	-2.296	0.048	-48.1	0.000
hthg	0.271	0.011	25.4	0.000
htag	0.288	0.012	24.3	0.000
sa	1.592	0.101	15.8	0.000
fouls	-0.005	0.001	-5.13	0.000
cards	0.007	0.004	1.96	0.049
home1h	0.100	0.021	4.67	0.000
away1h	0.113	0.023	4.86	0.000
offset	2.398			
Level-1 var.	1			
Level-2 var. (int.)	0.021	0.034		
Level-2 var. (slope)	1.158	0.537		
Level-2 covariance	-0.152	0.156		

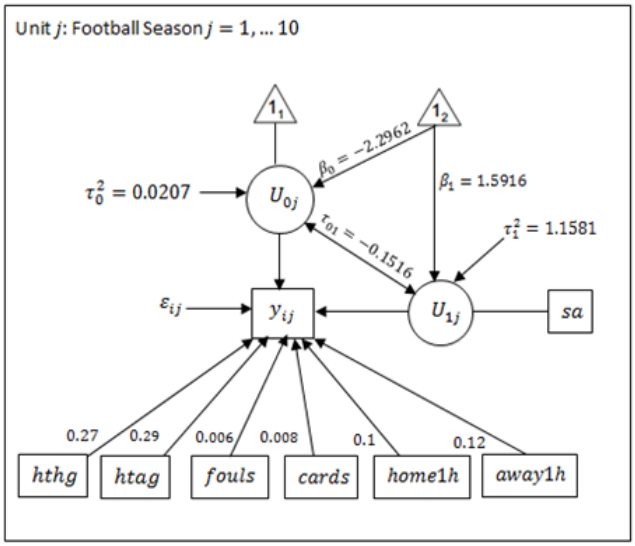

Figure 3: Path diagram for 2-level random coefficient model

Comparing these estimates to the random intercept model, the fixed effects estimates have not changed substantially but the estimates of the covariance matrix are quite different. The fraction of total residual variance attributed to the level-1 units is 0.459 and the fraction of total variance attributed to the level-2 random intercept and random slope are 0.01 and 0.531 respectively. This implies that 45.9% of the total variance is accounted for by level-1 variations between matches, 1% of the variance is accounted for variations between season intercepts and 53.1% of the variance is accounted for variations between season slopes. The empirical Bayes predictions for the random intercepts and the random slopes of the ten seasons are provided in Table 5 along with the posterior standard deviation.

$$U_j = \begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.0207 & -0.1516 \\ -0.1516 & 1.1581 \end{bmatrix} \right)$$

Table 5: Posterior means and posterior standard deviations

Season	P.M. 1	S.D. 1	P.M. 2	S.D. 2
1	-0.0044	0.0399	-0.0223	0.1254
2	0.0005	0.3989	-0.0477	0.1260
3	-0.0063	0.0392	0.0762	0.1215
4	-0.0010	0.0391	-0.0193	0.1216
5	-0.0081	0.0404	0.0249	0.1299
6	-0.0253	0.0401	0.1127	0.1263
7	-0.0008	0.0398	0.0341	0.1246
8	-0.0151	0.0409	0.1674	0.1346
9	0.0585	0.0374	-0.3433	0.1009
10	0.0827	0.0364	-0.6210	0.0831

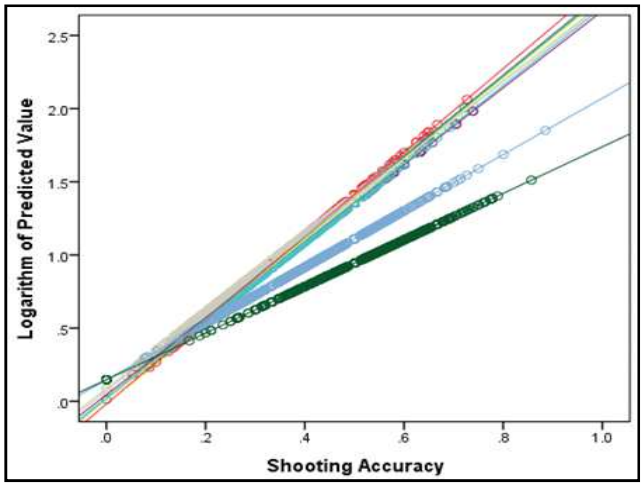

Figure 4: Log of predicted values against shooting accuracy

Table 6: Parameter estimates, standard errors and p-values

Parameter	Coef.	S.E.	Z	$P>\|z\|$
Constant	-2.295	0.053	-43.2	0.000
hthg	0.271	0.011	25.1	0.000
htag	0.288	0.012	24.1	0.000
sa	1.388	0.130	10.6	0.000
fouls	-0.006	0.001	-5.48	0.000
cards	0.009	0.004	2.29	0.022
home1h	0.099	0.021	4.64	0.000
away1h	0.111	0.023	4.78	0.000
offset	2.398			
Level-1 var.	1			
Level-2 var. (int.)	0.011	0.056		
Level-3 var. (int.)	0.019	0.037		
Level-3 var. (slope)	0.334	0.117		
Level-3 covariance	0.074	0.051		

The three-level random coefficient Poisson model given by:

$$\log \mu_{ijk} = \log m_{ijk} + \boldsymbol{x}'\boldsymbol{\beta} + U_{0jk} + \boldsymbol{z}'\boldsymbol{V}_k$$

$\boldsymbol{V}_k$ holds the random intercept V_{00k} and slope V_{10k} at level-3 and U_{0jk} is the random intercept at level-2. $\boldsymbol{z}'$ holds the

predictor, x_{3ij} and $\boldsymbol{x}'$ and $\boldsymbol{\beta}$ are the same as in previous models. The log-likelihood of the parsimonious three-level random coefficient model is -11243.8. The parameter estimates $\beta_0, \beta_1, \ldots, \beta_7$ and the estimated variances τ_0^2, θ_0^2, θ_1^2 and θ_{10} are displayed in Table 6.

This implies that 73.3% of the total variance is accounted for by level-1 variations between matches, 0.8% of the variance is accounted for by level 2 variations between seasons, 1.4% is accounted for by level 3 variations between football league intercepts and 24.5% is accounted for by level 3 variations between football league slopes.

Figure 5 displays the path diagram to present the structure of the implemented 3-level random coefficient model.

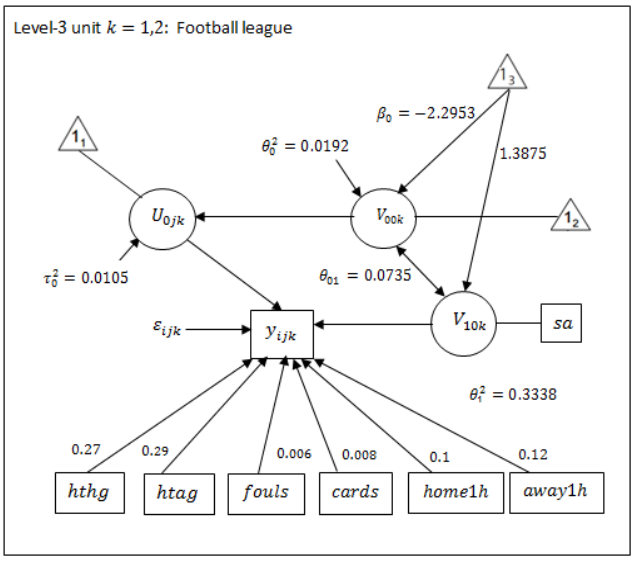

Figure 5: Path diagram for 3-level random coefficient model

$$U_{0jk} \sim \mathcal{N}(0, 0.011)$$

$$\boldsymbol{V}_k = \begin{bmatrix} V_{00k} \\ V_{10k} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.019 & 0.074 \\ 0.074 & 0.334 \end{bmatrix} \right)$$

The empirical Bayes estimates for the random intercept U_{0jk} for $j = 1,2, \ldots, 10$ along with the posterior standard deviations are given in Tables 7, along with the Bayes estimates for the random effects V_{00k} and V_{10k} for $k = 1,2$ and their posterior standard deviations given in Table 8.

Table 7: Posterior means and posterior standard deviations

Season	Posterior Mean	Posterior St. Deviation
1	-0.1370	0.0428
2	-0.0089	0.0419
3	0.0460	0.0419
4	-0.0040	0.0426
5	0.0080	0.0427
6	0.0527	0.0425
7	0.0097	0.0428
8	0.0758	0.0422
9	0.0664	0.0425
10	-0.2573	0.0475

Table 8: Posterior means and posterior standard deviations

League	P.M. 1	S.D. 1	P.M. 2	S.D. 2
1	0.0096	0.0328	0.1002	0.0705
2	0.0259	0.0336	0.1508	0.0729

The 3-level random coefficient model provides the best fit because it has the lowest AIC value.

4. Conclusion

Football is a game that has matured over the years, where football players run faster, they shoot harder, they dribble quicker and, they pass the ball more accurately. As a result, game practices including offside traps, pressing and triangular passing have evolved over time. Changes in these techniques are the main reason why goal scoring has gradually decreased from an average of approximately 4.5 goals per game in 1900 to an approximate average of 2.6 goals more than 100 years later. Goal scoring has remained essentially stable in the last two decades. Results in chapter 4 confirm the latter statement, since goal scoring is not affected much by the football season during which the game was played.

This paper presents a proper methodology to model count data in the presence of nested data. The three level random coefficient model which included shooting accuracy both as a main effect and as a random effect showed that 73.3% of the total variation is accounted for by variation at level-1, 0.8% is accounted for by variation at level-2, 1.4% is accounted for by variation at level-3 intercepts and 24.5% is accounted by variation in level-2 slopes. Moreover, shooting accuracy, number of fouls, number of red and yellow cards booked by referees, number of goals scored during the first half by the home team and by the away team and whether the home team is winning/losing at half time where all found to be significant predictors of the number of goals scored per match.

References

Bryk, A.S., and Raudenbush, S.W. (1992) Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. Journal of the Royal Statistical Society, Series A, 135, 370-384

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.

LIBERATO CAMILLERI studied Mathematics and Statistics at the University of Malta. He received his PhD degree in Applied Statistics from Lancaster University. His research specialization areas are related to statistical models, which include Generalized Linear models, Latent Class models, Multilevel models and Mixture models. He is an associate professor and Head of the Statistics department at the University of Malta.