Soc Indic Res (2019) 146:249–261
https://doi.org/10.1007/s11205-018-1891-6

A Generalized Estimating Equation in Longitudinal Data to Determine an Efficiency Indicator for Football Teams

Anna Crisci[1] · **Luigi D'Ambra**[2] · **Vincenzo Esposito**[3]

Accepted: 26 March 2018 / Published online: 30 March 2018

Abstract Over the years football has attracted enormous interest from various fields of study, attracting attention both for its sporting and social aspects. Professional business operators consider football an important industry with enormous potential both in terms of its size and growth, and also because of indirect benefits due to the popularity gained by investors and management of football teams. The focus of the analysis has been on what characterizes most football clubs, and determines their particular economic and financial needs. The aim of this paper is to establish an efficiency measurement for football team financial resource allocation. In particular, we analysed the impact that the income statement, Net equity and Team value variables have on the points achieved by football teams playing in "Serie A" championship (Italian league). The method used in our study is a generalized estimating equation (GEE) for longitudinal count data. In addition we consider a coefficient of determination in the GEE approach based on Wald Statistics, and we propose a modified Mallow's Cp for choosing the best model. Finally we propose an AFRSport index based on the differences between observed and theoretical points, in order to identify those teams that efficiently employ their financial resources.

Keywords Financial indicator · Generalized estimating equations · Best subset · C_P Mallows

✉ Anna Crisci
anna.crisci@unipegaso.it

Luigi D'Ambra
dambra@unina.it

Vincenzo Esposito
vincenzo.esposito@quadrans.it

[1] Department of Law and Economic Sciences, Pegaso Telematic University, Naples, Italy

[2] Department Economics, Management and Institutions, University of Naples, Federico II, Naples, Italy

[3] Quadrans S.R.L, Naples, Italy

1 Introduction

In Italy, football (soccer) is an element of tradition and sociality. At a national and international level this sport represents a very important industry, and even emergent countries like China are preparing to raise the intensity of significant investments in the game.

Professional business operators consider football an important industry with enormous potential in terms of growth and also for the indirect benefits gained by investors and management due to the popularity of football teams.

An important research program called "Football Report 2017", now in its 7th edition, was commissioned by the Italian Football Association and implemented by PricewaterhouseCoopers (PWC). It evidences how the financial numbers relative to Italian professional football are important. The "Football Report 2017" is just one of the documents that the Federation makes systematically available to its stakeholders, and from the Football Report 2017 we find the following:

- € 26.6 billion: Aggregate turnover of worldwide football (46% of total revenue in the global sports business);
- 2718: the total number of sponsors of the clubs participating in the Top 10 European Leagues, of which 22.2% are foreign;
- 1.1 billion: the aggregate data for fans and followers present in the social networks of clubs taking part in The Top 10 European League competitions;
- Tax and social security contributions of Italian professional football in 2014 amounted to € 1.073 million;
- The value of aggregate production of Italian professional football has grown in the last 5 years, going from 2625.1 million in 2014–2015 to the current 2857.7 million.

In the football world, major consulting companies provide statistical data relating exclusively to athletic performance and sports results. The recipients of such data can be placed in two main categories. The first concerns professional football players, sports clubs, coaches, sports directors, etc. Such information is sold, in some cases, for payment. The second category is represented by media outlets, which release statistical reports to fans and sports people.

There are numerous statistics that regard rewards for the best players (e.g. highest goal scorer, number of assists, number of successful passes, etc.) and the performance of individual football teams (e.g. number of matches won, goals scored, etc.). There is, however, a lack of measurement of the efficiency of individual teams in this area.

For this reason, the aim of this paper is to measure the efficiency of the allocation of financial resources by football teams. In particular, we analyse the impact that some variables, including Income statement, Net equity and Team value, have on points made by football teams participating in the series A championship, by means of a generalized estimating equation (GEE) for longitudinal count data.

The GEE approach was developed by Liang and Zeger (1986) and Zeger and Liang (1986), to produce more efficient and unbiased regression estimates when analyzing longitudinal or repeated measures research designs with non-normal response variables. GEE is able to cope with correlated data within subjects (teams). The main idea behind GEE is to generalize and extend the usual likelihood equation for a Generalized Linear Model by including the covariance matrix of the responses. The biggest advantage of the GEE is that we do not need to

specify the whole distribution of the response. Only the mean structure, the mean–variance relationship and specification of the covariance structure need to be defined.

Moreover, GEE are versatile and the processing of observational results leads to a very interesting theoretical development. Indeed, it is possible to specify or estimate the correlation matrix in order to estimate more efficient regression coefficients than those obtainable under the hypothesis of independence.

The GEE model we are going to propose will provide a financial indicator of the efficiency of football teams on the basis of the variables mentioned above. In particular, the GEE model provides the theoretical points that individual football clubs should realize through efficient employment of productive factors.

The differences between observed and theoretical points represent the ability of individual team to efficiently mix the economic and financial variables considered.

We define the relative residual of the points (positive or negative) as AFRSport (Allocation Financial Resources Sport) Index. The AFRSport index has allowed us to identify the teams that efficiently employ their financial resources. The proposed AFRSport index can also be used in other sports where there is a ranking.

In addition to the introduction, this paper consists of four further sections. In Sect. 2, a summary of the Generalized Estimating Equation is described. In Sect. 3, some criteria for selecting a working correlation structure and for selection of the best subset are shown. In Sect. 4, the data and results are described. We leave some final conclusions in Sect. 5.

2 Summary of the Generalized Estimating Equation Method (GEE)

Let $\boldsymbol{y_i} = \left(y_{i1}, \ldots, y_{iT}\right)'$ be a vector of responses from n subject, with T observation for the ith subject, i=1,…,n. For each y_{it} a vector of covariates $\boldsymbol{x_{it}} = \left(x_{it1}, \ldots, x_{iTK}\right)'$ for t=1,…,T, is available.

The expected value and variance of measurement y_{it} on subject i can be expressed using a generalized linear model:

$$E(y_{it}|\boldsymbol{x_{it}}) = g\left(\boldsymbol{x'_{it}\beta}\right) = \mu_{it}$$

where g is the non-linear response function, and g^{-1} is a known inverse link function. $\boldsymbol{\beta}$ is an unknown $K \times 1$ vector of regression coefficients with the true value as β_0. The conditional variance of y_{it} given x_{it} is specified as $Var(y_{it}|\boldsymbol{x_{it}}) = \upsilon\left(\mu_{it}\right)\phi$, where υ is a known variance function of μ_{it} and ϕ is a scale parameter which may need to be estimated. Mostly, υ and ϕ depend on the distributions of outcomes. For instance, if y_{it} is continuous, $\upsilon\left(\mu_{it}\right)$ is specified as 1, and ϕ represents the error variance; if y_{it} is count, $\upsilon\left(\mu_{it}\right) = \mu_{it}$ and ϕ is equal to 1. Also, the variance–covariance matrix for $\boldsymbol{y_i}$ is noted by $\boldsymbol{V}_i = \phi \boldsymbol{A}_{\boldsymbol{i}}^{\frac{1}{2}} \boldsymbol{R_i}(\boldsymbol{\alpha})$, $\boldsymbol{A}_i = Diag\left\{\upsilon\left(\mu_{i1}\right), \ldots, \upsilon\left(\mu_{iT}\right)\right\}$ and the so-called "working" correlation structure $\boldsymbol{R_i}(\boldsymbol{\alpha})$ describes the pattern of measures within the subject, which is of size $T \times T$ and depends on a vector of association parameters denoted by $\boldsymbol{\alpha}$. An iterative algorithm is applied for estimating α using the Pearson residuals $e_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{\upsilon(\mu_{it})}}$ calculated from the current value of $\boldsymbol{\beta}$.

Also, the scale parameter ϕ can be estimated by:

$$\hat{\phi} = \frac{1}{n-K} \sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2$$

The parameters β are estimated by solving:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i'[V(\hat{\boldsymbol{\alpha}})]^{-1}\boldsymbol{S}_i = 0$$

where $\boldsymbol{S_i} = \left(\boldsymbol{y_i} - \boldsymbol{\mu_i}\right)$ with $\boldsymbol{\mu_i} = \left(\mu_1, \ldots \mu_{iT}\right)'$ and $(\hat{\boldsymbol{\alpha}})$ is a consistent estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{D}_i = \partial \boldsymbol{u_i}/\partial \boldsymbol{\beta}'$. Under mildregularity conditions $\hat{\boldsymbol{\beta}}$ is asymptotically distributed with a mean β_0 and covariance matrix estimated based on the sandwich estimator (White 1980):

$$\hat{\boldsymbol{V}}_i^{\boldsymbol{R}} = \left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1} \sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{S}_i\boldsymbol{S}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}$$

In GEE models, if the mean is correctly specified, but the variance and correlation structure are incorrectly specified, then GEE models provide consistent estimates of the parameters and also the mean function, while consistent estimates of the standard errors can be obtained via a robust "sandwich" estimator. Similarly, if the mean and variance are correctly specified but the correlation structure is incorrectly specified, the parameters can be estimated consistently and the standard errors can be estimated consistently with the sandwich estimator.

The robust covariance matrix adjusts for the loss of efficiency due to possible misspecification of variance function.

The dependence structure among repeated measurements of the same individual are dealt with via the definition of the "working" correlation matrices.

The working correlation matrixes used in GEE are:

- Indipendent (R_{IN}) assume that correlations for distinct measurements of the same individual are zero;
- Exchangeable (R_{EX}): assume that correlation between all measurements of the same individual are equal; this matrix requires a single parameter estimation;
- Autoregressive of order 1 (R_{AR-1}): repeated measurements have a first-order autoregressive relationship. A characteristic of the R_{AR-1} is that the magnitude of the (positive) correlations quickly decreases over time;
- Unstructured (R_{UN}): an unstructured working correlation matrix has no explicit pattern, but rather every coefficient correlation is allowed to be different. All T($T - 1$)/2 correlations are estimated.

3 Criteria for Selecting a Working Correlation Structure

Unlike the GLM method, which is based on the maximum likelihood theory for independent observations, the GEE method is based on the quasilikelihood theory and no assumption is made about the distribution of response observations. Therefore, AIC (Akaike's Information Criterion), a widely used method for model selection in GLM, is not directly applicable to GEE.

The *QIC* (Quasilikelihood under the Independence model Criterion) statistic proposed by Pan (2001), and further discussed by Hardin and Hilbe (2003), is analogous to the familiar AIC statistic used for comparing models fit with likelihood-based methods:

$$QIC = -2Q(\hat{\boldsymbol{\mu}};\boldsymbol{I}) + 2\boldsymbol{trace}\left(\hat{\Omega}_{\boldsymbol{I}}^{-1}\hat{\boldsymbol{V}}_{\boldsymbol{i}}^{\boldsymbol{R}}\right)$$

where $\boldsymbol{I}$ represents the independent covariance structure used to calculate the quasilikelihood, $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{x}_{it}\hat{\boldsymbol{\beta}})$. The coefficient estimates $\hat{\boldsymbol{\beta}}$ and robust variance estimator $\hat{\boldsymbol{V}}_{\boldsymbol{i}}^{\boldsymbol{R}}$ are obtained from a general working covariance structure. Another variance estimator $\hat{\Omega}_I$ is obtained under the assumption of an independence correlation structure.

QIC can be used to find an acceptable working correlation structure for a given model. However, Hardin and Hilbe (2003) recommend the use of QIC only to choose from otherwise equally suitable structures.

Other criteria for selection of the working correlation are available. Rotnitzky and Jewell (1990) describe an approach to appraise the adequacy of the assumed correlation matrix based on the fact that the asymptotic distribution of a modified working Wald statistic is the linear combination of independent χ_1^2 random variables. This criterion is called the Rotnitzky and Jewell criterion (*RJC*).

Furthermore, Hin and Wang (2009) propose a correlation information criterion (*CIC*) that modifies *QIC* and substantially improves its performance.

Gosho et al. (2011) devised an objective criterion for evaluating the appropriateness of the correlation structure. The proposed criterion measures the discrepancy between the covariance matrix estimator and the specified working correlation matrix.

Hin et al. (2007), Hin and Wang (2009), and Gosho et al. (2011) compared the performances of criteria previously mentioned for selecting the working correlation structure.

In this paper, we consider the QIC criterion because it allows the selection of the covariates and working correlation structure simultaneously.

3.1 Criteria for Selection of the Best Subset

When trace $\hat{\Omega}_{\boldsymbol{I}}^{-1}\hat{\boldsymbol{V}}_{\boldsymbol{i}}^{\boldsymbol{R}} \approx trace(\boldsymbol{I}) = K$, there is a simplified version of *QIC*, called QIC_u (Pan 2001),

$$QIC_u = -2Q(\hat{\boldsymbol{\mu}};\boldsymbol{I}) + 2K$$

QIC and the related QIC_u statistics can be used to compare GEE models and aid model selection. QIC_u approximates *QIC* when the GEE model is correctly specified. Models do not need to be nested in order to use *QIC* or QIC_u to compare them. QIC_u should not be used for selecting a working correlation structure.

When using *QIC* or QIC_u to compare two structures or two models, the model with the smaller statistic is preferred.

Now we discuss the use of the coefficient of determination (Natarajan et al. 2007) for the GEE method in order to measure the strength of association between the response variable and covariates. We propose a modified Mallows' Cp for the choice of one or the best subsets.

For this purpose, we carried out a review of ordinary least squares.

Consider a standard linear regression model with the model, in which the ith individual has response Y_i and a K covariate vector $\boldsymbol{x_i}$:

$$E\left(Y_i|\boldsymbol{x_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} = \boldsymbol{x}_i'\beta$$

and variance

$$Var\left(Y_i|\boldsymbol{x}_i\right) = \sigma^2$$

Let $\hat{\beta}_k$ be the maximum likelihood estimator of β_k for k = 1,…,K and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$ be the maximum likelihood estimator of σ^2.

In this paper, we consider the unbiased estimator of σ^2, that is, $\hat{s}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-K-1}$ where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK}$.

To test $H_0 : \beta_1 = \beta_2 = \cdots \beta_K = 0$ versus $H_1 : \beta_k \neq 0$for at least one k, one can use the Wald statistic

$$Q = \left[C\hat{\beta}\right]' \left[C\widehat{Var}\hat{\beta}C'\right]^{-1} \left[C\hat{\beta}\right]$$

where $\boldsymbol{C}$ is a $(K-1) \times K$ matrix with its first column having all 0s, and its last $(K-1)$ columns being the $(K-1)$ identity matrix. Finally $\widehat{Var}\hat{\beta}_k = \hat{\sigma}^2 (X'X)^{-1}$.

Christensen (1996) shows that the coefficient of determination R^2, equals:

$$R^2 = \frac{Q/(n-K-1)}{1 + Q/(n-K-1)}$$

we consider the following coefficient of determination:

$$\tilde{R}^2 = \frac{\tilde{Q}/(n-K-1)}{1 + Q/(n-K-1)}$$

where

$$\tilde{Q} = \left[C\hat{\beta}\right]' \left[C\widehat{Var}(\hat{\beta})C'\right]^{-1} \left[C\hat{\beta}\right]$$

is the Wald statistics with the GEE robust covariance matrix estimated under the null and denoted by $\widehat{Var}\hat{\beta}$. The idea is that $\tilde{R}^2$ transforms the Wald statistic $\hat{Q}$ to a more intuitively appealing (0.1) scale.

It is important to know that unlike linear regression using OLS, there is no guarantee that a model with additional covariates would have a larger $\tilde{R}^2$

Now, we can propose partial $\tilde{R}^2$ by:

$$\tilde{R}^2_{partial} = \frac{\tilde{R}^2_{(x_1,x_2,\ldots,x_K)} - \tilde{R}^2_{(x_1,x_2,\ldots,x_{K-1})}}{1 - \tilde{R}^2_{(x_1,x_2,\ldots,x_{K-1})}}$$

where $\tilde{R}^2_{(x_1,x_2,\ldots,x_K)}$ considers all covariates in the model and $\tilde{R}^2_{(x_1,x_2,\ldots,x_{K-1})}$ considers all covariates except x_K.

The aim of this paper is to select the best subset models from among the 2^K models. In order to select the best model, we propose a modified Mallows' Cp ($\hat{C}_p$) in GEE that is closely related to $\tilde{R}^2$ defined above:

$$\hat{C}_p = (n-K)\frac{1-\tilde{R}^2_p}{1-\tilde{R}^2_K} + 2p - n \quad \text{with} \quad p \leq K$$

where $\tilde{R}^2_p$ is calculated by considering the Wald statistics with the GEE robust covariance matrix and p regressors, while $\tilde{R}^2_K$ is calculated by considering the Wald statistics with the GEE robust covariance matrix and the complete set of K regressors.

Both *QIC* and $\hat{C}_p$ consider the possible difference in the number of parameters among different models. Moreover, the use of the *QIC* allows you to choose only one subset that has the smallest *QIC*, in the case of the $\hat{C}_p$ it is possible to locate multiple subset optimals as being the choice tied to a near value of p.

Mallows (1973) also suggested that a good model has a negative or small $\hat{C}_p - \mathrm{p}$.

4 Data and Results

4.1 Sample

The data used for our study was obtained from the financial statements filed by the Serie A football teams. The analysis was conducted on all the filed balance sheets. The period of study concerned five consecutive sports seasons from season 2010/2011 up to 2014/2015.

The focus of the analysis is to verify the impact that the income statement, Net equity and Team value variables have on the points achieved by football teams.

This choice also arises from the need of an assessment model that leads to the identification of an economic-financial measurement indicator for the efficiency of football teams.

The independent variables considered are: Wages; Depreciation Expense of multi-annual player contracts; Revenue net of player capital gain; Team-value; Net Equity.

4.1.1 Wages (W)

They refer to labor costs, specifically footballers' salaries. In particular, wages are the total cost of labor and include both social security contributions and severance pay. The salary logarithm of each team was considered. Forrester and Simmons (2002) show that in European football high wage expenditures clearly increase the chances of success on the pitch. Estimated at the beginning of each football year, all salaries are expressed in thousands of euros and do not include any bonuses.

4.1.2 Depreciation Expense of Multi-annual Player Contracts (DEM)

The depreciation expense of multi-annual player contracts are carried out a cost with an amortization plan. In the particular case of football club financial statements, the purchase of a football player is considered to be an immaterial immobilization, since it is the company's "right" to be the exclusive recipient of an athlete's sporting performance for a certain number of years. This investment is a cost shared for a period of time equal to the duration of the contract that the company has signed with the player. In our study we consider the log of depreciation Expense of multi-annual players.

4.1.3 Revenue net of Player Capital Gain (RNC)

Like all companies, football clubs have different categories of income: characteristic and accessories:

- Typical revenues

  Revenue from the stadium, television rights, sponsorships, football rights, participation rights in European competitions;

- Revenue accessories: capital gains
  The log of RNC considers both Typical Revenues and Revenue accessories.

4.1.4 Team-Value (TV)

This is the only variable used that is not derived from financial statements but is an estimated value made by transfermarkt.[1] It should be emphasized that the value of each player is something attributed and varies according to the declaration source. In other words, it is not true data. Understanding the monetary policy of a club and the far-sightedness of the transfer market is, however, significant. We consider the log of Team-value.

4.1.5 Net Equity (NE)

Net equity is the difference between assets and liabilities and all the resources that the company has as a form of internal financing. Equity may be affected positively by contributions from shareholders (capital increases, retained earnings, etc.) and the profits generated by the company.

Operating losses, the repayment of capital to members, results in a decrease in shareholders' equity. We consider the log of Net Equity.

As regards the responses (y_{it}), they consist of the points achieved, with games won and tied, by the football teams participating in the Serie A championship. Moreover, the responses within subject are not independent (they are correlated with each other), but they are independent across subjects.

It is important to emphasize that the aim of the work is to analyze how some income statement variables including Net equity and Team value affect the points achieved by football teams. Other variables such as player, injured players and coaching statistics are not considered within this particular context, but may be considered in future research that is not purely financial/economic but linked to individual player characteristics, in addition to team organization and management factors.

4.2 Results

The objective of the present work is to analyse, by means of generalized estimating equations (GEE) for longitudinal count data, the impact that some income statement variables including Net equity and Team value have on points achieved by football teams in the Serie A championship. We consider, therefore, the GEE poisson that estimates the same model as the standard poisson regression. Unlike that seen in poisson regression, GEE poisson allows for dependence within subjects, such as in longitudinal data, although its use is not limited to just panel data. We use GEE with robust variance estimates to model within-team correlation. We selected GEE instead of a mixed model as we are interested in understanding the influence of overall budget variables rather than individual team level effects. The GEE are appropiate for analyzing longitudinal data with relatively small between subject variation.

[1] The transfertmarket is one of the most important sites in the world, which newspapers refer to in order to assess the market value of football players in real time.

Table 1 Results *QIC* for different working correlation matrixes

Correlation	*K*	*QIC*
Exchangeable (R_{EX})	5	87.48
Unstructured (R_{UN})	5	89.16
Independent $\left(R_{IN}\right)$	5	88.72
AR-1 (R_{AR-1})	5	93,37

Variables: W, DEM, RNC, TV, NE

With the above GEE model established, we now calculate the *QIC* value in order to select the best working correlation structure among four structures, exchangeable, unstructured, independent and AR-1. The calculation results are presented in Table 1. As previously discussed in Sect. 3, the best working correlation structure in the GEE model is the one with the smallest *QIC* value. Therefore, we decide to apply the established GEE model under Exchangeable working correlation structure.

The exchangeable working correlation structure assumes that the correlation (0.7047) between any pair of measurements is the same. R_{EX} is often used as a practical choice in small samples, since R_{EX} is very parsimonious with only one parameter. In our case, we have a short panel, in which the budget variables did not change substantially over time.

Now, let's begin the choice of the best subsets using the criteria described in Sect. 3.1.

In Table 2 we report the Wald statistics, $\tilde{R}^2$, *QIC*, QIC_u and modified $\tilde{C}_p$, for all 2^K models.

Following the *QIC* criterion, the model with the smaller statistics is the Model 15 (M15), with variables RNE and NE. By contrast the modified Mallow's $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$ leads us to choose those subset whose $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$ is close to the number of variables, then Model 4 (M4), Model 8(M8). Moreover, we consider also Model 15 (M15), Model 24(M24), Model 25(M25) and Model 29(M29) with negative $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$.

Note that in the models with negative $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$ the presence of the variables Ne and RNC. This confirms the importance of these variables.

As these are potential models, our attention is on Model 15, Model 25 and Model 29, which have the smallest $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$ value. Of these models, Model 29 is not chosen, both for the *QIC* criterion and because it has a higher negative $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$ than the other models.

In other words, the choice falls on Model 25, which presents the lowest $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$, the best $\tilde{\boldsymbol{R}}^2$, and whose *QIC* does not differ much from Model 15. Table 3 shows the output of Model 25.

The most significant variables are RNE and NE. The variable DEM is not significant (p value > 0.05). In fact, the contribution of this variable in the explanation of the response variable is not relevant, the partial $\tilde{\boldsymbol{R}}^2$ (see Sect. 3.1) is 0.067 (about 6.7%). Despite the lack of contribution this variable is kept as it refers exclusively to multi-annual depreciation of rights for the purchased football players. It is therefore important because buying stronger players involves higher depreciation in the balance sheet. Moreover, this variable measures the ability of football companies to make new investments. However, both semi-robust standard error and standard error (SE) are approximatively the same. This could indicate that the true correlation structure for the GEE is correctly modeled using exchangeable model assumption.

In addition, since there is no formal diagnostic tool available in the GEE framework to verify the adequacy of the model, we could consider a Q–Q plot based on the χ^2

Table 2 Goodness of fit statistics for GEE models

	Models					Wald statistic	$\tilde{R}^2$	*QIC*	QIC_u	$\tilde{C}_p$
M1	W					93.73**	0.49	111.33	108.58	38.53
M2	DEM					30.91**	0.24	139.60	133.22	104.11
M3	TV					83.53**	0.46	108.36	106.56	46.07
M4	RNC					166.18**	0.63	94.14	94.87	1.77
M5	NE					17.59**	0.15	163.02	153.43	127.25
M6	W	DEM				92.68**	0.49	116.00	112.02	40.59
M7	W	TV				96.58**	0.50	108.06	106.02	37.88
M8	W	RNC				166.56**	0.63	101.81	96.09	3.00
M9	W	NE				101.99**	0.52	91.53	91.14	34.31
M10	DEM	TV				88.21**	0.48	110.31	108.48	43.83
M11	DEM	RNC				169.00**	0.64	98.17	96.37	2.12
M12	DEM	NE				47.07**	0.33	118.62	114.87	83.18
M13	TV	RNC				161.84**	0.63	97.22	95.29	4.76
M14	TV	NE				75.62**	0.44	99.53	98.92	53.19
M15	RNC	NE				184.43**	0.66	83.82	85.16	−3.12
M16	W	DEM	TV			97.13**	0.51	111.15	108.51	38.82
M17	W	DEM	RNC			159.43**	0.63	104.39	98.47	7.03
M18	W	DEM	NE			106.34**	0.53	93.40	92.67	32.88
M19	W	TV	RNC			164.85**	0.63	102.00	97.36	4.99
M20	W	TV	NE			101.41**	0.52	92.65	92.99	35.99
M21	W	RNC	NE			163.24**	0.63	87.44	86.62	5.59
M22	DEM	TV	RNC			160.67**	0.63	98.72	96.92	6.56
M23	DEM	TV	NE			79.20**	0.46	101.24	100.88	52.19
M24	TV	RNC	NE			180.21**	0.66	86.37	87.12	−0.37
M25	DEM	RNC	NE			202.52**	0.68	85.77	86.66	−7.16
M26	W	DEM	TV	RNC		156.66**	0.63	103.79	99.24	9.46
M27	W	DEM	TV	NE		107.09**	0.53	94.01	99.23	33.73
M28	W	DEM	RNC	NE		170.67**	0.64	87.77	87.29	4.25
M29	DEM	TV	RNC	NE		192.78**	0.67	87.25	88.36	−2.94
M30	W	TV	RNC	NE		169.89**	0.64	88.36	88.56	4.53
M31	W	DEM	TV	RNC	NE	171.38**	0.65	87.48	89.30	6

**Significant at 5%

Table 3 Poisson GEE population-averaged model with exchangeable structure (model 25)

Points	Coeff.	Semirobust SE	Z	**P > |z|**
DEM	−0.0546	0.0725 (0.070)	−0.75	0.452
RNC	0.4523	0.0805 (0.083)	5.62	0.000
NE	0.1808	0.0458 (0.060)	3.95	0.000
Cons	−5.9522	0.6266	−9.50	0.000
Wald Statistic 202.52; Prob > Chi-square 0.000				

(.) Shows model based standard error (SE)

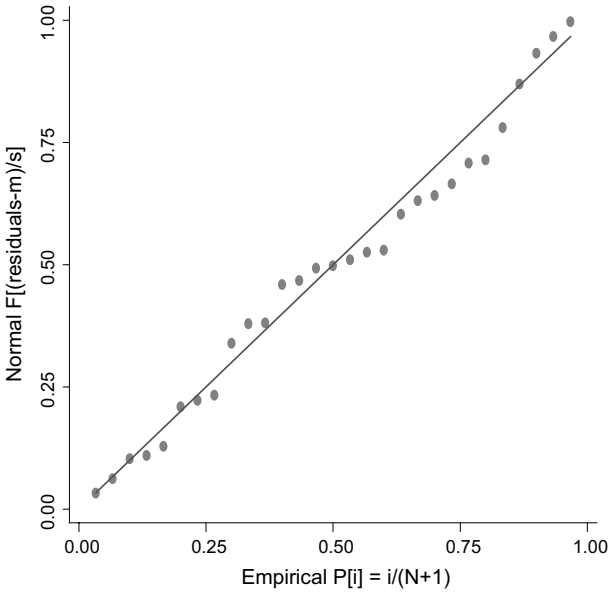

Fig. 1 Poisson plot for residuals. The plot is based on Model 25

distribution. By using the same notation as Park and Lee (2004), let $\boldsymbol{e_i}$ be the $T \times 1$ vector of the ith subject residual vector, $\boldsymbol{W}_i$ be the variance matrix of $\boldsymbol{e_i}$. When $\boldsymbol{W}_i$ is known and the mean model is correct,$q_i = \boldsymbol{e_i'} \boldsymbol{W}_i^{-1} e_i$, i = 1,…,n, are approximately distributed as the χ^2 distribution with T degrees of freedom.

However, when the number of responses from the same subject differs from subject to subject due to unbalanced observations (t_i), the degrees of freedom differ, and it is not possible to construct a Q–Q plot based on the χ^2 distribution. In this case, we consider the simple fourth root transformation proposed by Hawkins and Wixley (1986) to achieve approximate normality, so that:

$$q_i^N = \frac{q_i^{1/4} - (t_i - 0,5)^{1/4}}{(8\sqrt{t_i})^{-1/2}}$$

is approximately distributed as standard normal distribution $N(0, 1)$. Figure 1 shows the performance of residual plots based on the transformation by Hawkins and Wixley (1986). The plot shows a good fit of the model 25, the residuals are quite close to the reference line.

The GEE model provides the theoretical points that individual football clubs should realize when their productive factors are employed efficiently.

The differences between observed and theoretical points represent the ability of individual team to efficiently mix the considered financial variables.

The relative residual of the points (positive or negative) is defined by the AFRSport (Allocation Financial Resources Sport) Index. The AFRSport index finds its explanation in the entrepreneurial ability of sports management to allocate resources more efficiently.

The AFRSport index allowed us to build a ranking by referring exclusively to the teams that participated in all 5 championships. Figure 2 shows the AFRSport values.

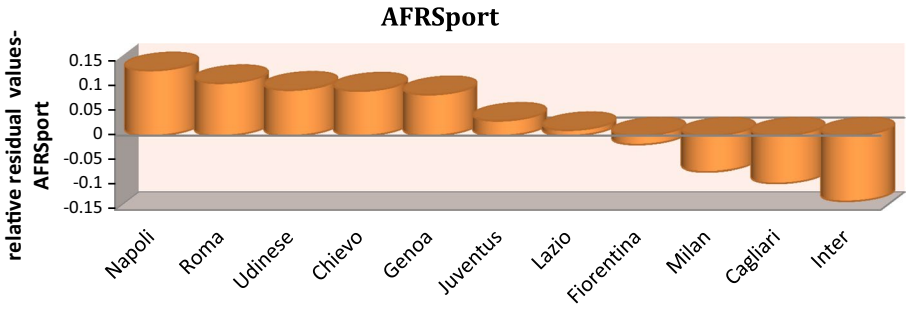

Fig. 2 Efficiency index (AFRSport) of the football teams-year 2010–2015

In particular, the y- axis indicates the relative residual of the points. For example, the AFRSport index of Inter equals: -0.15, and this means that Inter obtained point that was 15% less than predicted point, so has failed with regard to optimal use of financial resources.

5 Conclusion

In this paper, we have proposed an efficiency measurement for the allocation of football team financial resources. We have analyzed the impact that some income statement, Net equity and Team value variables have on points achieved by football teams playing in the Serie A championship. The method used for our study is a generalized estimating equation (GEE) for longitudinal count data.

Moreover, following (Natarajan et al. 2007) we consider a coefficient of determination based on Wald statistics and we propose Mallow's Cp for the choice of the best model. We campare the Mallow's Cp with the classical maesure in GEE, that is QIC.

Both *QIC* and $\tilde{C}_p$ consider the possible difference in the number of parameters among different models. Moreover, the use of the *QIC* allows you to choose only one optimal subset that has the smallest QIC, in the case of the $\tilde{C}_p$ it is possible to locate multiple subset optimals.

The GEE model provides the theoretical points that individual football clubs should realize through efficient employment of productive factors.

The differences between observed and theoretical points represent the ability of individual companies to efficiently mix the economic and financial variables considered.

We define the relative residual of the points (positive or negative) as AFRSport (Allocation Financial Resources Sport) Index. The AFRSport index has allowed us to identify the teams that efficiently employ their financial resources.

Acknowledgements We would like to thanks the anonymous referee for carefully reading our manuscript and for giving such constructive comments which substantially helped improving the quality of the paper.

References

Christensen, R. (1996). *Plane answers to complex questions. The theory of linear models* (2nd ed.). New York: Springer.

Forrester, D., & Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: The Case of English soccer. *Journal of the Royal Statistical Society: Series D (The Statistician), 51*(2), 229–241.
Gosho, M., Hamada, C., & Yoshimura, I. (2011). Criterion for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data. *Communications in Statistics Theory and Methods, 40*(21), 3839–3856.
Hardin, J., & Hilbe, J. (2003). *Generalized estimating equations*. London: Chapman and Hall.
Hawkins, D. M., & Wixley, R. A. J. (1986). A note on the transformation of Chi squared variables to normality. *The American Statistician, 40,* 296–298.
Hin, L. Y., Carey, V. J., & Wang, Y. G. (2007). Criteria for working-correlation-structure selection in GEE: Assessment via simulation. *The American Statistician, 61*(4), 360–364.
Hin, L. Y., & Wang, Y. G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine, 28*(4), 642–658.
Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22.
Mallows, C. L. (1973). Some comments on CP. *Technometrics*. https://doi.org/10.2307/1267380.
Natarajan, S., Lipsitz, S., Parzen, M., & Lipshultz, S. (2007). A measure of partial association for generalize estimatiing equations. *Statistical Modelling, 7*(2), 175–190.
Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics, 57*(1), 120–125.
Park, T., & Lee, S. Y. (2004). Model diagnostic plots for repeated measure data. *Biometrical Journal, 46,* 441–452.
Rotnitzky, A., & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika, 77*(3), 485–497.
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. https://doi.org/10.2307/1912934.
Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42,* 121.