The study of relationship between financial performance and points achieved by Italian football championship clubs *via* GEE and diagnostic measures

Lo studio della relazione tra risultati finanziari e punti realizzati delle squadre di calcio di serie A tramite GEE e misure di diagnostica

Anna Crisci, Sarnacchiaro Pasquale e Luigi D'Ambra

Abstract
Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. The main goal of any Football Championship club is to achieve sport results. The study of the relationship between sport and economic results attracts the interest of many scholars belonging to different disciplines. Very informative is considered the connection, over short or long periods of time, between the points in the championship and the resource allocation strategies. The aim of this paper is to give a interpretation of this last link using the Generalized Estimating Equation (GEE) for longitudinal data. Some diagnostic measures and graphical plots for checking the adequacy of GEE method will be presented and used.

Abstract
Il calcio in Italia è un fenomeno sociale che coinvolge intere comunità e continua ad aumentare il suo valore sociale ed economico. Lo studio della relazione tra i risultati sportivi ed economici riscuote l'interesse di tantissimi studiosi appartenenti a diverse discipline. Particolarmente stimolante è risultato il dibattito che lega, per ciascuna squadra di calcio, i punti in classifica alle capacità imprenditoriali del management sportivo in termini di allocazione delle risorse finanziarie e sportive. Obiettivo del presente lavoro è quello di dare un contributo in termini di interpretazione di quest'ultimo legame attraverso l'utilizzo delle Equazioni di Stima Generalizzate (GEE) per dati longitudinali. Alcune misure diagnostiche e metodi grafici per testare l'adeguatezza del metodo GEE saranno illustrati e utilizzati.

Key words: Italian Football championship clubs, sports and economic results, generalized estimating equations, Regression diagnostics

[1] Crisci Anna, Pegaso Telematic University, anna.crisci@unipegaso.it
Sarnacchiaro Pasquale, Univ. of Rome Unitelma Sapienza, pasquale.sarnacchiaro@unitelmasapienza.it
Luigi D'Ambra, University of Naples Federico II, dambra@unina.it

1. Introduction

Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. The main goal of any Football Championship club is to achieve sport results. Nevertheless, football has also become one of the most profitable industries, with a significant economic impact in infrastructure development, sponsorships, TV rights and transfers of players. Very informative is considered the connection between the points in the championship and the resource allocation strategies.

The Generalized Estimating Equation (GEE) methodology has been introduced to extend the application of generalized linear models to handle correlated data. For repeated measures, nowadays GEE represents a method based on a quasi-likelihood function and provides the population-averaged estimates of the parameters.

The aim of this paper is to give an interpretation of the link between the points in the championship and the resource allocation strategies using the GEE. In particular, we analyze 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 (2010-2015), by GEE for count data.

There are six sections in this study. A summary of GEE has been introduced in the second section. Section 3 deals with how to choose a working correlation structure and model selection in GEE. The Influential observation, Leverage and Outlier in GEE have been discussed in section 4. In section 5 the case study has been presented. Next, concluding remarks are presented in the final section of the paper.

2. Summary of the Generalized Estimating Equation method (GEE)

Let $\boldsymbol{y_i} = (y_{i1}, \dots, y_{it_i})'$ be a vector of responses value and let $\boldsymbol{X_i} = (\boldsymbol{X}_1', \dots, \boldsymbol{X}_n')'$ be a $t_i \times K$ matrix of covariates, with $\boldsymbol{x_{it}} = (x_{it1}, \dots, x_{itK})'$, $i = 1,2, \dots ,n$ and $t = 1,2, \dots,T$. To simplify notation, let $t_i = t$ without loss of generality.

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

$$E(y_{it}|\boldsymbol{x_{it}}) = \mu_{it}$$

Suppose that the regression model is $\eta_{it} = g(\mu_{it}) = \boldsymbol{x}_{it}^T\boldsymbol{\beta}$ where g is a link function and $\boldsymbol{\beta}$ is an unknown $K \times 1$ vector of regression coefficients with the true value as $\boldsymbol{\beta}_0$. The $Var(y_{it}|\boldsymbol{x_{it}}) = v(\mu_{it})\phi$, where v is a known variance function of μ_{it} and ϕ is a scale parameter which may need to be estimated. Mostly, v and ϕ depend on the distributions of outcomes. For instance, if y_{it} is continuous, $v(\mu_{it})$ is specified as 1, and ϕ represents the error variance; if y_{it} is count, $v(\mu_{it}) = \mu_{it}$ and ϕ is equal to 1. Also, the variance-covariance matrix for $\boldsymbol{y_i}$ is noted by $\boldsymbol{V}_i = \phi\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R_i}(\boldsymbol{\alpha})$, $\boldsymbol{A}_i = diag\{v(\mu_{i1}), \dots, v(\mu_{iT})\}$ and the so-called "working" correlation structure $\boldsymbol{R_i}(\boldsymbol{\alpha})$ describes the pattern of measures within the subjects, 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 $rp_{it} = \frac{y_{it}-\hat{\mu}_{it}}{\sqrt{v(\mu_{it})}}$ calculated from the current value of $\boldsymbol{\beta}$ (see section 4). Also, the scale parameter ϕ can be estimated by: $\hat{\phi} = \frac{1}{n-K}\sum_{i=1}^{n}\sum_{t=1}^{T} rp_{it}^2$. 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 $\boldsymbol{s}_i = (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$ with $\hat{\boldsymbol{\mu}}_i = (\mu_1, \dots \dots \mu_{iT})'$and $(\hat{\boldsymbol{\alpha}})$ is a consistent estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{D}_i' = \boldsymbol{X}_i'\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_i = diag\ (\partial_{\mu_{i1}}/\partial_{\eta_{i1}} \dots \dots, \partial_{\mu_{it}}/\partial_{\eta_{it}})$. Under mildregularity conditions $\hat{\boldsymbol{\beta}}$ is asymptotically distributed with a mean $\boldsymbol{\beta}_0$ and covariance matrix estimated based on the sandwich estimator:

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

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.

3. Criteria for choosing a working correlation structure and model selection

Unlike the GLM method, which is based on the maximum likelihood theory for independent observations, the GEE method is based on the quasi-likelihood 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 [8], and further discussed by Hardin and Hilbe [7], is analogous to the familiar AIC statistic used for comparing models fit with likelihood-based methods:

$$QIC = -2Q(\hat{\boldsymbol{\mu}}; \boldsymbol{I}) + \mathbf{2}\boldsymbol{trace}(\hat{\boldsymbol{\Omega}}_I^{-1}\widehat{\boldsymbol{V}_i^{\boldsymbol{R}}})$$

where $\boldsymbol{I}$ represents the independent covariance structure used to calculate the quasi-likelihood, $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{x}_{it}\hat{\boldsymbol{\beta}})$. The coefficient estimates $\hat{\boldsymbol{\beta}}$ and robust variance (estimator $\hat{\boldsymbol{V}}_i^{\boldsymbol{R}}$ are obtained from a general working covariance structure. Another variance estimator $\hat{\boldsymbol{\Omega}}_I$ 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. When trace $\hat{\boldsymbol{\Omega}}_I^{-1}\hat{\boldsymbol{V}}_i^{\boldsymbol{R}} \approx trace\ (\boldsymbol{I}) = K$, there is a simplified version of QIC, called QIC_u [8]: $QIC_u = -2Q(\hat{\boldsymbol{\mu}}; \boldsymbol{I}) + 2K$. 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. When using QIC and related QIC_u to compare two models, the model with the smaller statistic is preferred.

4. Regression diagnostics: Residuals, Influential and leverage points

Model checking is an important aspect of regression analysis with independent observation [9]. Unusual data may substantially alter the fit of the regression model, and regression diagnostics identify subjects which might influence the regression relation substantially. Therefore, GEE approach also needs diagnostic procedures for checking the model's adequacy and for detecting outliers and influential observations. Graphical diagnostic plots can be useful for detecting and examining anomalous features in the fit of a model to data.

Regression diagnostic techniques that are used in the linear model [3] or in GLM [4] have been generalized to GEE. Venezuela *et al.* [10] described measures of local influence for generalized estimating equations. Here, we extend such diagnostic measures of the regression model in GEE approach (Table 1). The diagnostic measures are numerous and can be classified into five groups:

a) Measures based on the prediction matrix

In GEE the Hat matrix is $\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{\frac{1}{2}}$ where $\boldsymbol{W} = diag(\boldsymbol{W}_1, \dots, \boldsymbol{W}_n)$ is a block diagonal weight matrix whose ith block corresponds to the ith subject. The leverage h_{it} as the ith diagonal element of the Hat matrix. Thus, h_{it} represent the high-leverage of ith observation y_i in determinig its own predicted value. It ignores the information contained in y. High-leverage can be rewritten by considering the Mahalonobis Distance (MD):

$h_{it} = MD_{it}^2 + \frac{1}{N}$ where $MD_{it} = \sqrt{(w_{it}^{1/2}\boldsymbol{x}_{it} - \bar{x})'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{x}_{it}' w_{it}^{1/2} - \bar{x})}$ and $\bar{x}$ is weighted mean. Cut off point: $2K/N$.

b) Measures based on residual

One method of detecting model failures is examining the residuals. There are many ways to compute residuals. The Pearson residual is a simple residual scaled by standard deviation of y_{it}. Pearson standardized residuals have been computed in order to have unit asymptotic variance. Anscombe residual have been introduced to make the distribution of the residuals as normal as possible [2].

c) Measures based on Influence Function

In this group we find CD [3, 4]; $\mathrm{SIC}_{\mathrm{it}}$ and $\mathrm{SC}_{\mathrm{it}}$.

d) Measures based on the volume of confidence ellipsoids

A measure of the influence of the ith observation on the estimated regression coefficients can be based on the change in volume of confidence ellipsoids with and without the ith observation. We consider:

a) The Andrews –Pregibon Statistic (AP) [1]: AP measures the volume of the confidence ellipsoid. It provides considerable information not only on outlying and influential observations but also on the remoteness of observations in the parameter space. We extend the AP in GEE approach. It is worth noting that this statistics does not assume the linear model and it therefore has a more general applicability. Indeed, it may be used not only as confirmative but also an exploratory tool and thus it may be extended to any data set independently of a model hypothesis.

b) The quasi-likelihood distance: Let $Q(\hat{\beta})$ is the quasi-likelihood estimate of the GEE parameters β using all response values and $Q(\hat{\beta}_{(it)})$ is the corrisponding estimate evaluate with the y_{it} observation deleted. A measure of the influence of the ith observation on $\widehat{\boldsymbol{\beta}}$ can be based on the distance between $Q(\widehat{\boldsymbol{\beta}})$ and $Q(\widehat{\boldsymbol{\beta}}_{(it)})$: $QD = 2[Q(\widehat{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}_{(it)})]$.

e) Measures based on total influence.

The overall influence [6] is based on the simple fact that potentially influential observations are outliers as either X-outliers, y-outliers, or both (see Table 1). Hadi recommends using "mean(HD_{it}^2)+ c$\sqrt{var(HD_{it}^2)}$" as a cut-off point for Hadi's measure, where c is an appropriately chosen constant such as 2 or 3.

5. Case study

The data used for our case study was obtained from the financial statements filed by the Serie A football teams. The period of study concerned the championship 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. We have started by previous paper where we selected the best model through *Cp* Mallows [5]. The independent variables considered in the final model are: Depreciation Expense of multi-annual player contracts(DEM); Revenue net of player capital gain (RNC); Net Equity (NE). Later, for this model, considering the diagnostic measures presented in section 4, we can note that the teams, Roma, Udinese and Genoa, exceed the cut of value of some measures (see table 2). In particular, the Roma team to the championship 2013-14 exceeds the cut off values related to the measures in table 2. For this reason, we consider a new GEE model with exchangeable work correlation structure (α = 0.612), without the observation related to Roma team to the championship 2013-14. The results are described in table 3.

Table 2: Diagnostic Measures

Teams	***Pearson***	***Leverage*** ***(h_{it}> 0,2454)***	***Cook-Distance*** ***(CD_{it}> 0,07272)***	***Hadi***
Roma (2013-14)	2,3079	0,3426	0,3966	0,7505 (HD^2>0,64)
Udinese (2012-13)		0,3083		0,4757 (HD^2>0,45)
Genoa (2010-11)		0,2558		

Table 3: The Poisson GEE Population-Averaged Model with Exchangeable Structure

Points	***Coef***	***St.err.***	***Z***	***P >| z|***
DEM	-0,0720	0,039	-1,84	0,066
RNC	0,3589	0,059	6,08	0,000**
NE	0,0539	0,017	3,17	0,002**
Cons	-2,2083	0,665	-3,32	0,009
Wald Stat.	93,56	Prob>chi^2 0,0000		

** significant at 5%.

Finally, we can note that the minimal working residual, computed by using correlation matrix, is obtained when we delete Roma Team. Concluding, we have discussed and reviewed the various measures which have been presented for studying outliers, high leverage points, and influential observations in the context of GEE approach. As an illustration, a data set about impact that some budget variables have on points achieved by football teams in the Serie A championship [5], has been presented, applying the methods developed in Section 4.

References

1. Andrews, D.F. and Pregibon, D. (1978). Finding the ouliers that matter. *J. Roy. Statist. Soc. Ser. B* **40** 85-93.
2. Anscombe, F. J. (1953). Contribution to the discussion of H. Hotelling's paper. *J. Roy. Statist. Soc. Ser. B* **15** 229-230.
3. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**,15–18.
4. Cook, R. D and Thomas, W. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, 76, 741–750.
5. Crisci, A., D'Ambra, L. and Esposito, V. (2018). A Generalized Estimating Equation in Longitudinal Data to Determine an Efficiency Indicator for Football Teams. Social Indicators Research. https://doi.org/10.1007/s11205-018-1891-6
6. Hadi, A.S (1992). A new measure of overall potential influence in linear regression. Computational Statistics and Data Analysis **14**, 1–27.
7. Hardin, J., & Hilbe, J. (2003). *Generalized estimating equations*. London: Chapman and Hall.
8. Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics,* **57**(1), 120–125.
9. Thomas, L.R., 1997. Modern Econometric: An Introduction. Addison-Wesley.
10. Venezuela M.K., Botter D.A., Sandoval M.C. (2007) Diagnostic techniques in generalized estimating equations, Journal of Statistical Computation and Simulation, 77:10, 879-888, DOI: 10.1080/10629360600780488

Table 1: Regression Diagnostics

Measures and Formula	***Interpretation and Cut off point***
a) Hat matrix h_{it} is the *i*-th diagonal element of the Hat matrix $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$	It allows to identify high leverage point $h_{it} \geq 2K/N$
$h_{it} = MD_{it}^2 + \frac{1}{N}$ where $\mathrm{MD_{it}} = \sqrt{(\mathrm{w}_{\mathrm{it}}^{\frac{1}{2}}\mathbf{x}_{\mathrm{it}} - \bar{\mathrm{x}})'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{x}'_{\mathrm{it}}\mathrm{w}_{\mathrm{it}}^{\frac{1}{2}} - \bar{\mathrm{x}})}$	High-leverage point can be computed by using the Mahalonobis Distance (MD) $h_{it} \geq 2K/N$
b) Residuals	
Square Pearson residuals $rp_{it}{}^2 = \sum_{i=1}^{n}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)'\boldsymbol{\Lambda}_i{}^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)$	rp_{it} is a simple residual scaled by standard deviation of y_{it} . Residuals are evalueted at the current value of $\boldsymbol{\beta}$. The matrix $\boldsymbol{\Lambda}_i{}^{-1}$ can be replaced by $\boldsymbol{V}_i$ and $\boldsymbol{R}_i$ in order to consider the correlation within subjects (working residual)
Pearson standardized residuals $(rpsd)_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{v(\mu_{it})(1 - h_{it})}}$	$(rpsd)_{it}$ is standardized in order to have unit asymptotic variance
Anscombe residuals for poisson distribution $r_{it}^A = \frac{\frac{3}{2}(y_{it}^{2/3} - \hat{\mu}_{it}^{2/3})}{\hat{\mu}_{it}^{1/6}}$	r_{it}^A: Anscombe (1953), proposed a residual using the function $G(y)$ in place of y where $G(\cdot)$ is chosen to make the distribution of as normal as possible. For univariate generalized linear models $G(\cdot)$ is given by: $G(\cdot) = \int \frac{1}{V^{\frac{1}{3}}(\mu)} \partial\mu$
c) Influence function	
Cook distance (CD) $(CD)_{it} = rpsd_{it}^2 \frac{h_{it}}{K(1 - h_{it})}$	$(CD)_{it}$ is a measure to detect clusters with a strong influence on parameter estimates $(CD)_{it} > 4/N$
$SIC_{it} = (\boldsymbol{N} - \mathbf{1})(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{x}'_{it}(\boldsymbol{y}_{it} - \boldsymbol{x}_{it}\widehat{\boldsymbol{\beta}}_{(it)})$ $SC_{it} = N(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x}'_{it}\frac{r_{it}}{1 - h_{it}}$ where $r_{it} = (\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)$	SIC_{it} and SC_{it} are easier to interpret; they are proportianal to the distance between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(\mathrm{it})}$
d) Volume of confidence ellipsoid	
Andrews/Pregibon Statistic (AP) $AP_{it} = \frac{\lvert \boldsymbol{X}^{*\prime}_{(it)}\boldsymbol{X}^{*}_{(it)} \rvert}{\lvert \boldsymbol{X}^{*\prime}\boldsymbol{X}^{*} \rvert}$	It provides considerable information not only on outlying and influential observations but also on the remoteness of observations in the parameter space. Small value of AP_{it} calls for special attention

The Quasi likelihood distance (QD): $QD = 2[Q(\widehat{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}_{(it)})]$	
e) Overall influence; $HD^2 = \frac{K}{(1-h_{it})}\frac{d_{it}^2}{(1-d_{it}^2)} + \frac{h_{it}}{(1-h_{it})}$ where $d_{it}^2 = \frac{\boldsymbol{r}_{it}^2}{\boldsymbol{r}'\boldsymbol{r}}$	HADI is based on the simple fact that potentially influential observations are outliers as either X-outliers, y-outliers, or both. $HD^2 >$ mean(HD_{it}^2)+ c$\sqrt{var(HD_{it}^2)}$