UNIVERSITÀ CATTOLICA del Sacro Cuore
Università Bocconi MILANO
POLITECNICO MILANO 1863
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
UNIVERSITÀ DEGLI STUDI DI MILANO
mathesia

Smart Statistics for Smart Applications

Book of Short Papers SIS2019

SIS 2019
SMART STATISTICS FOR SMART APPLICATIONS

SIS Società Italiana di Statistica

Editors: Giuseppe Arbia, Stefano Peluso,
Alessia Pini and Giulia Rivellini

PUBLISHED BY PEARSON

WWW.PEARSON.COM

Giugno 2019 ISBN 9788891915108

Analysis of the financial performance in Italian football championship clubs *via* GEE and diagnostic measures

Analisi delle performance finanziaria delle squadre di calcio di serie A via GEE e misure di diagnostica

Maria Kelly Venezuela[1], Anna Crisci[2], Luigi D'Ambra[2] , D'Ambra Antonello[3]

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 an interpretation of this last link using the Generalized Estimating Equation (GEE) for longitudinal data. Some diagnostic measures and simulate envelope 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, Generalized Estimating Equations, Diagnostic Measures, Simulated envelope

[1] Insper Institute of Education and Research, São Paulo, Brazil

[2] University of Naples Federico II, Department Economic, Management, Institutions

[3] University of Campania "Luigi Vanvitelli", Department of Economics

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) [5] 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 financial indicators have on points made by football teams participating in the series A championship (2010-2015), by GEE for count data.

2. Overview Generalized Estimating Equation method

Let $\boldsymbol{y_i} = (y_{i1}, \dots, y_{it_i})'$ be a vector of response values and let $\boldsymbol{X_i} = (\boldsymbol{x_{i1}}, \dots, \boldsymbol{x_{it_i}})'$ be a $t_i \times K$ matrix of covariates, with $\boldsymbol{x_{it}} = (x_{it1}, \dots, x_{itK})'$, i = 1,2, … ,n and t = 1,2, …,t_i. 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 $\eta_{it} = g(\mu_{it}) = \boldsymbol{x}'_{it}\boldsymbol{\beta}$ where g is a link function and $\boldsymbol{\beta}$ is an unknown $K \times 1$ vector of regression coefficients. 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. The variance-covariance matrix for $\boldsymbol{y_i}$ is noted by $\boldsymbol{V}_i = \phi \boldsymbol{A}_{\boldsymbol{i}}^{\frac{1}{2}} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}_{\boldsymbol{i}}^{\frac{1}{2}}$, $\boldsymbol{A}_i = diag\{v(\mu_{i1}), \dots, v(\mu_{iT})\}$ 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 $\boldsymbol{\alpha}$.
The parameters $\boldsymbol{\beta}$ are estimated by solving: $U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}'_i [V(\widehat{\boldsymbol{\alpha}})]^{-1} \boldsymbol{s_i} = 0$ where $\boldsymbol{s_i} = (\boldsymbol{y_i} - \widehat{\boldsymbol{\mu}}_i)$ with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} = (\mu_1, \dots, \dots \mu_{iT})'$and $(\widehat{\boldsymbol{\alpha}})$ 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 $\widehat{\boldsymbol{\beta}}$ is asymptotically distributed with a mean $\boldsymbol{\beta}_0$ and covariance matrix estimated based on the sandwich estimator:

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

3. Choice of best model: adjusted $\boldsymbol{R}_{\boldsymbol{adj}}^{\mathbf{2}}$ and modified Mallows' Cp ($\widetilde{\boldsymbol{C}}_{\boldsymbol{p}}$) based on Wald Statistics

In this section we discuss the use of the adjusted coefficient of determination R_{adj}^2 for the GEE and modified Mallows' Cp for the choice of one or the best subsets. We show that the adjusted $\tilde{R}_{adj}^2$ based on Wald Statistics is:

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

where

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

is the Wald statistics with the GEE robust covariance matrix estimated under the null and denoted by $\widetilde{Var}(\widehat{\beta})$ and $\boldsymbol{C}$ 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.

We select the best subset model from among the 2^K models. In order to select the best model, we consider a modified Mallows'Cp ($\tilde{C}_p$) [4] in GEE that is closely related to $\tilde{R}^2$, where:

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

later,

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

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

Finally, Cantoni *et al.* [1] have proposed an extension of the Mallows'Cp for GEE approach, by:

$$GC_p = WRSS_A - N + 2dfc$$

where $WRSS_A = \sum_{i=1}^{n}(\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_i)'\,\widehat{\boldsymbol{A}}_i^{-1}(\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_i)$ and $dfc = tr(\boldsymbol{H}^{-1}\boldsymbol{Q})$, $\mathbf{H} = n^{-1}\sum_i \boldsymbol{D}_i'\, V_i^{-1}\boldsymbol{D_i}$ and $\mathbf{Q} = n^{-1}\sum_i \boldsymbol{D}_i'\, A_i^{-1}\boldsymbol{D_i}$. The matrix $\widehat{\boldsymbol{A}}_i^{-1}$can be replaced by $\boldsymbol{R}(\boldsymbol{\alpha})$ in order to consider the within correlations.

4. Regression diagnostics and simulated envelope

Model checking is an important aspect of regression analysis with independent or dependent observations [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 [2] or in GLM [3] have been generalized to GEE. Venezuela *et al*. [6] described measures of local influence for generalized estimating equations. Here, we extend the diagnostic measures based on Cook distance, leverage and standardized residuals, of the regression model in GEE approach. Moreover, in order to identify possible outlier observations in the dataset and examine the adequacy of the fitted model, a suggestion is to plot the lth-ordered absolute values of the Pearson standardized residuals calculated from the fitted model to the dataset.

5. Results

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 some financial indicators have on the points achieved by football teams.

We consider the following independent variables: Wage(W), Depreciation Expense of multi-annual player contracts (DEM), Revenue net of player capital gain(RNC), Net equity(NE). In addition, we have considered, on the bases a bivariate descriptive analysis, also the square effect of DEM (DEM^2), given the non- linear relationship between Point and DEM. Finally, the interaction between DEM and NE (DEM*NE) also was considered. We consider the QIC criterion in order to select the best working correlation structure among three structures: independent, exchangeable and AR-1. The results showed values of the QIC for these correlation structures very similar to each other, even though the AR1 structure was slightly smaller, it generates some difficulties in the choice of the best working correlation. Later, we have carried out a descriptive analysis for the within-subject correlation of variable Point by Year. We can note that the correlations are decreasing by Year and this would lead to choice of the AR-1 working correlation. Now, let's begin the choice of the best subsets using the criteria described in Sect. 3. In particular we selected the best subset within 2^K models. The choice falls on Model describe in table1, whose $\tilde{\boldsymbol{C}}_{\boldsymbol{p}}$ is close to the number of variables (5.34) and lower GC_p by Cantoni

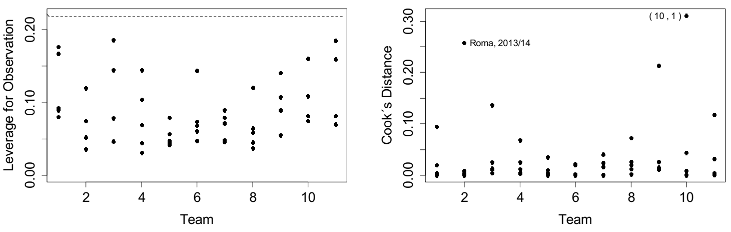
(35,67). Moreover, it shows the best R^2_{adj} (0.55) and the lowest *QIC* (87.904). The Table 1 shows the output of this model. We have all significant variables, with the most important variable is DEM.

Table 1. Coefficient estimates of the Poisson regression model using AR(1) structure and with all observations.

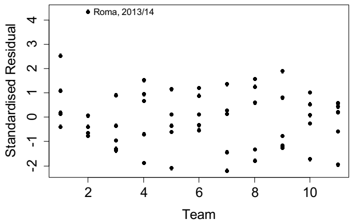
Coefficients	Estimate	Robust S.E.	Robust z	p.value
Intercept	-21.069	11.611	-1.815	0.0696
DEM	3.074	1.330	2.311	0.0208
RNC	0.278	0.083	3.371	0.0007
NE	-0.718	0.196	-3.659	0.0003
DEM^2	-0.114	0.041	-2.804	0.0051
DEM*NE	0.045	0.012	3.836	0.0001
Correlation	0.347			

Wald statistic 71.08, Prob > Chi-square 0.000

Figure 1 shows the diagnostic measures using AR1 structure. In Figure 1(a), there is no observation playing leverage role in matrix of covariates. Figure 1(b) shows two observations as influence observations by Cook's distance, but one of them, that is the Roma team to the championship 2013/14, is also considered an outlier by the standardized residual in Figure 1(c).



(a) (b)



(c)

Figure 1. Diagnostic measures for the Poisson regression model using AR(1) structure and with all observations.

The half-normal probability plot with simulated envelope (Figure 2) indicates a good fit with exception of one observation (Roma team to the championship 2013/14) that is outside the simulated envelope. Therefore, it can be concluded that the Poisson regression with AR1 correlation structure is adequate to explain the relation between the points achieved by football teams and financial variables.

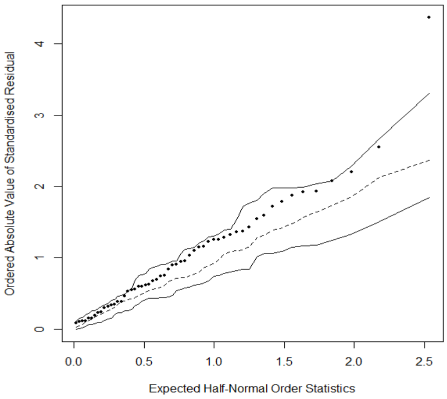


Figure 2. Half-normal probability plot with simulated envelope for the Poisson regression model using AR(1) structure and with all observations.

Just to be sure that, without the observation related to the Roma team for the

2013/14 championship, the fitted model would still be appropriate with the other observations, we consider a new estimation of the previous gee model with AR1 work correlation structure. Table 2 presents the new estimates of the coefficients described in the model, but without the observation related to Roma team to the championship 2013/14. Analyzing the results of this table, the same conclusion is maintained since all coefficients remain significant (p.value < 0.05). Then, the Roma team for the 2013/14 championship, there is not an influential point.

Table 2. Coefficient estimates of the Poisson regression model using AR(1) structure and without the observation related to Roma team to the championship 2013/14.

Coefficients	Estimate	Robust S.E.	Robust z	p.value
Intercept	-21.024	11.101	-1.894	0.058
DEM	2.967	1.294	2.293	0.022
RNC	-0.111	0.041	-2.744	0.006
NE	0.342	0.056	6.056	0.000
DEM^2	-0.711	0.208	-3.416	0.001
DEM*NE	0.044	0.012	3.562	0.000
Correlation	0.315			

References

1. Cantoni E, Flemming J, Ronchetti E. Variable selection for marginal longitudinal generalized linear models. Biometrics 2005;61(2):507–14.
2. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**,15–18.
3. Cook, R. D and Thomas, W. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, 76, 741–750.
4. 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
5. Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22.
6. 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