Using Quantile Regression to Examine Pay and Performance on Advanced Metrics in Major League Soccer

Clinton Warren, Illinois State University

Finance/Economics - Economics (Professional Sport)  Saturday, June 3, 2017
20-minute oral presentation (including questions)  3:10 PM
Abstract 2017-414  Room: Oxford

Soccer is on the verge of an analytics revolution that is impacting clubs worldwide. With the volumes of data collected by companies like Opta and ProZone leagues, clubs, and media have begun analyzing and utilizing soccer data in decision making in a manner previously unimagined (Wahl, 2015). Specifically, Major League Soccer (MLS) has provided brief reports on the usage of performance analytics by clubs such as New York Red Bulls, Portland Timbers, and expansion Atlanta United FC (Baer, 2016). Additionally, academic literature has begun to examine and develop advanced performance metrics in soccer (Auer & Hiller, 2015). This study seeks to provide an exploratory examination of a series of emerging player performance metrics and their impact on salary in MLS.

A large body of research exists examining the determination of player earnings in professional sport (Berri & Jewell, 2004; Bucciol, et. al., 2014; Gupta, et. al., 2012; Shaw, 2014; and Vincent & Eastman, 2009). Within this area of scholarly focus, a number of papers have specifically examined pay in professional soccer. In a study of pay dispersion in German soccer, Franck and Nüesch (2011) found very high and very low levels of income inequality seemed to positively impact team performance. However, in a study of Italian soccer, Bucciol, et. al. (2014) found pay dispersion had a negative effect on a team’s ability to win matches. Deutscher and Büschemann (2016) found players exhibiting volatile performance, rather than consistent performances, were paid a premium in the German Bundesliga. While Hall, Szymanski, and Zimbalist (2002) found higher team payrolls were causal of team performance success in English soccer. This study seeks to extend the examination of pay and performance to MLS.

Historically, the modeling of earnings in professional sport has utilized a log-linear, ordinary least squares (OLS) regression equation in which typical measures of player performance are set as explanatory variables of the natural log of player salary (nLOGSAL). However, Leeds (2014) states quantile regression (QR) is an important alternative to OLS and sport economists should make frequent use of the analytic technique. Specifically, QR allows for the study of conditional salary distributions rather than simply means and provides a way to check for errors of heteroscedasticity. Furthermore, Leeds (2014) suggests QR is underutilized in sport economics, and is particularly relevant in attempts to analyze concepts like bargaining power. This study seeks to understand if players are paid at a rate unusually higher or unusually lower than their counterparts based on their performance on advanced performance metrics. As such, QR is the favored approach to analyze these data. However, as the first study to take this approach to player performance on advanced metrics in MLS, an OLS regression will be conducted and compared with the QR results.

In soccer, common player performance metrics that may be predictive of salary are goals (G), assists (A), and minutes played (MIN). This paper advances the study of earnings in soccer by including the additional metrics of shots taken (S), expected goals (xG), expected assists (xA), unassisted goal percentage (uAGP), key passes (KP), and touch percentage (TP). The OLS regression equation including these independent variables with nLOGSAL set as the dependent variable is represented as follows:

\[ n\text{LOGSAL} = \alpha + \beta G + \beta A + \beta \text{MIN} + \beta S + \beta \text{xG} + \beta \text{xA} + \beta \text{uAGP} + \beta \text{KP} + \beta \text{TP} + \varepsilon \]

To utilize QR, this study expands the above equation according to the specifications identified by Leeds (2014) and models the data across salary distributions based on quantiles. In so doing, this paper is better able to analyze MLS club proficiency in investing in traditional and advanced performance metrics at various levels of roster construction. Using terminology from Leeds (2014), this allows for a better analysis of players who are paid “unusually high” or “unusually low” based on their production. The QR model specifies that the 0th quantile of the conditional distribution of nLOGSAL as a linear function of each independent variable. As such, the QR equation in this study
takes the following form:

$$Q_0(nLOGSAL) = \alpha + \beta(0)G + \beta(0)(0)A + \beta(0)GP + \beta(0)MIN + \beta(0)S + \beta(0)xG + \beta(0)xA + \beta(0)uAGP + \beta(0)KP + \beta(0)TP + \varepsilon(0)$$

This study utilizes a data set of 437 MLS out field players and their performance and salary statistics from the 2015 league season. Salary data were obtained from the MLS Player's Union and performance data were obtained from an open source data file made available by American Soccer Analysis.

Results of this study indicate a series of important findings. First, the QR model outperformed the OLS regression. Second, there were significant differences in pay on the advanced performance metrics included in the study at each quantile. The OLS regression equation was significantly predictive ($p < .001$) of MLS player salary with approximately 32.6% of the variance explained by the model. However, of the ten variables included in the model only G ($p = .024$) and TP ($p < .001$) were significant predictors. Holding all else constant, a one G increase yielded a 6.98% increase in player salary and an increase of one TP led to a 12.34% increase in salary. Of the other independent variables in the model, only MIN ($p = .096$) approached significance. Results of the QR equation indicated a number of nuanced differences in player salary in the conditional distributions analyzed. First, the 10th quantile indicated MIN ($p = .037$), KP ($p = .002$), and TP ($p = .002$) were significant predictors of nLOGSAL. An increase in TP yielded the largest increase in pay (2.16%) in this quantile. Second, the 25th quantile showed xG ($p = .014$), KP ($p = .038$), and TP ($p = .050$) were significant predictors. Both xG (1.24%) and TP (1.88%) indicated meaningful increases in player salary in this conditional distribution. Third, the 50th quantile indicated MIN ($p = .001$), KP ($p = .029$), A ($p = .023$), TP ($p = .003$) were significantly predictive of salary. In this quantile an increase in TP led to a 5.39% increase in salary. The other significant predictors yielded less than a 0.5% change in salary. Fourth, the 75th quantile yielded two significant predictors of salary: G ($p = .011$) and TP ($p < .001$) with both variables impacting salary positively by more than 11%. Finally, the 90th quantile indicated only MIN ($p = .003$) and TP ($p < .001$) were significant predictors of nLOGSAL with an increase in TP leading to a 26.37% increase in salary.

In all, results of the QR model indicate TP is the most impactful player performance metric across the conditional distributions of salary. This study contributes to the body of literature in the field of pay and player performance by utilizing QR to examine pay and performance in MLS. Further, this is the first paper to analyze advanced performance metrics in MLS while providing an initial explanation of how individual clubs are compensating players for their performance on those metrics.