Modeling Major League Soccer Salary Data

Nathán Goldberg
Harvard College

May 2017

1 Research Question

What role, if any, do performance metrics play in apportioning salaries to players in Major League Soccer? What role *should* they play?

2 Background and Motivation

In their 2013 book about statistics in soccer, *The Numbers Game*, [9] Chris Anderson and David Sally detail several ways in which soccer has been more reluctant than other sports to embrace the tools of statistical analysis. The most important reason is perhaps the disdain for numbers shared by many coaches and front offices, who would rather rely on "conventional wisdom" in their decision-making process.

I believe that the prospect of saving money would be enough to convince front offices to adopt data-driven approaches. An analysis of salary data in Major League Soccer is ideal for this purpose because the data is publicly available and American sports are generally receptive of the use of statistics as a competitive advantage. Given the marginal role of statistical analysis in soccer, there is reason to believe that salaries are apportioned based on "conventional wisdom" proxies for performance rather than on-field performance metrics. This paper aims to identify the inefficiencies in the MLS salary market, allowing teams to exploit them to their advantage.

3 Understanding the Data

The data and salary rules in this paper refer to the 2016 MLS season. Most variables were collected or updated during the international break of June 2016, meaning the analysis excludes players who joined the league during the summer transfer window. The only variable in the analysis that directly measures on-field performance, MLS Fantasy Points, were collected at the end of the season.

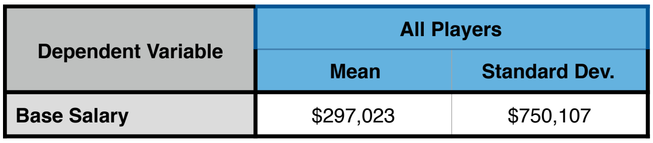
3.1 MLS Salary Rules

MLS has many convoluted salary rules, but our analysis will focus only on these three:[1] [2]

- *Salary Cap*: Each club is allowed to pay at most **$3,660,000** in salaries. The number itself is not as important as the understanding that there is an incentive for teams to pay their players efficiently. There are anywhere between **18** and **30** players on each roster.
- *Designated Players (DPs)*: Each club can register **3** DPs, who count at most **$457,500** against the salary cap, regardless of their actual salary. Many clubs reserve these spots to lure international stars that will promote the club's global brand, sell jerseys, and fill stadiums, although there are other reasons why a club would choose to make someone a DP.
- *Minimum Salary*: The minimum salary is **$62,500**, although several rules allow clubs to pay players below the minimum. For our analysis, we will assume everyone earns at least the minimum.

3.2 MLS Salary Data (Courtesy of the MLS Players' Union)

The dataset used for this analysis comes from the May 2016 salary release from the MLSPU [3]. After cross-referencing the salary dataset with the MLS Fantasy dataset, there are **545** players in the complete dataset. The distribution of salaries is very heavily right-skewed, in large part because of the high DP salaries (Fig. 1).



Dependent Variable	All Players	
	Mean	Standard Dev.
Base Salary	$297,023	$750,107

Table 1: Summary Statistics for Base Salary

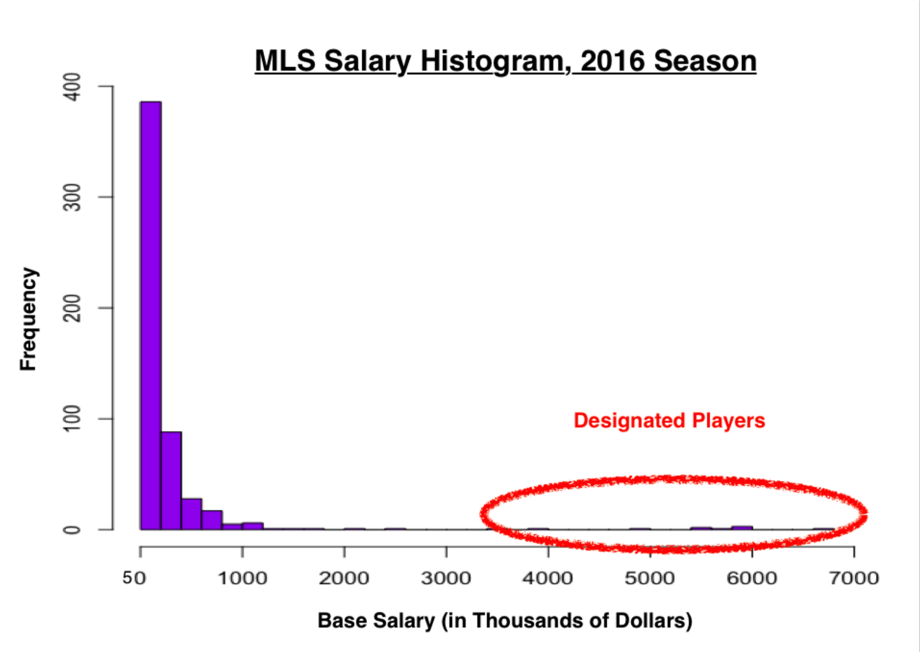
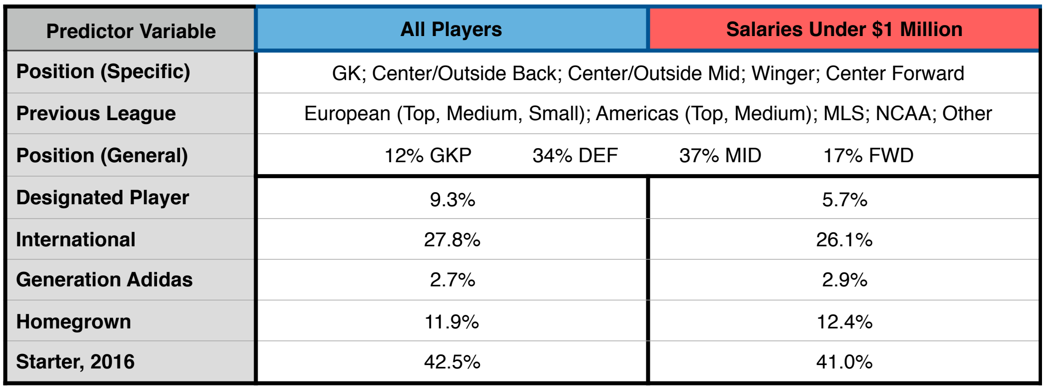


Figure 1: Distribution of salaries is extremely right-skewed

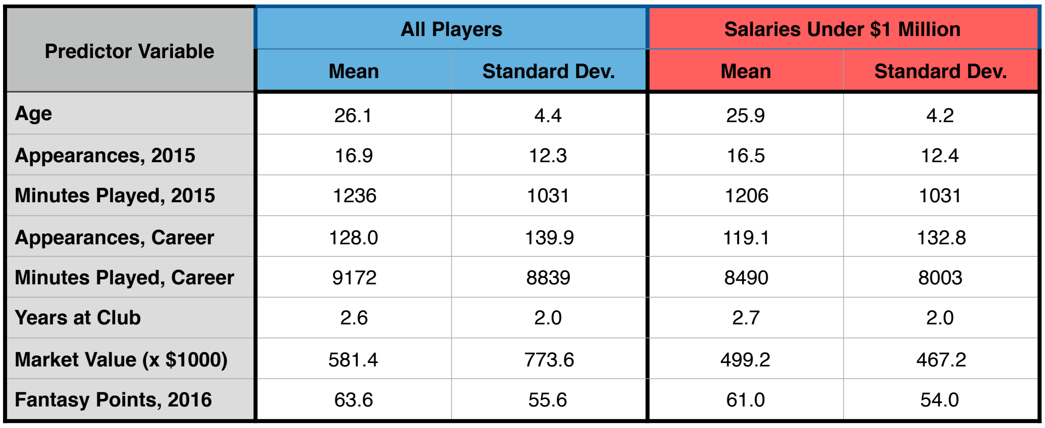
3.3 Available Predictor Variables

Most data points were collected manually from soccer websites like MLS Soccer [4], Transfermarkt [5], and WhoScored [6]. Official MLS Fantasy Points for the 2016 season come from an open-source GitHub repository of MLS data [7]. For a description of all the variables, refer to Appendix A.



Predictor Variable	All Players	Salaries Under $1 Million
Position (Specific)	GK; Center/Outside Back; Center/Outside Mid; Winger; Center Forward	
Previous League	European (Top, Medium, Small); Americas (Top, Medium); MLS; NCAA; Other	
Position (General)	12% GKP 34% DEF 37% MID 17% FWD	
Designated Player	9.3%	5.7%
International	27.8%	26.1%
Generation Adidas	2.7%	2.9%
Homegrown	11.9%	12.4%
Starter, 2016	42.5%	41.0%

Table 2: Summary Statistics for Categoric Data



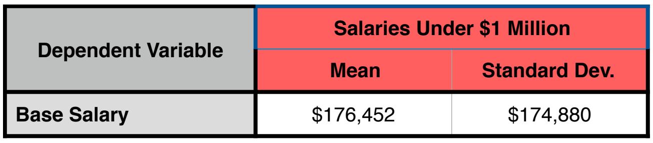
Predictor Variable	All Players		Salaries Under $1 Million	
	Mean	Standard Dev.	Mean	Standard Dev.
Age	26.1	4.4	25.9	4.2
Appearances, 2015	16.9	12.3	16.5	12.4
Minutes Played, 2015	1236	1031	1206	1031
Appearances, Career	128.0	139.9	119.1	132.8
Minutes Played, Career	9172	8839	8490	8003
Years at Club	2.6	2.0	2.7	2.0
Market Value (x $1000)	581.4	773.6	499.2	467.2
Fantasy Points, 2016	63.6	55.6	61.0	54.0

Table 3: Summary Statistics for Numeric Data

4 Research Design

4.1 Removing Designated Players

Because a large part of the skewness of the salary data comes from high-paid DPs, I removed all players earning a base salary above **$1 million**. This decision is justified, I believe, because DP salaries do not necessarily represent inefficiencies within the salary cap system, since their salaries do not count fully against the cap. More importantly, the main purpose of many DPs is not as much to perform well on the field as it is to sell jerseys and to promote the global brand of the club. For that reason, their high salaries might be justified by their ability to generate revenue off the field. Finally, I chose a salary cut-off as opposed to simply removing all DPs because not all DPs are household names who fill stadiums; there are other reasons for clubs to register a DP, such as the prospect of making a larger profit on a future sale to an overseas club. DPs who earn under $1 million are thus still expected to justify their salaries with their on-field performance. After removing all players above the salary cut-off, there are **525** players in the final dataset.



Dependent Variable	Salaries Under $1 Million	
	Mean	Standard Dev.
Base Salary	$176,452	$174,880

Table 4: Summary Statistics for Base Salary, Players Earning Under $1 Million

4.2 Choosing a Model

Here are several types of models, along with their disadvantages, that I considered before settling on the final version:

1. A Random Forest model on the log-transform of base salary with all available predictor variables explains a high percentage of the variability in salaries. However, I learned from working directly with an MLS front office that the idea of following a series of nodes to determine a salary was not as intuitive as piecing a salary together from all its components, as is the case with (generalized) linear models.

2. An Ordinary Linear Regression model has a high Adjusted R^2, but even the distribution of the log-transform of the salaries below $1 million is heavily right-skewed, which would violate one of the assumptions of simple linear regression (Fig. 2).

3. A Gamma model, a kind of generalized linear model (GLM) that deals well with skewed data, fit well by measure of its residual deviance, a goodness-of-fit metric that extends the sum of squares metric to models that determine fit based on maximum likelihood estimation instead of linear regression [8]. However, a likelihood ratio test shows that a Generalized Additive Model, a different kind of GLM that can also model skewed data, is a better predictor than the Gamma model.

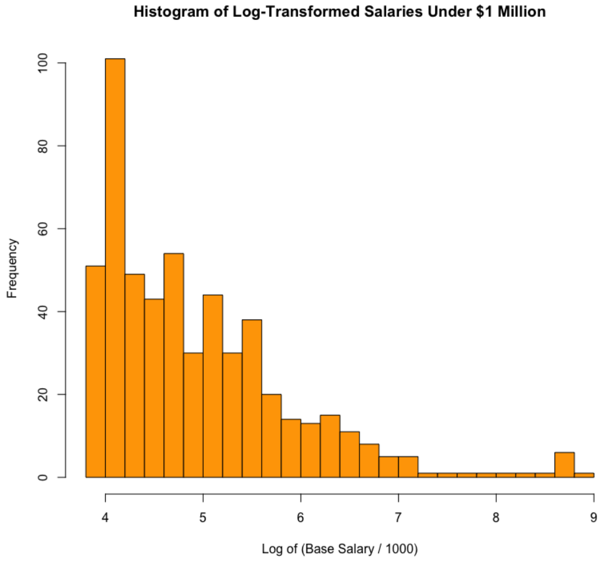


Figure 2: Distribution of log transformed salaries is still heavily right-skewed

4.3 Generalized Additive Model

The GAM is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions (s_i) of the predictor variables (X_i). A link function (g) relates the dependent variable ($E(Y)$) to the predictor variables. In this particular case, the link function used is the log-link function. Apart from being one of the most common link functions, it allows for higher variability among larger observations, which suits a right-skewed analysis well. The smooth functions of the predictor variables are estimated non-parametrically, which means they are generated from the data itself and not built to fit any specific family (e.g. quadratic or logarithmic) [11]. For our GAM with a log-link function, where Y is the log-transform of salary and X_i is the vector of predictor variables, the relationship between X and Y is expressed below:

$$log[E(Y|X_1, X_2, ..., X_n)] = \beta_0 + s_1(X_1) + s_2(X_2) + ... + s_n(X_n)$$

The two main advantages of the GAM are that it can handle skewed data and that the contributions from the predictor variables to the linear predictor do not need to be linear themselves. For example, the relationship between Market Value and its contribution to the linear predictor is better captured by a smooth function than it is by a straight line (Fig. 3).

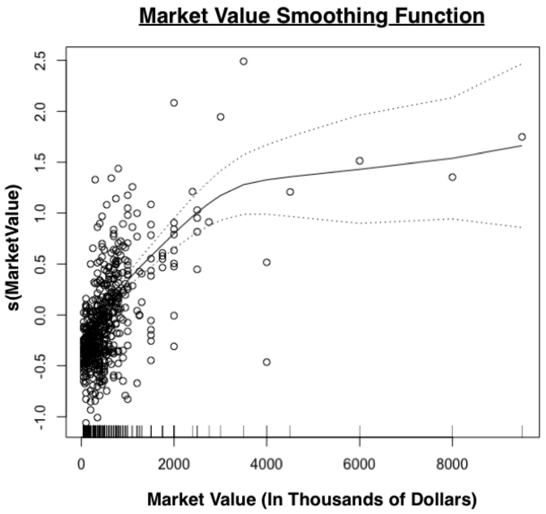


Figure 3: Smoothing function shows how contributions to predictor variable need not be linear.

5 Analysis of Results

Using a GAM with the log-link function, I fit a model using only non-performance based predictor variables (i.e. every variable in the dataset except for Fantasy Points)[1]. After performing model selection steps to only keep the variables that significantly improved the model's predictive ability, the following eight variables (or functions of variables) are significant predictors:

- Position (General)
- Generation Adidas
- International
- Designated Player
- Previous League
- s(Years at Club)
- s(Minutes, Career)
- s(Market Value)

Perhaps surprisingly, a GAM with only these eight variables as predictors of base salary fit very well according to two measures of fit:

1. The ratio of its residual deviance to its degrees of freedom is much lower than 1 ($RD/DF << 1$). The rule of thumb is that this indicates a good fit [10]
2. The Adjusted R^2 of a regression between the fitted and actual values is **61.6%** (Fig. 4)

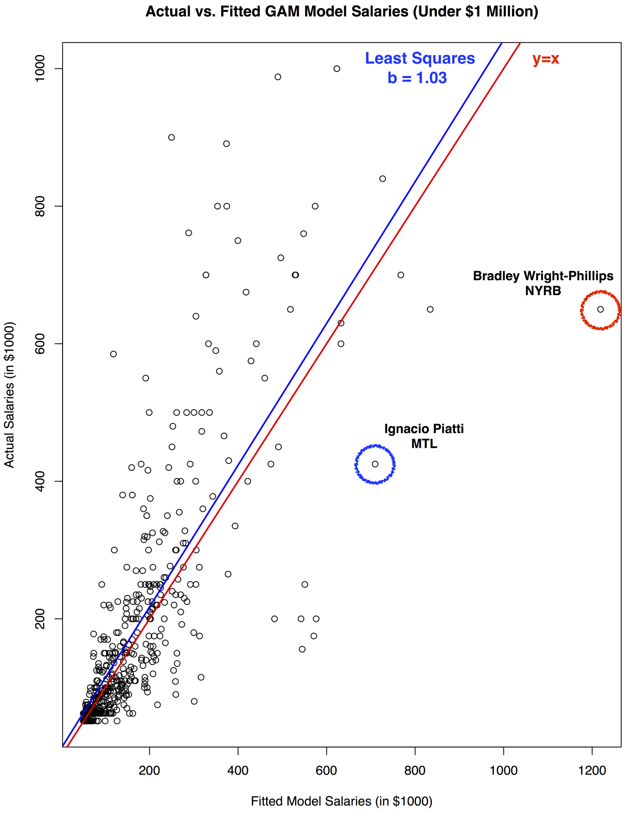


Figure 4: Blue line shows Least Squares Regression line; Red line shows y=x

Here is the upshot from this section: If we can create a truly well-fitting salary model without any direct on-field performance metrics, then that means MLS teams are not paying players according to their on-field performance. Instead, they are paying players according to "conventional wisdom" proxies for on-field performance, such as Market Value, the Previous Leagues their players played in before joining their current club, and whether or not their players are International.

[1] Although Market Value could be considered a proxy for performance to some extent, it is still not a *direct* on-field performance metric.

6 Fantasy Points-Above-Replacement Model

Since we have determined that MLS salaries can be predicted relatively well without any on-field performance metrics, there is clearly an opportunity to improve efficiency in salary practices for any team that wants to take a data-driven approach. The model I propose here apportions salaries solely according to on-field performance metrics, in the form of official MLS Fantasy Points, through a Points-Above-Replacement system.

6.1 Building a Basic PAR Model

This particular PAR Model can be constructed in six steps:

1. Find salary S_d available for distribution
   $S_d = S_{total} - (\#Players) * (S_{min})$
2. Normalize Fantasy Points FP across positions so no position is favored over another
   $FPN_i = FP_i * (\bar{FP} \div \bar{FP}_{POSi})$
3. Define Replacement Rate R_r as % of Players at S_{min}
   $R_r = (\#Players_{Smin}) \div (\#Players_{total})$
4. Define Replacement Fantasy Points FP_r as FP at percentile R_r
5. For each player i, find Fantasy Points Above Replacement PAR_i
   $PAR_i = FPN_i - FP_r$
6. For each player i, calculate PAR Salary S_{PARi}
   $S_{PARi} = S_{min} + (S_d \div \sum_{n=1}^{526} PAR_k) * PAR_i$

6.2 Individual Example: Bradley Wright-Phillips

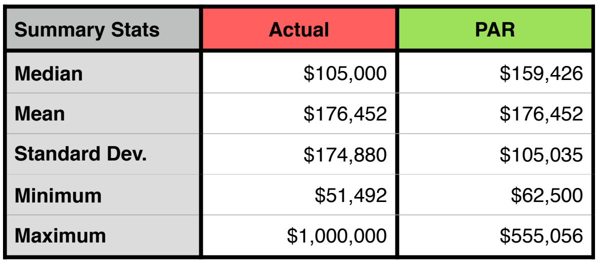
In order to more easily conceptualize all the steps, we will see how the model determines the salary of Bradley Wright-Phillips, the New York Red Bulls forward with the highest PAR Salary:

1. $S_d = \$92,637,149 - (525) * (\$62,500) = \$59,824,649$
2. $FPN_{BWP} = 223 * (60.98 \div 52.55) = 258.78$
3. $R_r = 92 \div 525 = 17.5\%$
4. $FP_r = FP$ at the $17.5th$ Percentile $= 1.82$
5. $PAR_{BWP} = 258.78 - 1.82 = 256.96$
6. $S_{BWP} = \$62,500 + (S_d \div 31,209.79) * 256.96 = \mathbf{\$555,055.50}$

Bradley Wright-Phillip's actual base salary is **$650,000**, and his position on Figure 5 is highlighted in red.

6.3 PAR Model Fit

The Points-Above-Replacement model and the actual salaries have an equal mean that is higher than their medians, but the model has a higher median than the actual salaries, meaning that although the distribution of PAR salaries is still right-skewed, it is less skewed than the original distribution. The PAR model also shows a narrower range, which is just over half as wide as the original.



Summary Stats	Actual	PAR
Median	$105,000	$159,426
Mean	$176,452	$176,452
Standard Dev.	$174,880	$105,035
Minimum	$51,492	$62,500
Maximum	$1,000,000	$555,056

Table 5: Summary Statistics for Actual Base Salaries and PAR Model

The only comparable goodness-of-fit measure between the non-performance based GAM and the Points-Above-Replacement model is the Adjusted R^2 of the regression between the fitted PAR and actual values, which in this case is **15.2%** (Fig. 5)

We can see from the plot below and the low Adjusted R^2 value that there seems to be little correlation between actual and PAR salaries. If we believe Fantasy Points to be a good measure of on-field performance (which is by no means a given), we come to the same conclusion as before: MLS teams in general do not pay players according to their on-field performance. This model offers a first step for those teams interested in tying salaries to performance more closely.

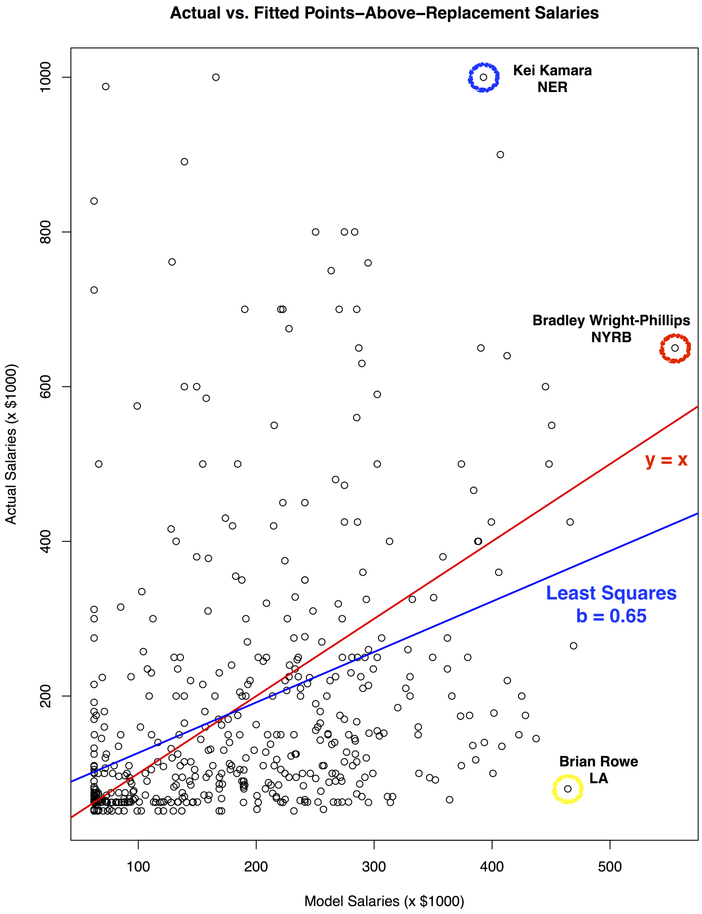


Figure 5: Blue line shows Least Squares Regression line; Red line shows y=x

7 Uses, Limitations, and Extensions

7.1 Potential Uses

After showing that salaries in MLS are determined largely by "conventional wisdom" proxies for player performance, I hope my analysis serves the purpose, above all, of convincing teams across the league that there is an opportunity to gain a competitive advantage by valuating players according to their on-field performances.

Beyond that, my basic PAR model offers a simple blueprint upon which to build more complex models of player valuation based around on-field performance. These models can provide an extra dimension of consideration to the player scouting process, as well as a rational way to restructure each team's payroll and perhaps save some money. Additionally, a comparison of actual and PAR salaries can identify options for arbitrage that can make teams more efficient with their funds in the transfer market (Fig. 6). It is worth mentioning that clubs do not have to pay players the full extent of their model salary; they only have to pay them more than any other team would offer!

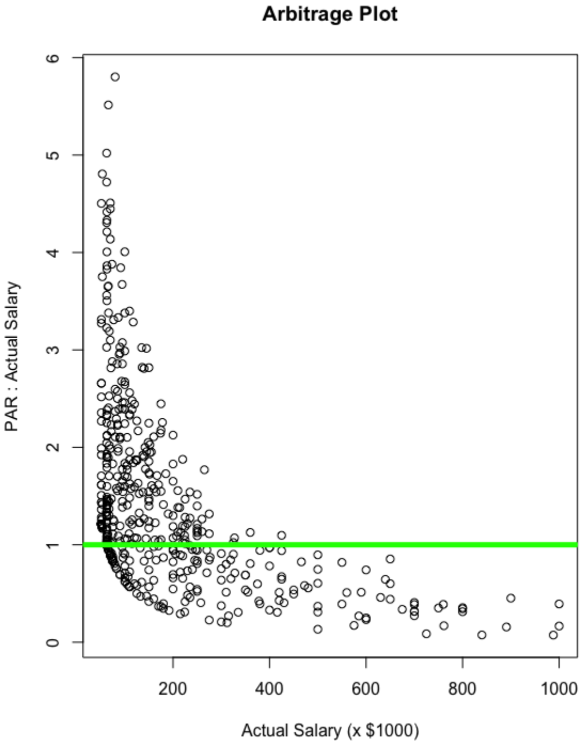


Figure 6: Points above the green line are players who have lower actual than PAR salaries

7.2 Limitations and Possible Extensions

Because the collection of soccer data was done by hand and was very intensive, this analysis focused only on 2016 data. More years' worth of data would make for a more robust analysis.

Because detailed on-field performance data in soccer is expensive and not readily available, the only direct performance metric available were Fantasy Points, which are not necessarily the best way to measure performance. Access to other kinds of data would make it possible to build other models that do not rely on Fantasy Points. Furthermore, the basic PAR model assumes that all Fantasy Points are equally valuable, but it is possible to explore relationships between salary and Fantasy Points that are not linear.

However, perhaps the biggest limitation is that, to my knowledge, there is no robust data on the correlation of any performance metrics to win percentage, so even with access to other types of performance data, we would need an analysis of which metrics are actually relevant. That way, instead of a Fantasy Points-Above-Replacement model, we can build a Wins-Above-Replacement model that has a much wider scope.

Finally, while I have argued that teams would be smart to place more emphasis on on-field performance metrics, a good model might incorporate both on and off-the-field data, especially in order to make predictions about new players as opposed to prescriptions on known players. For example, a player's age might factor into how well he is expected to perform, and so developing a model that could account for that as well would be ideal.

References

[1] https://www.mlssoccer.com/post/2016/03/04/major-league-soccer-releases-2016-roster-and-compet

[2] https://www.mlssoccer.com/league/official-rules/mls-roster-rules-and-regulations.

[3] https://www.mlsplayers.org/salary_info.html.

[4] https://www.mlssoccer.com/rosters/2017.

[5] https://www.transfermarkt.com/major-league-soccer/startseite/wettbewerb/MLS1.

[6] https://www.whoscored.com/Regions/233/Tournaments/85/Seasons/6137/Stages/13276/PlayerStatistics/USA-Major-League-Soccer-2016.

[7] https://github.com/coffenbacher/mls-data.

[8] https://en.wikipedia.org/wiki/Deviance_(statistics).

[9] Chris Anderson and David Sally. *The Numbers Game.* Penguin, 2013.

[10] Mark Glickman. Stat 149 lecture.

[11] Huimin Liu. Generalized additive model, December 2008.

A Predictor Variable Descriptions

1. *Positional Group:* Goalkeepers, defenders, midfielders, and forwards are assigned to their positional group by the MLS Players Union, which releases the salary data.

   - *Position:* Furthermore, each player is coded by his specific position on the field (e.g. *Right Back*, *Center Midfielder*) as specified on their profile on the soccer website Transfermarkt

2. *Previous League:* The league that the player last played in before joining the team they currently play in. For simplicity, leagues are grouped together based on strength, so players coming from the first divisions of Spain, Germany, and England, for example, are all coded as *Europe-Top*. Players coming from college are coded as *NCAA*.

3. *Designated Player:* Binary variable indicating whether or not each player is under a DP contract.

4. *Generation Adidas:* Binary variable indicating whether or not each player is under a Generation Adidas contract. Generation Adidas is a program, sponsored by Adidas, which rewards a handful of college soccer standouts each year with professional contracts that are paid by the company, as opposed to by the club that signs them.

5. *Homegrown:* Binary variable indicating whether or not each player is under a Homegrown contract. Homegrown players are those that played for the club's youth academy before signing a professional contract, and there are rules that allow teams to pay several Homegrown players under the league's minimum salary.

6. *International:* Binary variable indicating whether or not each player is an international player (i.e. not an American or Canadian citizen).

7. *Starter (2016):* Binary variable indicating whether a player was considered a starter for his team at the halfway mark of the 2016 season. All players who had played more than half of the available minutes for their before the June international break are considered starters.

8. *Age:* A players age in years as of June 1st, 2016.

9. *Market Value:* Market value in thousands of dollars as indicated by Transfermarkt.

10. *Appearances (2015):* Number of appearances made by each player during the previous calendar year, regardless of team.

11. *Minutes (2015):* Number of minutes played by each player during the previous calendar year, regardless of team.

12. *Appearances (Career):* Number of appearances made by each player during their entire career, regardless of team.

13. *Minutes (Career):* Number of minutes played by each player during their entire career, regardless of team.

14. *Years at Club:* Number of years player has spent at the club for which he is currently contracted.

15. *Fantasy Points-Above-Replacement (2016):* The number of fantasy points scored by each player above the replacement level during the 2016 season, normalized by positional group.