English Premier League Multiple Regression Analysis

Andrei Velasevic

21/12/2020

Abstract

In the following report, an analysis was performed on data from the 2019/20 English Premier League season. The goal of the report was to see if any conclusive mutliple linear regression model can be fit to predict the rank of a midfielder in the league, given various predictors. A model was found by the BIC stepwise regression method to have five statistically relevant predictors. Afterwards, a look was taken at correlations between variables and multiple conclusions were made about the model.

Key Words

English Premier League (EPL): England's top tier soccer league, where matches last a total of 90 minutes in regulation.
Midfielder: A position in soccer in which the player is positioned in the middle of the field and connects the defense with the offense in progressive play.

Introduction

The English Premier League (EPL) which is arguably the most competitive soccer league in the world generates billions of dollars of revenue annually and is the highest viewership of any top tier soccer league. It consists of many of the world's best players and managers competing in a 38 game long season.

The goal of this specific study is to report conclusions that can be drawn from analyzing data from the EPL. Specifically, a multiple linear regression is used to predict the rank of a midfield player in the league, based on various predictors taken from the league, such as (goals scored, assists, clean sheets, etc...). The goal is also to see which predictors are more significant in predicting the rank compared to others.

The data set that this report is referring to is an observation of the season-long information on players and their performance in the English Premier League. A comma separated values table of players and their performance in numerous categories will be assessed and analyzed. A general model will be fit with all potential predictors, following will be a method of stepwise regression to deduce a single model. Finally correlations will be looked at of the final model's predictor variables.

Methodology (Data and Model)

Data:

The data presented in this report draws from real data taken from footystats.com based on the results of the 2019/20 season of the English Premier League. The data is all based on observations of players and their results/performances in the whole season. The data sets are formatted as comma separated files in excel, with the original set of players consisting of 615 observations of 46 variables. The data set contains

all players registered in the league from the 2019/20 season, including all positions and players who did not play in the Premier League (but perhaps played in other leagues). In this report, the main focus is players who contributed to the English Premier League. The population of estimation is all midfielders in the EPL, the sample is the players who were registered and competed in the 2019/20 season. A big strength of the data set is that there is relevant data for every player and there are no issues with accessibility. All of the data was observed and formatted in a csv file.

Upon extracting the data from the website, it was modified to have less irrelevant variables and certain rows removed (due to N/A information). From the original data set, rows were removed of players who were not midfielders as the main focus of this report is analysis of midfielders and how they ranked. Also, midfielders who had *-1* as the input for their rank were also removed (this is due to players who were registered, but did not play). Finally, whole columns were removed that did not provide relevant information to this topic. A list of the removed columns is provided in the appendix.

Table 1 shows a preview of the data set used in the analysis. In the preview, only the first six players are shown with the first eight observable variables.

Table 1: Midfielder Data

full_name	age	birthday	league	season	position	Current.Club	minutes_played_overall
Aaron Lennon	33	545529600	Premier League	2019/2020	Midfielder	Burnley	485
Aaron Mooy	30	653356800	Premier League	2019/2020	Midfielder	Brighton & Hove Albion	2090
Aaron Wan-Bissaka	22	880502400	Premier League	2019/2020	Midfielder	Manchester United	3071
Abdoulaye DoucourÃ©	27	725846400	Premier League	2019/2020	Midfielder	Watford	3166
Adam David Lallana	32	579225600	Premier League	2019/2020	Midfielder	Liverpool	373
Adama TraorÃ© Diarra	24	822528000	Premier League	2019/2020	Midfielder	Wolverhampton Wanderers	2605

Model:

The model used to perform an analysis on the effects of numerous variables on the rank in the league of the midfielder is a multiple linear regression model (MLR). A regression analysis in the form of MLR has a model in the following form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

Where Y represents the variable to be predicted (rank in league), β_0 is meaningless since a rank given no predictors cannot be explained reasonably. Coefficient betas represent the relative change in rank with respect to x_p which is the predictor variable. ϵ represents the error of the model, in other words the difference between the predicted values of Y and the observed values. The reason for selecting this model is that all the variables chosen to predict rank are numerical and are statistically relevant in predicting rank. Therefore a linear model with multiple predictors provides a plausible explanation of how rank would change with respect to the model. Originally a full model was fit in order to perform backward stepwise regression and get a final model with only relevant predictor variables. A backward AIC regression was used which is a likelihood-based criterion for assessing models and balances goodness of fit and a penalty for model complexity. It is defined as:

$$AIC = 2[-logL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p, \hat{\sigma^2}|Y) + K]$$

Where $K = p + 2$ is the number of parameters in the fitted model. In addition, a backward BIC regression was also used (Bayesion Information Criterion) in which:

$$BIC = -2logL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p, \hat{\sigma^2}|Y) + Klogn$$

BIC is useful since it penalizes complex models more heavily than AIC, meaning it favours simpler models.

Results

Firstly, a full model was created with all predictors.

A summary of both models chosen by the AIC backward regression and BIC backward regression models is shown in Tables 2 and 3 respectively.

Table 2: AIC Regression Model

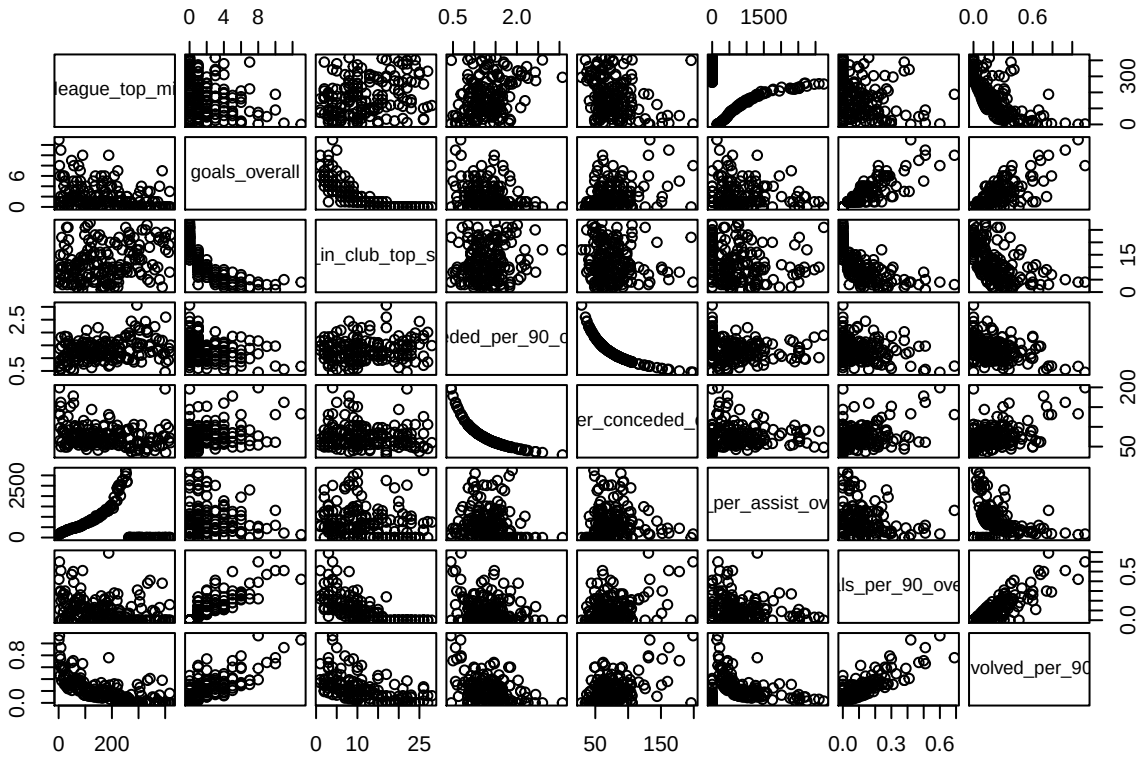
term	estimate	std.error	statistic	p.value
(Intercept)	102.2075638	54.4855615	1.875865	0.0624085
goals_overall	5.5708924	3.8969865	1.429539	0.1547064
rank_in_club_top_scorer	1.9546311	0.9648868	2.025762	0.0443715
conceded_per_90_overall	58.5781577	22.1875214	2.640140	0.0090680
min_per_conceded_overall	1.0305694	0.3439254	2.996491	0.0031450
min_per_assist_overall	-0.0264143	0.0061718	-4.279857	0.0000313
goals_per_90_overall	837.2216778	83.1670933	10.066742	0.0000000
goals_involved_per_90_overall	-857.3472041	41.1245020	-20.847601	0.0000000

According to the model created according to backwards AIC, and a significance level of 0.05, there are six significant variables in predicting the rank of a midfielder. These are goals overall, rank in the club (top scorer), conceded per 90 overall, minutes per assist overall, goals per 90 overall, and goals involved per 90 overall. According to this model, the only statistically insignificant predictors are goals overall and rank in club (top scorer). The proportion of variation in rank explained by the model is 0.7599. The AIC, $correctedAIC$, and BIC values are 1931.9438, 8, 1430.4775, and 1960.4782.

Table 3: BIC Regression Model

term	estimate	std.error	statistic	p.value
(Intercept)	123.4185247	53.7958404	2.294202	0.0230022
conceded_per_90_overall	62.0712741	21.9503007	2.827810	0.0052499
min_per_conceded_overall	1.1407951	0.3355797	3.399475	0.0008413
min_per_assist_overall	-0.0282314	0.0059747	-4.725130	0.0000048
goals_per_90_overall	848.9420241	62.9207770	13.492237	0.0000000
goals_involved_per_90_overall	-855.4543405	41.1078390	-20.810005	0.0000000

According to the BIC model, the model consists of the predictor variables conceded per 90 overall, minutes per conceded overall, minutes per assist overall, goals per 90 overall, and goals involved per 90 overall. In this model, all of the predictor variables are statistically significant in predicting the rank of a given midfielder. 0.7563 of the variation in rank is explained by the model. The AIC, $correctedAIC$, and BIC values are 1932.6757, 6, 1431.2093, and 1954.869.



According to a plot of the correlations, there is indication of high correlation between goals per 90 overall to goals overall (0.8648), goals involved per 90 to goals overall (0.7554), and goals involved per 90 to goals per 90 overall (0.8448). There is also a noteworthy strong negative correlation between rank in club and goals overall (-0.7456).

Discussion

Summary:

To find an ideal multiple linear regression model in estimating a prediction for a midfielder's rank in the league, backwards stepwise regression was performed on the data, and narrowed down possibilities to two potential models (one done by Aikaike's Information Criterion "AIC", and the other by Bayesian Information Criterion "BIC"). Observations were noted of various criteria from the models, and potential covariate predictors were observed. The figure with the correlations of variables was modeled after the AIC chosen model, regardless of whichever model was chosen. This is due to the fact that the AIC model had more potential predictor variables. Having more predictors in a valid model implies that bias is reduced, however variance of the estimated coefficients increases.

Conclusions:

After performing both AIC backward regression and BIC backward regression on a full model with all potential predictors, and comparing correlations in predictors, there can be several conclusions to be made from the analysis

By observing the values of R^2_{adj}, AIC, $correctedAIC$, and BIC a it can be said that the more accurate model is the one done by BIC. As could have been guessed, assists and goals are among the statistically

relevant predictors for rank at the end of the season. Interestingly, even defense of a midfielder is considered (according to the model), where goals conceded per game is a statistically significant predictor. When referring to the figure of correlations, it is worth looking into a strong negative correlation between rank in club and goals overall. The predictors chosen through the BIC method are conceded goals per game, minutes per conceded overall, minutes per assist overall, goals per game overall, and goals involved in per game overall. All of the predictors are statistically significant at a level of 0.05.

Weaknesses & Next Steps:

A potential weakness in the model is high correlation between predictor variables may result in an inaccurate model. There were in total three variables in the model that showed relatively high correlation. There are other potential models that could perhaps provide a better fit. There are also various types of variable selecting methods that could influence the outcome of the fit which were not looked at throughout this report.

Appendix

Removed columns from original dataset:
- minutes played away
- minutes played home
- appearances away
- appearances home
- goals away
- goals home
- assists away
- assists home
- penalty misses
- penalty goals
- clean sheets away
- clean sheets home
- conceded away
- conceded home
- goals per 90 away
- goals per 90 home
- min per match
- min per card overall

Github Repo: https://github.com/Andreivel23/STA304-Final-

References

FootyStats. (2020, 12 8). Premier League. Retrieved from footystacks.org: https://footystats.org/england/premier-league