Sir Alex Ferguson's Continued Tenure Could Have Sustained Manchester United's EPL Dominance*

An Analysis of an Increase in League Points

Aryaman Sharma

April 13, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/Ary4m3n/manchester-united.git

1 Introduction

Manchester United, an English football (soccer) team based in Manchester, England, has been one of the richest, most renowned and most supported clubs in the whole world. Founded in 1878, Manchester United is know for its note-worthy history in football. Manchester United's distinct history has been dominated by two long-serving managers, Sir Matthew Busby and Sir Alex Ferguson (*Manchester United* 2024). Sir Matthew Busby was the manager of Manchester United between 1945 and 1969, where he was most known for rebuilding the team after 23 of 44 players died after a plane crash in Munich in 1958.

Sir Alex Ferguson managed the club between 1986 and 2013 where he led the team to an unparalleled spell of dominance in the English Premier League (*Manchester United* 2024). The English Premier League (EPL) since the 1992-1993 season has been the top-tier league in the English football league system (*Premier League* 2024). Under Sir Alex Ferguson, Manchester United won 12 Premier League titles between 1992 and 2013, and dominated the rest of the 19 teams in the league. Sir Alex Ferguson is also well renowned to have nurtured the young talent in Cristiano Ronaldo, arguably now deemed to be one of the greatest players of all time.

Sir Alex Ferguson retired after the 2012-2013 season and that marked the now seen downfall of the club in recent times. Since the 2012-2013 season, Manchester United has seen 8 different managers and a significant drop in the stature of the club around trophies and wins (*Manchester United - Manager History* 2024). For context, the club won 38 trophies under Sir Alex and merely 5 trophies in total since his retirement under 8 different managers (*How Many Trophies Have Manchester United Won?* 2023). Additionally, the club has not won the English Premier League since Sir Alex's last season, 2012-2013, whereas they won 13 titles when under Sir Alex.

The aim of this paper is two-folded, where in the paper will first analyze the differences in performance between 2004-2013 (Sir Alex's era) and 2014-2023 (Post Sir Alex's era). Then, we will use this knowledge to build two Bayesian Models to predict and further analyze the estimand: Average League Points in each season for 2004-2013 and 2014-2023. The paper will also illustrate *how much* better Manchester United's position would have been in 2014-2023 if Sir Alex Ferguson would not have retired, helping us contribute positively to the debate about him being the greatest manager of all time.

The structure of this paper comprises four sections: Data, Model, Results and Discussion. In the Data section (Section 2), we discuss the data source and the process of measuring and cleaning the datasets. In the Model section (Section 3), we discuss the two Bayesian Models used in the paper, their justifications and how it was constructed. This section will also touch on the process of predicting the league points for 2014-2023 if Sir Alex were managing the club then. In the Results section (Section 4), we delve deeper into the trends observed. Finally, in the Discussion section (Section 5), we discuss the possible factors around contributing to Sir Alex's success, along with limitations and further research.

2 Data

The datasets under examination in this paper were obtained from the R package and API `worldfootballR` (Zivkovic 2022). Employing the open-source `R` programming language (R Core Team 2023), we conducted the cleaning and analysis procedures, leveraging several R libraries and packages such as `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `ggplot2` (Wickham 2016), `knitr` (Xie 2023), `readr` (Wickham, Hester, and Bryan 2023), `dplyr` (Wickham et al. 2023) and `arrow` (Richardson et al. 2024). We used `rstanarm` (Goodrich et al. 2022) for the models and `modelsummary` (Arel-Bundock 2022) for the model summary.

As mentioned above, in terms of estimand, the primary focus of this analysis is to assess the impact on league points Sir Alex would have had if he had been the manager between 2014 and 2023 as well. The three key predictors include the Success Rate per season (also interpreted as the Win Ratio or Win Percentage), the total goals scored and the goal difference (i.e. goals conceded deducted from goals scored). These will be talked more in detail about further into the paper.

In (Section 2.1) we provide an overview of the raw data obtained, (Section 2.2) delves deeper into the intricacies of the data cleaning process and (Section 2.3) outlines the measurement aspects for the paper.

2.1 Raw Data

As mentioned above, the raw data for this paper was collected from the R package and API `worldfootballR` (Zivkovic 2022). Team statistics per season for the English Premier League were obtained from the package. We obtained data per season, that implies that there were 20 files of raw data for 20 seasons we wish to analyze for 20 different teams in league each season. The dataset contained a lot of statistics that were not all relevant to us. Relevant statistics included the Squad Name, the finishing Rank, the number of Wins, Draws, Losses, Goals For, Goals Against, Goal Difference and the League Points. Other statistics included the Average Attendance per season, Expected Goals etc.

The next section (Section 2.2) will show the data cleaning process, further outlining the structure of the refined dataset employed in our analysis.

2.2 Cleaned Data

The indicators or statistics mentioned in (Section 2.1) were extracted from each raw data file and combined together to form two data tables for seasons between 2004-2013 and 2014-2023. Table 1 shows the cleaned data table for the seasons between 2004 and 2013 under Sir Alex. Table 2 in the appendix shows the cleaned data table for the seasons between 2014 and 2023 for reference.

Table 1: Cleaned Data showing Premier League statistics for Manchester United (2004-2013)

Year	Rank	Wins	Draws	Losses	Win Percentage (%)	Goals Scored	Goals Conceded	Goal Difference	League Points
2004	3	23	6	9	60.53	64	35	29	75
2005	3	22	11	5	57.89	58	26	32	77
2006	2	25	8	5	65.79	72	34	38	83
2007	1	28	5	5	73.68	83	27	56	89
2008	1	27	6	5	71.05	80	22	58	87
2009	1	28	6	4	73.68	68	24	44	90
2010	2	27	4	7	71.05	86	28	58	85
2011	1	23	11	4	60.53	78	37	41	80
2012	2	28	5	5	73.68	89	33	56	89
2013	1	28	5	5	73.68	86	43	43	89

There are 10 columns in total in the League Statistics data table:

1. The `Year` column refers to the Premier League season. Generally, a season starts in August and ends in May of the next year. Hence, seasons are generally represented as, for instance, 2003-2004. However, here for simplicity, the `Year` column refers to the year the season ends. So, the season 2003-2004 will be represented as 2004.
2. The `Rank` column points to the position Manchester United finished in, in the particular season. The Rank is going to between 1 and 20 because there are 20 teams in the league each year.
3. The `Wins` column contains the total number of wins by Manchester United in the particular season. There can be at maximum 38 wins because there are 38 games in total.
4. The `Draws` column contains the total number of draws by Manchester United in the particular season. There can be at maximum 38 draws because there are 38 games in total.

5. The `Losses` column contains the total number of losses by Manchester United in the particular season. There can be at maximum 38 losses because there are 38 games in total. Additionally, in total there are going to be 38 wins, draws and losses.
6. The `Win Percentage (%)` column, also referred to as the *Success Percentage* is calculated as the total number of wins divided by the total number of games, i.e. 38. It is shown as a percentage for simplicity of understanding. It is understood that higher the Win Percentage, the better the season was, and the better the league points will be.
7. The `Goals Scored` column outlines the number of goals scored by the team in the particular season. It is understood that this measure is a good indicator of how well the offense played in a particular season. The higher the goals scored, the better the offense played.
8. The `Goals Conceded` column outlines the number of goals conceded by the team in the particular season. It is understood that this measure is a good indicator of how well the defense played in a particular season. The lower the goals conceded, the better the defense played.
9. The `Goal Difference` column outlines the difference between the number of goals scored and goals conceded. It is understood that this measure is a good indicator of how well the team played as a whole. This value can be negative as well, implying that the team did bad in the season.
10. The `League Points` column indicates to the total number of points scored that season. This indicator is essential in determining where the teams end up in the league. A team receives 3 points for a win, 1 point for a draw and 0 points for a loss.

We will be using the *Win* or *Success Percentage*, *Goals Scored* and *Goal Difference* as our predictors to model for the *League Points* ahead, which will be explained in detail in (Section 3). Now, in (Section 2.3), we will delve into the measurement aspects of the data presented in the tables shown. Understanding this process is crucial for drawing meaningful insights from the data analysis results.

2.3 Measurement

According to the Premier League (*Statistics Explained* 2024), all the official performance data is collected and reported by Opta, a part of Stats Perform (*Stats Perform* 2024). All the data is collected by a team of three people which cover each match. These three people include two highly trained analysts who go through video-based collection system to gather data, and a quality control analyst who can rewind the video feed frame-by-frame to make sure the data collected is correct (*Statistics Explained* 2024). Additionally, the data collected is then subject to an exhaustive post-match check to ensure accuracy. This comprehensive process ensures that the data collected is complete and highly accurate. As the data analyzed in this paper is *not* collected and reported by the teams themselves, it is highly unlikely that a collection bias would exist. However, some might argue that some teams might be able to pressure the officials into tipping the statistics over in their favor, but this might be highly unlikely as well.

The `Win Percentage` or also referred to as the `Success Percentage` was calculated manually by taking taking a ratio of the wins in a season and the total number of games in a season (i.e. 38). As mentioned above, as the wins were reported by Opta (*Statistics Explained* 2024) and not by individual teams, hence once again the issues about any biases caused can be avoided.

There might be a concern when considering the fact that this paper analyses data for the past 20 years. There might have been a difference in data reporting back in 2004 when compared to right now. It is essential to note that, this difference would have a bigger impact on data for a minute to minute match data analysis. However, in this report we use data about each season. As this data includes most the number of points scored, wins, draws, losses and goals, we work under the assumption that this data should not be affected to a large extent. However, at the same time it is important to still acknowledge that data collection in the past could have impacted the data being analyzed and modeled with.

In (Section 3) we will describe the models used in this paper and provide with justification and hypothesis of what we expect from the models.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$
\begin{align}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \tag{1}\\
\mu_i &= \alpha + \beta_i + \gamma_i \tag{2}\\
\alpha &\sim \text{Normal}(0, 2.5) \tag{3}\\
\beta &\sim \text{Normal}(0, 2.5) \tag{4}\\
\gamma &\sim \text{Normal}(0, 2.5) \tag{5}\\
\sigma &\sim \text{Exponential}(1) \tag{6}
\end{align}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

Table 2: Cleaned Data showing Premier League statistics for Manchester United (2014-2023)

Year	Rank	Wins	Draws	Losses	Win Percentage (%)	Goals Scored	Goals Conceded	Goal Difference	League Points
2014	7	19	7	12	50.00	64	43	21	64
2015	4	20	10	8	52.63	62	37	25	70
2016	5	19	9	10	50.00	49	35	14	66
2017	6	18	15	5	47.37	54	29	25	69
2018	2	25	6	7	65.79	68	28	40	81
2019	6	19	9	10	50.00	65	54	11	66
2020	3	18	12	8	47.37	66	36	30	66
2021	2	21	11	6	55.26	73	44	29	74
2022	6	16	10	12	42.11	57	57	0	58
2023	3	23	6	9	60.53	58	43	15	75

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

How Many Trophies Have Manchester United Won? 2023. Sporting News Canada. https://www.sportingnews.com/ca/soccer/news/manchester-united-trophies-won-complete-list-silverware/uuyb6wc5o7isrfwa1pjssb07.

Manchester United. 2024. Encyclopædia Britannica, inc. https://www.britannica.com/topic/Manchester-United.

Manchester United - Manager History. 2024. WorldFootball. https://www.worldfootball.net/teams/manchester-united/9/.

Premier League. 2024. Premier League - facts; history. https://www.footballhistory.org/league/premier-league.html.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Statistics Explained. 2024. Premier League. https://www.premierleague.com/stats/clarification.

Stats Perform. 2024. Stats Perform. www.statsperform.com/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zivkovic, Jason. 2022. *worldfootballR: Extract and Clean World Football (Soccer) Data*. https://CRAN.R-project.org/package=worldfootballR.