Final Report

Dylan Li, Liam Quach, Brendan Callender

I. Introduction

The English Premier League (EPL) is the top tier of professional football (soccer) in England and is considered one of the most popular and competitive leagues in the world. The league is made up of twenty clubs (teams) that compete over a season for the Premier League title with new clubs added each year via a system of promotion and relegation. Each year, three new clubs are promoted from the second division based on the previous year's results with these promoted teams replacing the bottom three teams from the previous year's Premier League season.

Over the course of a season, each team plays a total of 38 matches, facing every other team twice—once at home and once away. Teams are rewarded points from each game as follows: 3 points for a win, 1 points for a draw, and 0 points for a loss. The team with the most points at the end of the 38-game season is crowned as the Premier League Champions.

add a little more maybe?

For our project, we are interested in exploring the following research questions:

1. What factors are associated with higher or lower point totals in the English Premier league?
2. Is spending more money in the off-season associated with earning more points the following season?
3. How do differences in expected metrics to actual metrics impact a clubs point totals?

II. Data Source & Methods

To answer our research questions, we collected English Premier League season-level data spanning from the 2017-2018 season up to the most recently completed 2023-2024 season. Data was collected from two sites: rbref.com and transfermarkt.com. The data collected from fbref includes performance related metrics for each team over the season as predictors as well as point totals for each team at the end of the season for our response variable. The performance metrics include total goals scored, total goals conceded, expected goals scored, expected goals

conceded, average % possession, shooting metrics and more. The data collected from transfermarkt includes data relating to each teams expenditure and sales with respect to buying and selling players in the trasnfermarkt. This data includes money spent, money gained from sales, net spend, number of players bought, number of players sold and more. Money related variables are measured in thousands of euros.

Predictors relating to season totals such as goals scored and goals conceded were scaled down to per 90/ per game values for better interpretability. This was achieved by dividing these metrics by the total games played which is 38.

Variable | **Role** | **Range of Values** |

«««< HEAD |————————————————|————————|————————| ======= |—-|—-|—-| »»»> e47b9c6bb14c4a849a391712757caaa4de6a8558 | Points | Response | (16, 100) | | Goals/90 | L1 | (0.52, 2.78) | | Goals Against/90 | L1 | (0.58, 2.74) | | Average Posession of Ball (%) | L1 | (35.4, 71.0) | | … | … | … | | Net Spend (in €1,000,000) | L1 | (-118.07, 562.39) | | Club Average Net Spend (in €1,000,000) | L2 | (-7.72, 139.39) | | Actual vs Expected Goals/90 Difference | L1 | (-0.37, 0.72) | | Actual vs Expected Goals/90 Against Difference | L1 | (-0.37, 0.72) |

: Description of Dataset Variables

See example rows of data below:

Table 1: Example Rows from Dataset

Club	Season	Pts	GF	GA	Poss	…	NetSpend	Mean_NetSpend
Chelsea	2017	70	1.63	1.0	55.6	…	65.9	139.0
Arsenal	2017	63	1.95	1.34	61.4	…	-9.55	100.0
Everton	2017	49	1.16	1.53	45.5	…	76.8	26.0

To analyze the data, we will employ multi-level regression models, also known as hierarchical linear models. This approach is well-suited for the structure of the dataset, in which we have repeat observations for different clubs over several seasons. (See figure below)

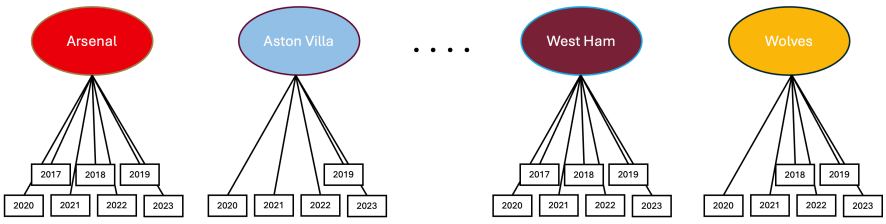


Figure 1: Multi-level Structure of Data

III. Results

Exploratory Data Analysis

This section presents the exploratory data analysis conducted to understand the key relationships between variables in the dataset. This exploratory data analysis was conducted before the model fitting process to gain an initial understanding of our research questions.

Figure 2 below shows the join distribution of goals scored per game and goals conceded per game, colored by season point totals. We there is a strong, negative correlation between goals scored per game and goals conceded per game. This means that teams who tend to score more, also tend to conceded less as well. When considering the season point totals, we see that decreasing the number of goals conceded per game is associated with higher point totals holding goals scored constant. Additionally, increasing the number of goals scored per game is assoicated with higher point totals holding goals conceded constant. Lastly, we see that jointly decreasing the number of goals conceded per game and increasing the number of goals scored per game is associated with the largest increase in season point totals.

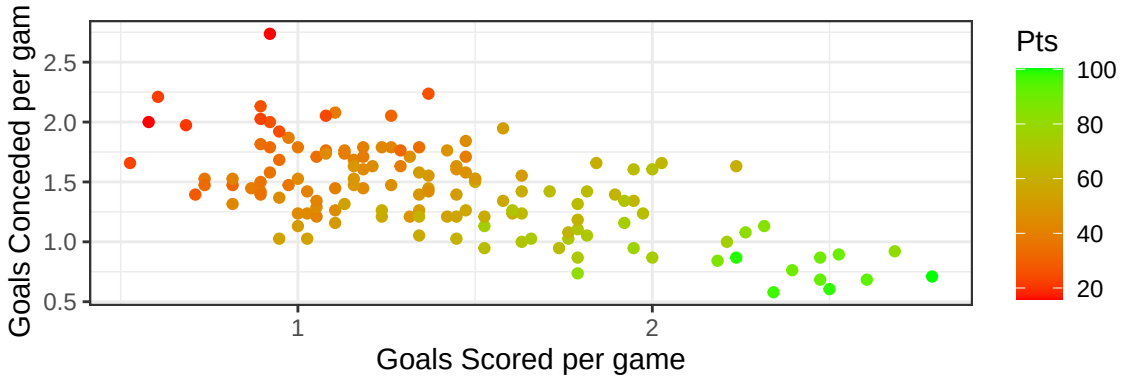


Figure 2: Impact of Goals Scored/90 and Goals Conceded/90 on Season Point Totals

Figure 3 below shows the relationship between average % possession of the ball and season point totals. The plot shows a strong, positive associated between % possession and season points with higher values for % possession associated with higher point totals. This makes sense intuitively because teams with more possession tend to have the ball more which reduces the chances of the opposing team scoring and gives your team more chances to score goals.

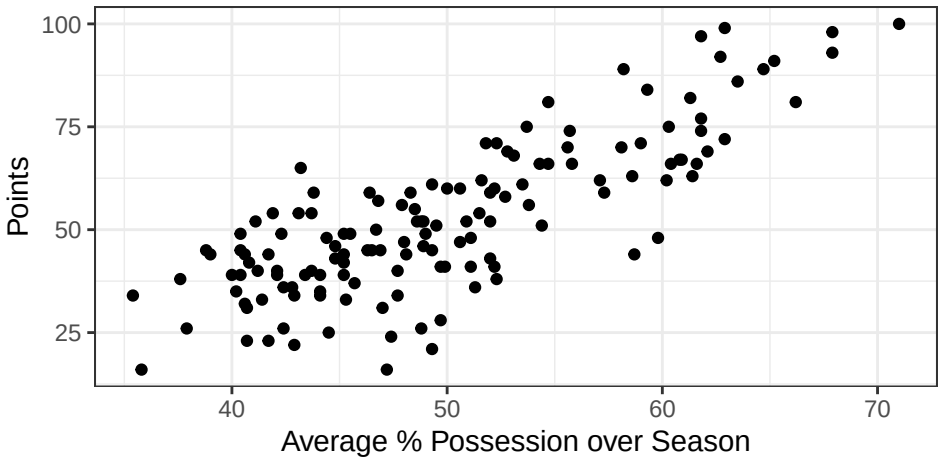


Figure 3: Season Point Totals by Season Average % Possession

Lastly, the plots below in *Figure 4* and *Figure 5* show the impact of spending on season point totals. The plot on the left shows the relationship between season point totals and the net spend of the club for individual seasons. Larger values for net spend represent a club spending more money on new players while smaller values indicate a club spending less money spent with negative values indicating a team made profit selling players in the market. From the plot, we see a weak positive association with teams who spend more money being associated with higher point totals. We also notice a major outlier in the plot with Chelsea in the 2023-2024 season. This is a valid data point and represents the season in which Chelsea had new owners invest large amounts of money into the team. This is not normal behavior for when teams get new owners and can serve as an example of how making too many changes to a team has a negative impact on performance.

In *Figure 5*, we see a much stronger positive associated between net spend and points when aggregated for each club. This demonstrates that consistent investment into a team over many seasons is more strongly associated with higher point totals than just a single season of large investment. (As they say… Rome wasn't built in a day)

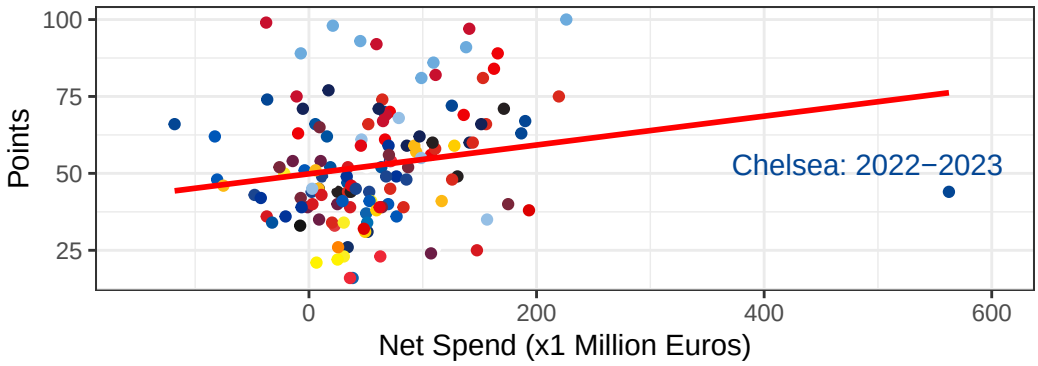


Figure 4: Season Point Totals by Single Season Net Spend

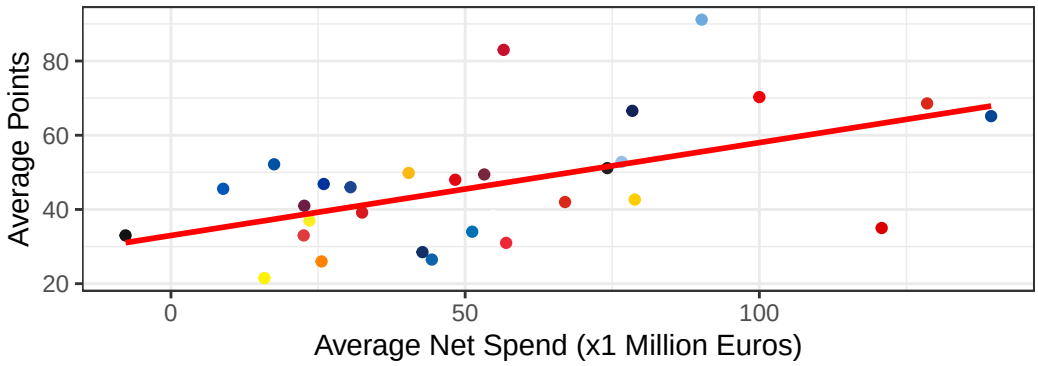


Figure 5: Club Average Points by Club Average Net Spend

ANOVA

After performing exploratory data analysis, an initial Analysis of Variance (ANOVA) test was performed to explore whether there was significant club-to-club variability in season point totals (See results in *Table 3* below). Looking at the p-value resulting from the ANOVA, we have significant evidence that at least 2 clubs have different mean point totals. This is supported by *Figure 6* below which shows the distribution of point totals by each club. We see clubs like Manchester City have very high point totals while clubs like West Brom and Norwich City have very low point totals.

Table 2: *ANOVA for Significance of Club-to-Club Variability*

	df	SSE	MSE	F-Statistic	P-Value
Club	29	37233	1283.89	12.848	< 0.0001

	df	SSE	MSE	F-Statistic	P-Value
Residuals	110	10992	99.93		

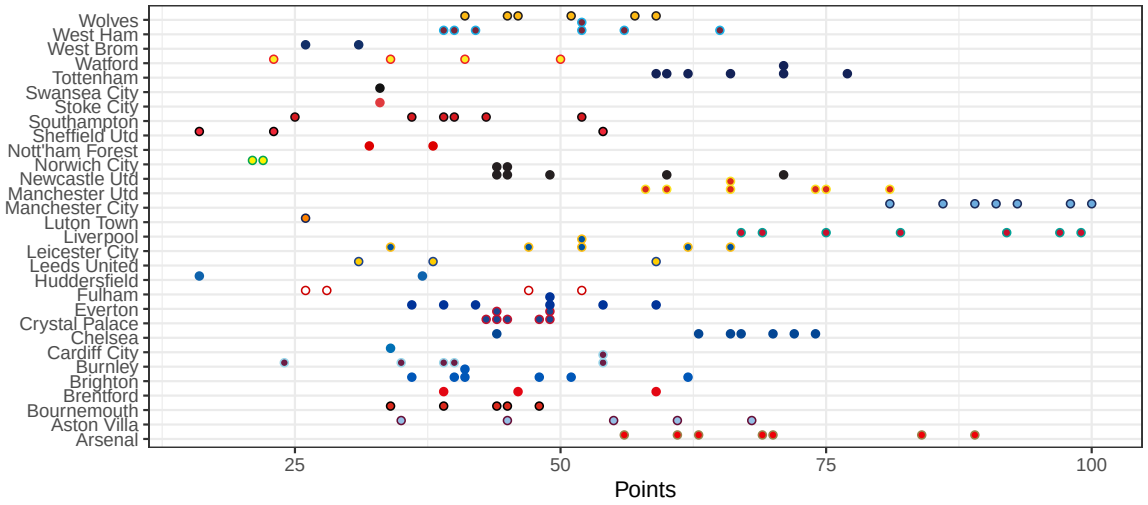


Figure 6: Distribution of Point Totals by Team

Null Model

After finding signficant club-to-club variability in the season point totals, we fit an initial null model which only includes the club as a random effect. The model can be written out as seen below:

$$Points_{ij} = \beta_{00} + u_j + \epsilon_{ij}$$

where $u_j \sim N(0, \tau_0^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$

The summary of the null model output can be found below in *Table 4*. Looking at the ICC of the model, we see that approximately 72% of the variation is at the club level while 28% of the variation is within each club. This matches what we saw in *Figure 6* above with clubs point totals being similar across the seasons. Additionally, we see the model seconds the fact that there is significant club-to-club variation with the confidence interval for τ_0 not containing 0. Lastly, we see the resulting predictions from the null model in *Figure 7 below*. The black points represent the mean points for each club while the red points represent the predicted points for each club. The distance from the red and black points represent the shrinkage that occurs when using a multi-level model. We can see a club like Luton Town has large shrinkage towards the mean because we only have one season of data for Luton.

Table 3: Null Model Summary

Parameter/Statistic	Estimate
σ^2	255.3
τ_0^2	99.6
95% CI for τ_0	(11.99, 21.18)
ICC	0.72

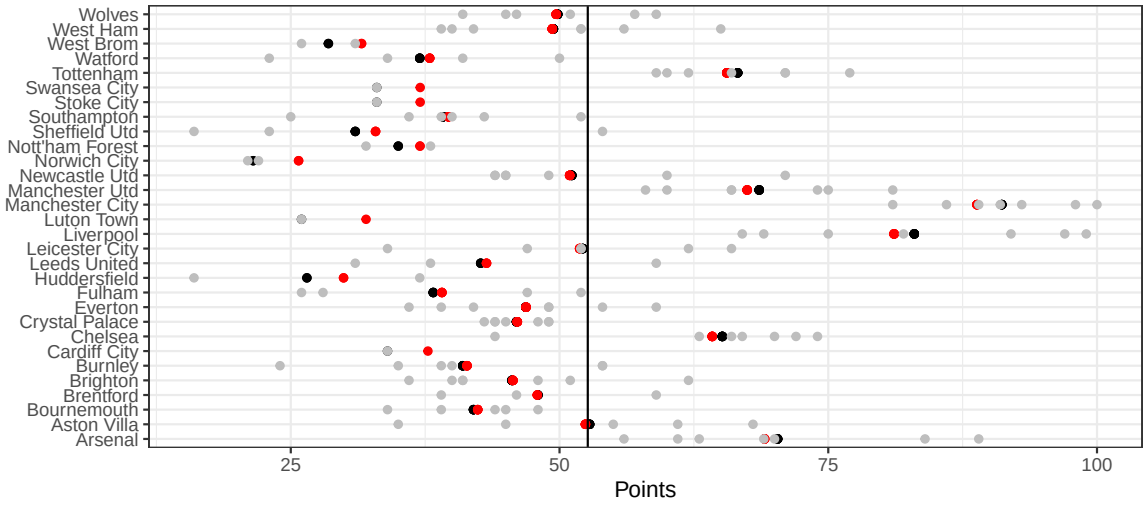


Figure 7: Null Model Predictions

Model Fitting Process

Below shows the process of how we arrived at our final model. We first began by including goals for and goals against in our first model due to our EDA showing a strong joint association with points. We found these variables to be significant so we included them in each subsequent model. From there we continued to explore new models by adding different level 1 predictors. If a predictor was significant, it was left in the model. After exploring level 1 predictors, we added our only level 2 predictor to the model which is average net spend for each club. We found this to be significant and moved on to adding random slopes to the model.

```
model1 <- lmer(Pts ~ GF + GA + (1 | Club), data = prem)
summary(model1)
```

Add Level 1 Predictors

```
model2_1 <- lmer(Pts ~ GF + GA + xG_diff + xGA_diff + (1 | Club), data = prem)
summary(model2_1)
```

```
model2_2 <- lmer(Pts ~ GF + GA + NetSpend + (1 | Club), data = prem)
summary(model2_2)
```

```
model2_3 <- lmer(Pts ~ GF + GA + Poss + (1 | Club), data = prem)
summary(model2_3)
```

```
model2_4 <- lmer(Pts ~ GF + GA + Age + (1 | Club), data = prem)
summary(model2_4)
```

Add Level 2 Predictors

```
model2_5 <- lmer(Pts ~ GF + GA + Mean_NetSpend + (1 | Club), data = prem)
summary(model2_5)
```

```
model2_6 <- lmer(Pts ~ GF + GA + NetSpend + Mean_NetSpend + (1 | Club), data = prem)
anova(model2_5, model2_6)
```

Random Slopes

```
model3_1 <- lmer(Pts ~ GF + GA + Mean_NetSpend + (1 + GF | Club), data = prem)
anova(model2_5, model3_1)

model3_2 <- lmer(Pts ~ GF + GA + Mean_NetSpend + (1 + GA | Club), data = prem)
anova(model2_5, model3_2)

model3_3 <- lmer(Pts ~ GF + GA + Mean_NetSpend + (1 + GA + GF | Club), data = prem)
anova(model2_5, model3_2)
```

Final Model

After the model fitting process we finished with a final model that can be written as follows:

$$Points_{ij} = \beta_{00} + u_j + \beta_1(G/90)_{ij} + \beta_2(GA/90)_{ij} + \beta_3(\overline{NetSpend})_j + \epsilon_{ij}$$

where $u_j \sim N(0, \tau_0^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. Additionally, all predictors have been grand-mean-centered (See interpretations below).

A summary table of the output from the final model can be found below in *Table 5* below. We see that the coefficients for the predictors included in the model match what we would expect after conducting our EDA. However, we do notice that there is only a small effect associated with clubs spending more money over many seasons after accounting for goals scored and goals conceded per game. Lastly, we notice that all the club-to-club variability in point totals can be explained by our model including goals scored, goals conceded, and average net spend. This matches our intuition due to how teams are awarded points: 3 for a win, 1 for a draw, and 0 for a loss. Simply put, more goals scored and less goals conceded means more wins which means more points.

Parameter | Estimate | Interpretation |

|————-|————-|————————————————|

σ^2 | 19.96 | 92% of Level 1 variability explained when compared to null model |
| | |
τ_0^2 | 0 | 100% of club-to-club variability explained by predictors |
| | |
β_{00} | 52.63 | Predicted point total for club with average goals scored, goals conceded, and net spend |
| | |
β_1 | 23.04 | Associated increase in predicted points with each 1 increase in goals scored per game after adjusting for goals conceded, net spend, and club. |
| | |
β_2 | -21.87 | Associated increase in predicted points with each 1 increase in goals conceded per game after adjusting for goals scored, net spend, and club. |
| | |
$50\beta_3$ | 1.55 | Associated increase in predicted points with each €50,000,000 increase in club average net spend after adjusting for goals scored, goals conceded, and club. |

: Final Model Summary

Model Diagnostics

Looking at the diagnostic plots for our final model, we see that the Linearity, Normality, and Equal Variance assumptions are not violated. Firstly, in *Figure 8* below, we can see a random scatter both above and below the horizontal line which indicates linearity is not violated. In the same plot, we also see that there is no obvious pattern of fanning in the data which means the equal variance assumption is not violated. Lastly, in *Figure 9* we see the points in the plot follow the diagonal line indicating the normality assumption is not violated. In these plots, we do notice potential outliers with both positive and negative residuals.

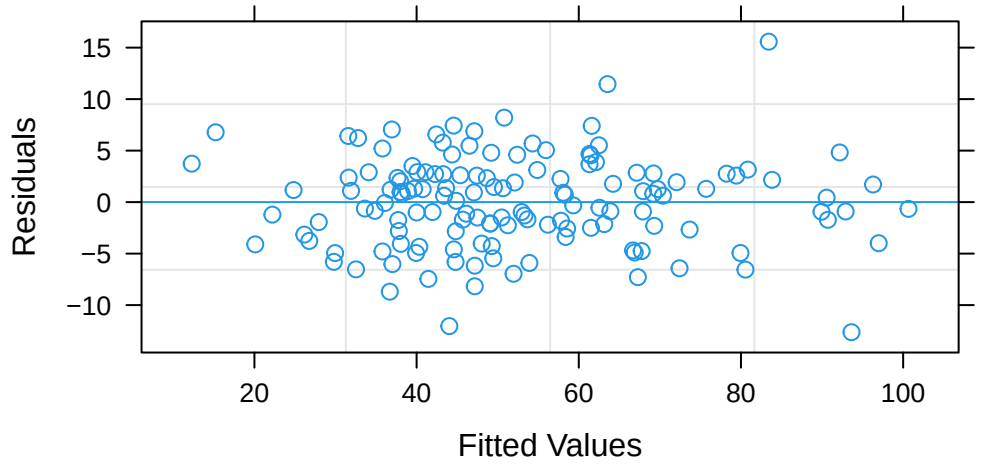


Figure 8: Residuals vs Fitted Plot

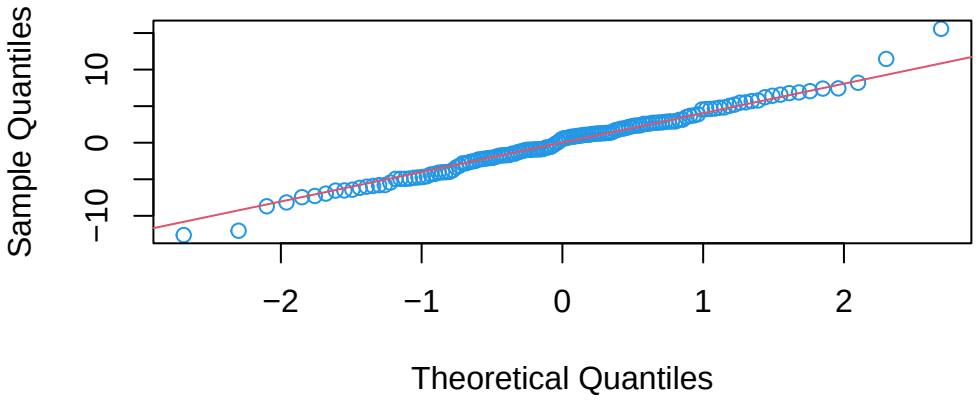
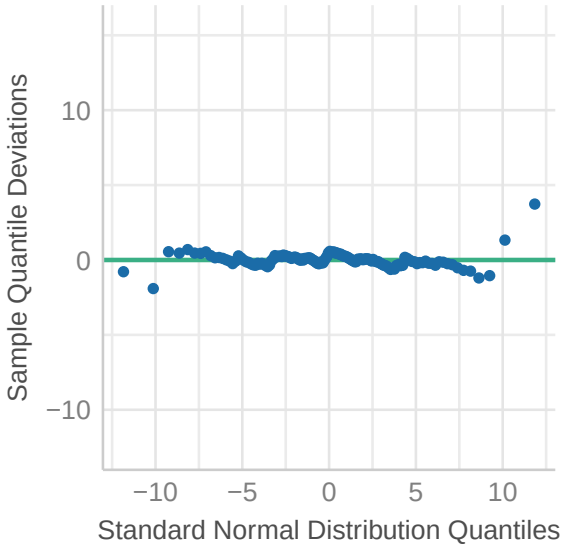


Figure 9: Normal-QQ Plot

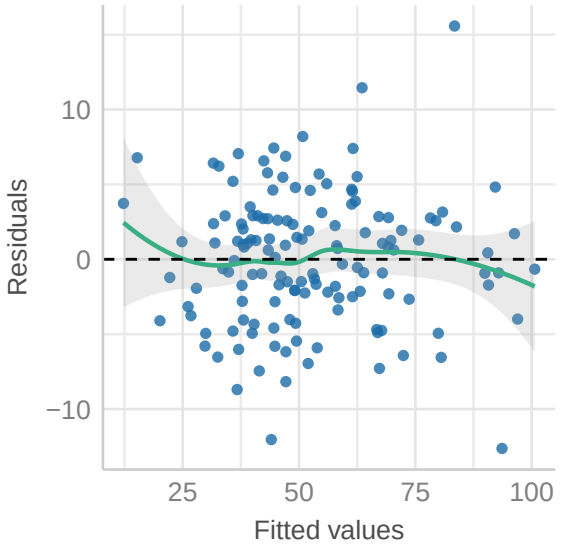
Normality of Residuals

Dots should fall along the line



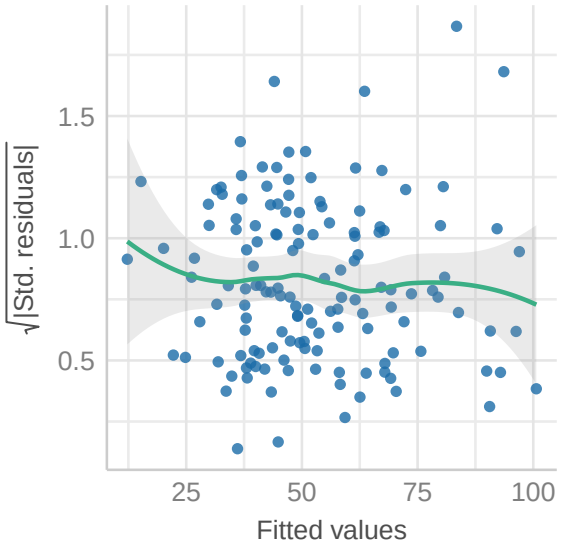
Linearity

Reference line should be flat and horizontal



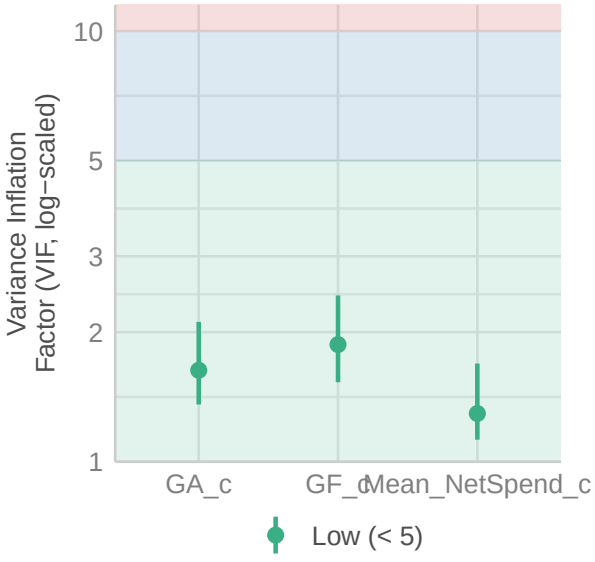
Homogeneity of Variance

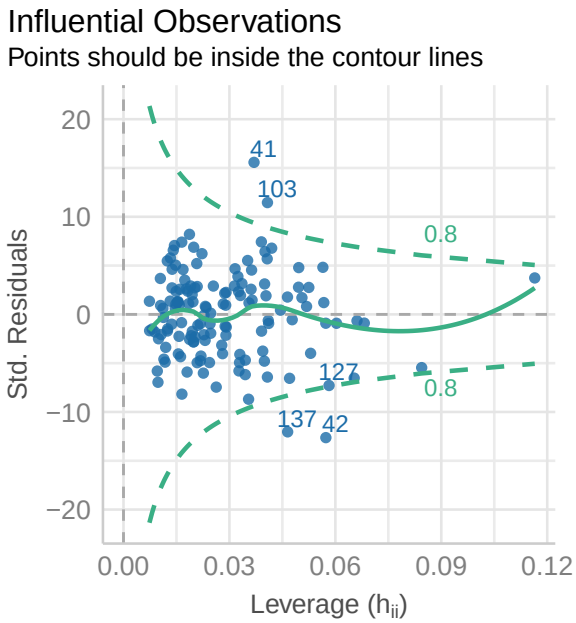
Reference line should be flat and horizontal



Collinearity

High collinearity (VIF) may inflate parameter uncertainty





Graph 1 shows the assumption for normality of residuals are met without problem. On Graph 2, we can see that the reference line deviates slightly at either end. However, due to the lack of distinct patterns and the deviation not being drastic, we conclude that the assumption of linearity is mostly met. Graph 3 shows that equal variance assumption is met as there is not any fanning of the residuals present. Graph 4 displays no signs of collinearity between our predictor variables. Graph 5 shows several influential outliers present. These outliers are within reason so we conclude that our models are not greatly affected by them.

V. Discussion

V. Discussion

Research Questions

Implications

Limitations and Next Steps

VI. Appendix

to do: add all model code stuff

add variable labels:

ANOVA

```
# A tibble: 2 x 6
  term         df  sumsq meansq statistic   p.value
  <chr>     <int>  <dbl>  <dbl>     <dbl>     <dbl>
1 Club         29 37233. 1284.       12.8  1.10e-23
2 Residuals   110 10992.   99.9      NA   NA
```