Premier League Players Analysis

PSTAT100: Data Science Concepts and Analysis

STUDENT NAME

- Agneya Poduval (aspoduval)
- Ibrahim Siri (isiri)
- Dongchi Xue (dongchi)
- Aarnav Gandhi (agandhi)
- Elliot Ly (elliotly)

1 Introduction

We chose a dataset from OpenML that collects observations from every player that has played in the English Premier League, including data on their offensive/defensive performances and their position. Since the data was collected from 1992-2023, we could use this dataset to extrapolate for the current season or future seasons. However, we don't think that this dataset could be used to generalize for other nations' leagues or even England's lower division leagues because of the high level of general play in the league. The hope would be that coaches could use some of this research to help their players know what to focus on in training.

We chose to research whether a player's big chances missed per game affect that player's losses per game and whether a player's interceptions per game and big chances per game created are related. Our first hypothesis is that if a player misses more big chances per game, then it will positively affect that player's losses per game. Our second hypothesis is that a player's number of interceptions per game is positively correlated with big chances per game.

We chose our first topic of analysis because we thought that a player who misses more chances per game than others would negatively affect their team's performance. However, we also wanted to see if the fact that the player had so many chances in the first place meant that the team they were playing for was naturally going to be better than other teams, meaning that their misses actually had little or no impact on their team's losses.

We chose our second topic of analysis because it seemed likely to us that players who intercept the ball more have more opportunities to create quick counterattacks and get the ball forward.

It is also important to split this research by position group because stats such as big chances and interceptions can vary wildly depending on where you play. Of course, a forward who is more often near the goal will have more big chances missed than a defender, who rarely ever gets the chance to shoot. Thus, our analysis is focused on finding any significant outliers or trends that could help players in each position group make fewer mistakes and play better.

2 Modeling Process

2.1 Hypothesis 1

For hypothesis 1, we first fit a simple linear regression model for forwards. After testing the initial model, we found that it did not meet most of the OLS model assumptions. Thus, we tried again after removing outliers through evaluation of Cook's Distance and the hat values. Still, the normality assumption was not met, so we additionally performed a Box-Cox transformation on the data. After checking the assumptions again, which will be explained later in this section, we found that this model was sufficient and ready to be used.

2.2 Hypothesis 2

For hypothesis 2, we decided to fit three separate linear regression models - one for each position on the field. This is because fitting simple linear models for each position would allow us to discover trends much easier than if we created a large interaction model that does not account for the categorical variable Position. This is also explored further in our report. Similar to our linear model for hypothesis 1, the initial model did not meet all of the assumptions, so we removed outliers by evaluating Cook's Distance and hat values and performed a Box-Cox transformation to help with the normality assumption. However, the models were still not performing as we hoped, so we decided to try many different transformations through guess-and-check. Finally, we found that the inverse of values for the forwards, the square root of the absolute value of the inverse for midfielders, and the log of the square root of the absolute value of the inverse for defenders produced the best results. Consider throughout the report that the raw values presented for these models have been transformed, and thus will be hard to interpret. Additionally, certain metrics will fail, such as the RMSE, which measures how much each model's predicted values deviate from actual values. However, after trying various different transformations, these were the best fitting models we could find.

2.3 Why Not Use Multiple Regression?

Table 1: Multiple Linear Model for Effect of INT/G and Position on Big Chances Created Per Game

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.007274	0.02302	-0.316	0.7521
Interceptions_pg	0.05863	0.01588	3.693	0.0002503
PositionMidfielder	0.121	0.02736	4.422	1.24e-05
PositionForward	0.09698	0.02767	3.505	0.0005041
Interceptions_pg:PositionMidfielder	-0.06966	0.02193	-3.177	0.001595

	Estimate	Std. Error	t value	Pr(>|t|)
Interceptions_pg:PositionForward	0.01555	0.03929	0.3957	0.6925

The model can easily be explained by presenting a regression equation:

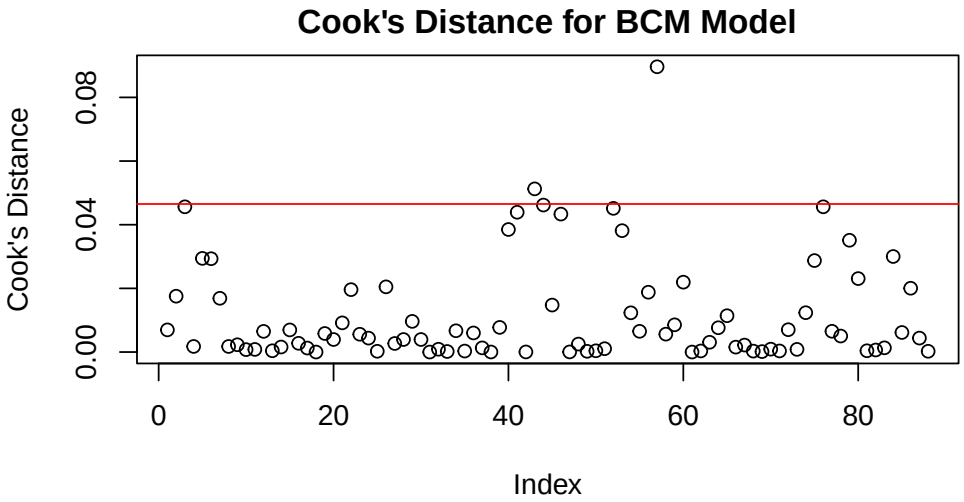
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$
$$\hat{y} = -0.0074 + 0.059 x_1 + 0.121 x_2 + 0.097 x_3 - 0.0697 x_1 x_2 + 0.0156 x_1 x_3$$

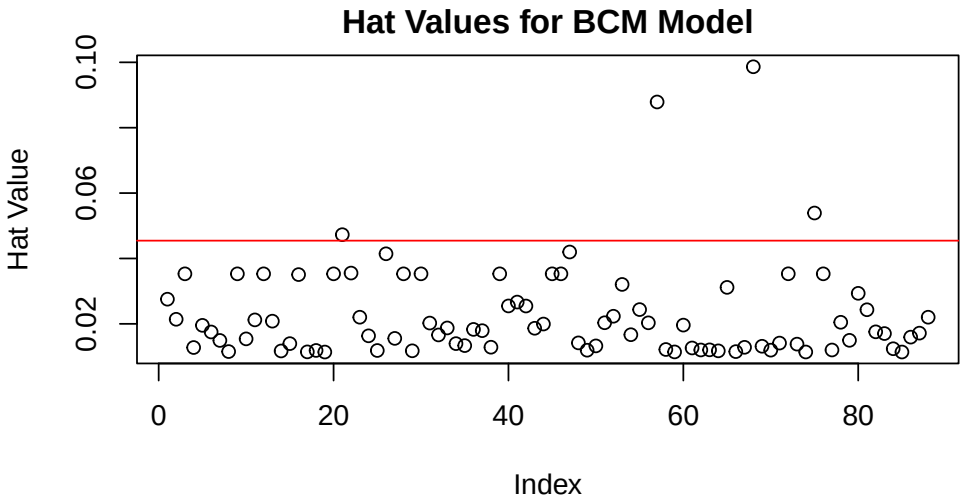
As you can see from this table and the amount of terms in the equation, there are many interaction terms that are hard to interpret. For example, what is the difference between the main effect of interceptions per game and the interaction term between interceptions per game and each position? None of these variables would have any real world significance because there are no levels to the factor that we are analyzing (position), meaning that creating numerical differences between each position would make no sense. In addition, the last interaction term was not significant. Thus, we would have had to perform additional transformations to an already cumbersome model. For these reasons, we decided not to explore further with this model.

3 Model Evaluation

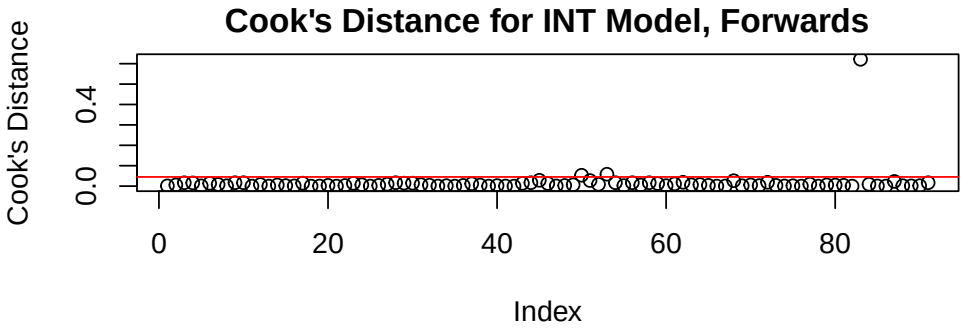
3.1 Outliers and Leverage Plots

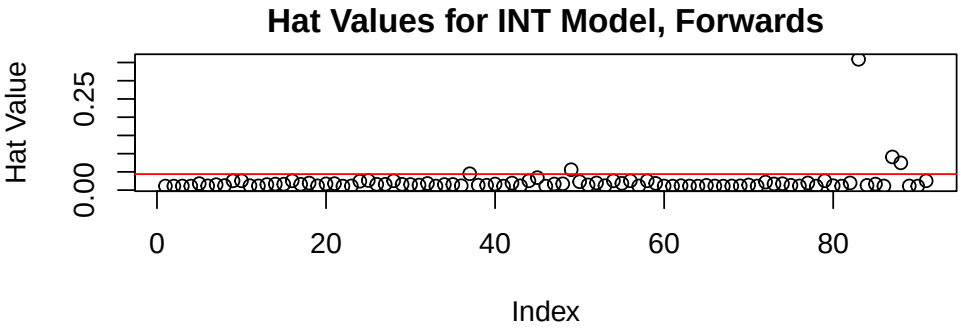
3.1.1 Hypothesis 1



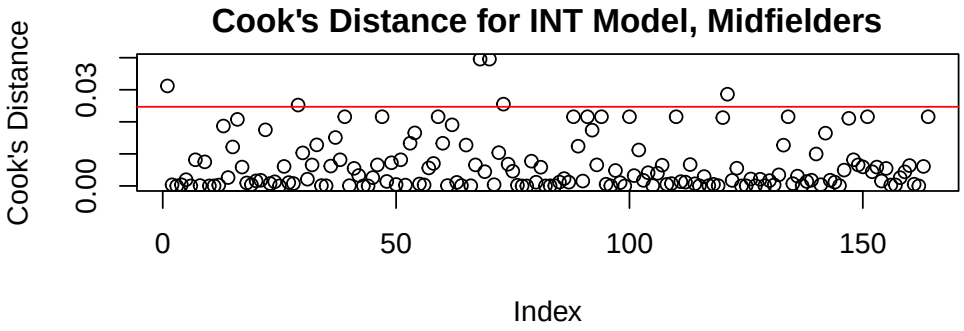


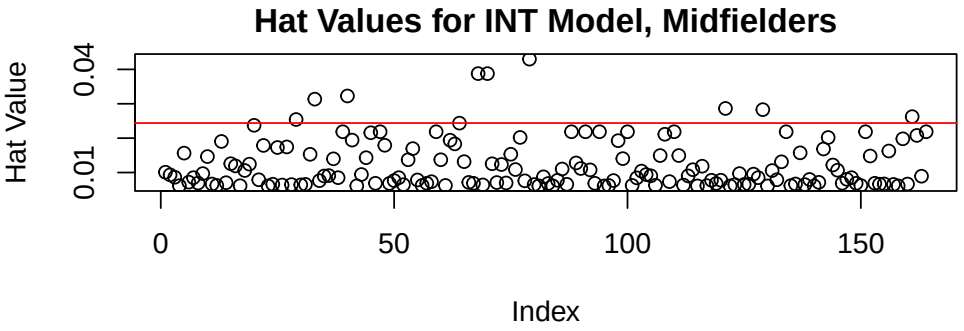
3.1.2 Hypothesis 2 - Forwards



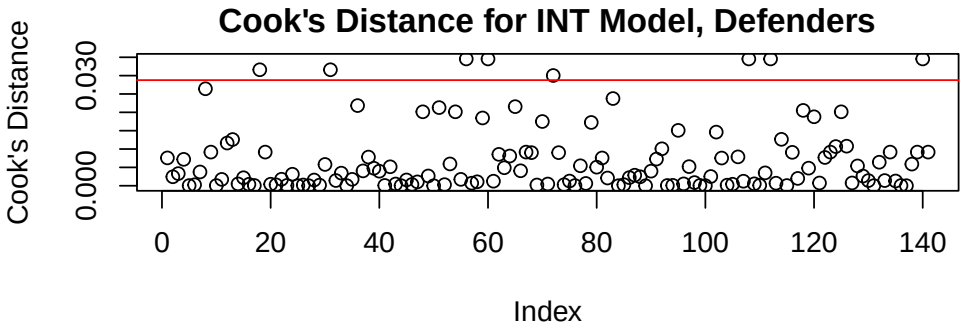


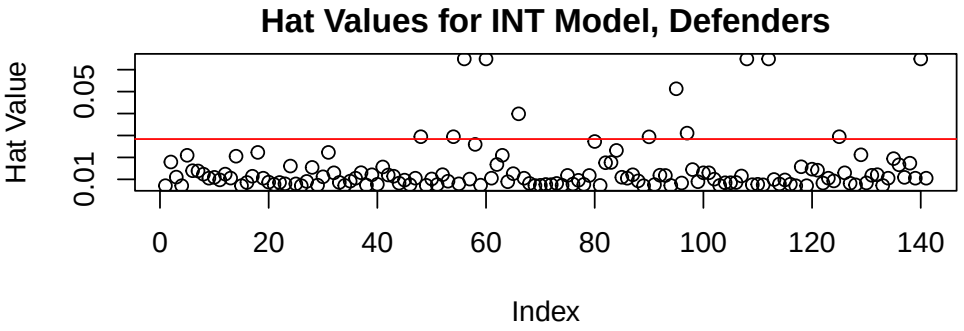
3.1.3 Hypothesis 2 - Midfielders





3.1.4 Hypothesis 2 - Defenders

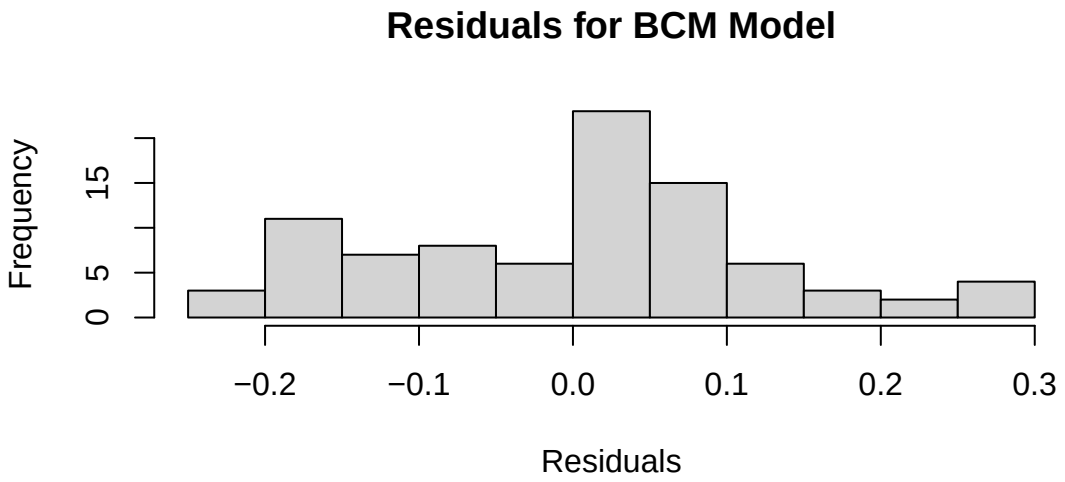




For all of the models for both hypotheses, we can see that after adjusting for outliers, there are no egregiously high Cook's Distance values or hat values. The most notable value is a point for the Forwards model with Cook's Distance of ~0.6 and a hat value of ~0.35. However, this did not impact our assumptions or model, and even if there are some deviations above the proposed line, none of them were strong enough to influence the outlier, as we can see when we check the OLS assumptions in the next section. As we already removed outliers at the beginning of the modeling process, we do not want to unnecessarily continue to remove outliers to fit our proposed model.

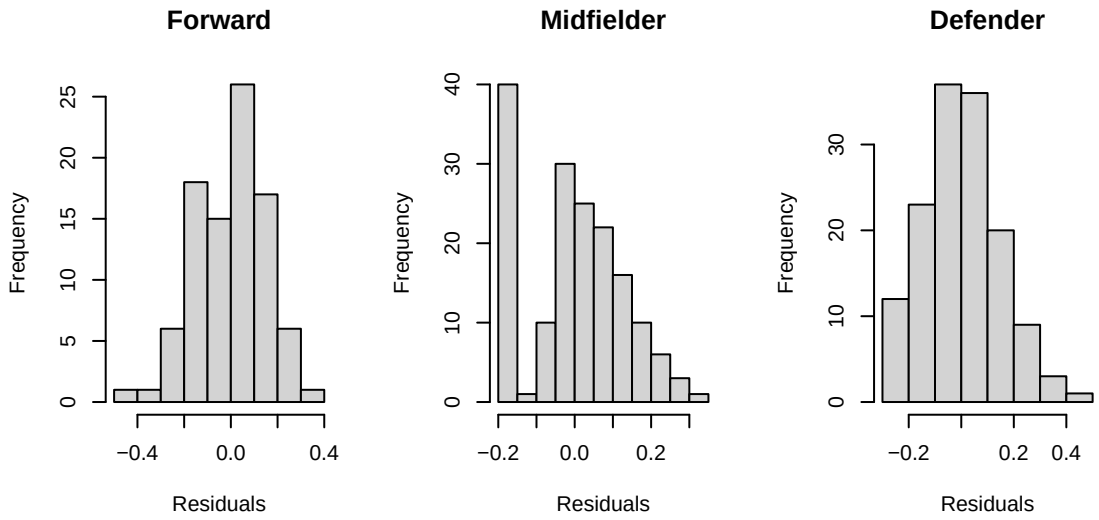
3.2 Histogram of Residuals for Final Models

3.2.1 Hypothesis 1



We can see that the residuals for the model for hypothesis 1 seem to be normally distributed with no extreme peaks or outliers, and we will confirm this when we run further diagnostic tests.

3.2.2 Hypothesis 2



We can see that the residuals for the model for forwards seem to be normally distributed with no extreme peaks or outliers. However, the residuals for the model for midfielders seems to have an extreme right skew, almost like it is missing the left side of the distribution. In addition, the residuals for the model for defenders seem to have a very slight right skew as well. All of these claims will be tested through diagnostic checking.

3.3 OLS Assumptions Tests

Table 2: P-values for OLS Assumption Tests

	Shapiro-Wilk	Breusch-Pagan	Durbin-Watson
Loss % Model	0.0620003	0.8133701	0.0000277
Big Chances Model (Forward)	0.0846857	0.8887937	0.2858406
Big Chances Model (Midfielder)	0.0000012	0.0000450	0.1265090
Big Chances Model (Defender)	0.0058248	0.1901617	0.7336850

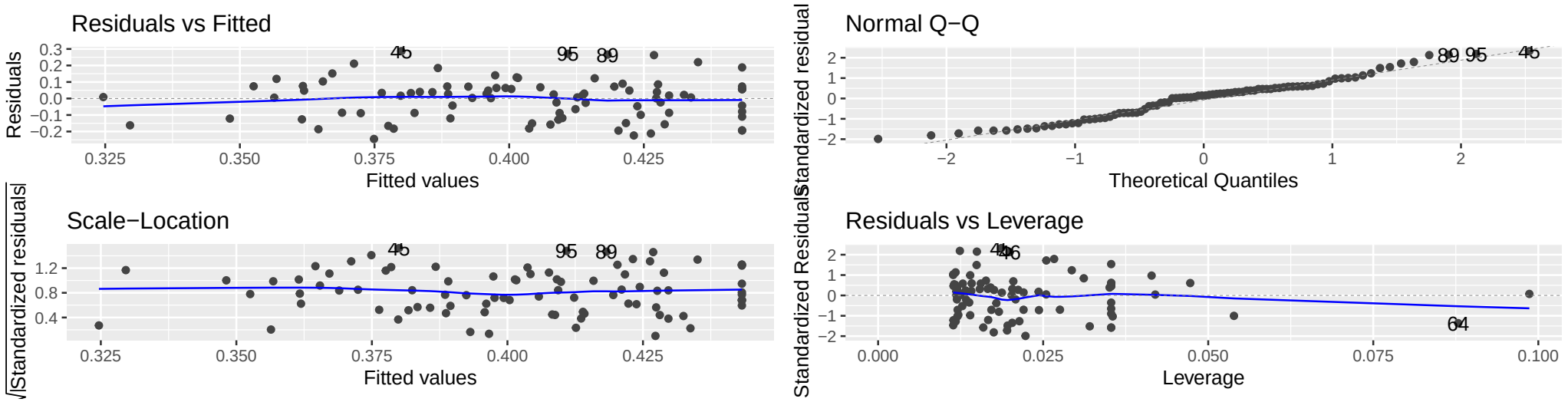
The Shapiro-Test checks whether the residuals are normally distributed, as we visualized in the previous section. As expected, the model for hypothesis 1 and the model for forwards both pass the normality assumption, but the midfielder and defender models do not, indicating that even after these transformations, a linear model may not be the best model to fit to this data.

The Breusch-Pagan test checks for heteroscedasticity with H_0 that the residuals have constant variance (homoscedasticity). If the p-value is small (ex. < 0.05), we would reject the null hypothesis, indicating heteroscedasticity. For our data, only one of the models fails this test, which is the midfielders model that also failed the Shapiro-Wilk test. This indicates that even after performing many transformations, a linear model may not be suited for the midfielder data.

The Durbin-Watson test for autocorrelation checks for a trend in the order of the data collection. In our case, only the model for hypothesis 1 fails this test. However, we do not believe it is enough to render the model unusable. Since we are only modeling loss % based on big chances missed per game, we end up ignoring many of the other important factors that can affect a player's performance such as player fatigue, weather conditions, or team strategy. These unaccounted variables might affect a player's or team's performance in a way that we cannot explain with this model alone. This is simply a limitation of our narrow hypothesis and research question, so we do not believe it is a flaw in the model itself.

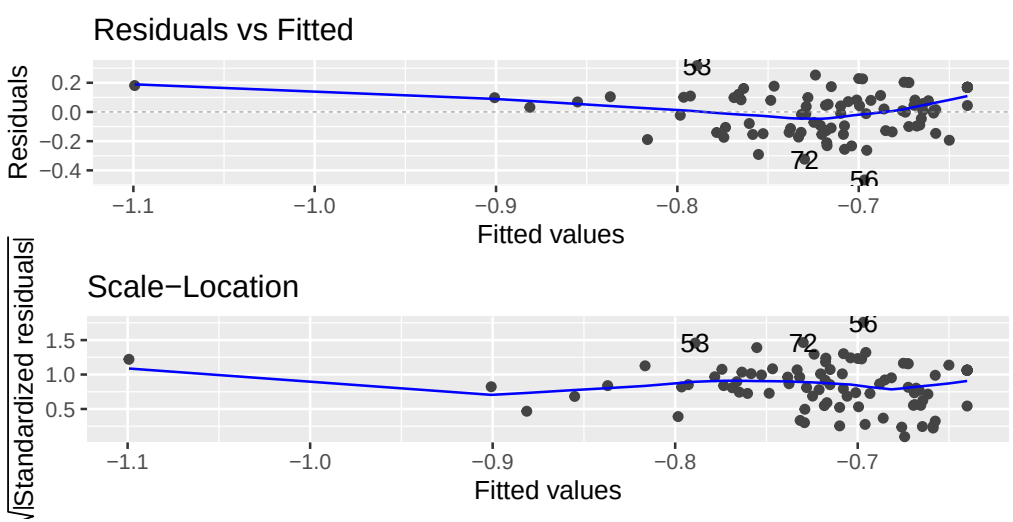
3.4 Final Checks

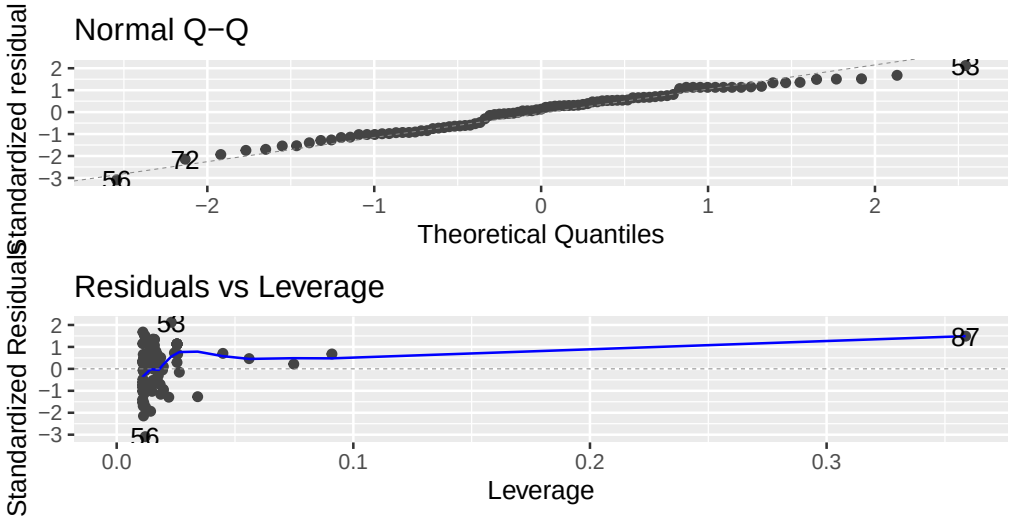
3.4.1 Hypothesis 1



We can see from the model evaluation plots that our assumptions are most likely satisfied. We can see that the the first plot of residuals versus fitted values shows no any distinct or unique patterns, and has relatively equal variability throughout, indicating that the linearity assumption is being fulfilled. However, there are still some values towards the right side that are potentially troubling. Either way, we have performed many transformations and model checks by this point, so it seems that this model performs decently. Furthermore, the Q-Q plot of the residuals indicates that the normality assumption is also being met since the points seem to follow the trend line. The Scale-Location plot shows variability along the standardized residuals, across the fitted values, through which we can conclude that the assumption of homoscedasticity is also being met. Finally, there are still some high leverage points, but nothing too extreme that would affect our results.

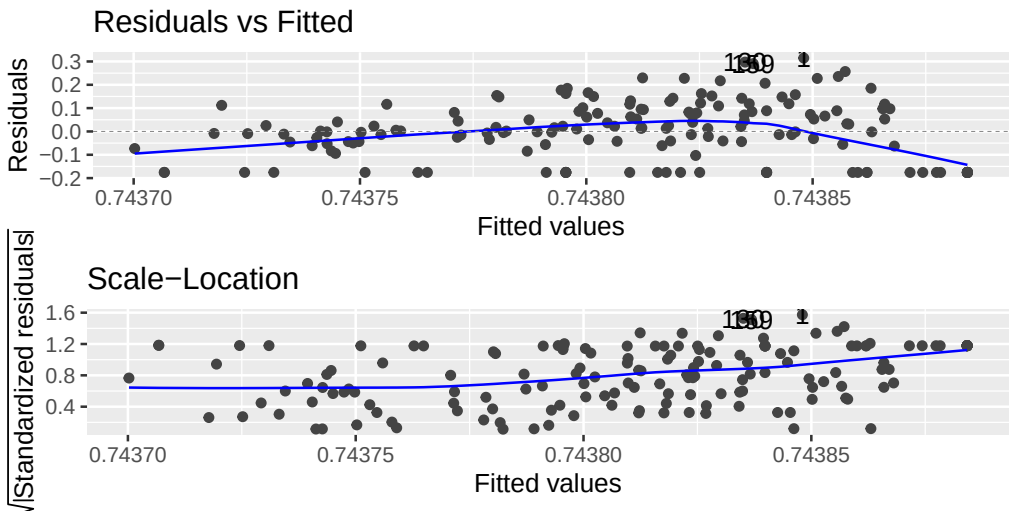
3.4.2 Hypothesis 2 - Forwards

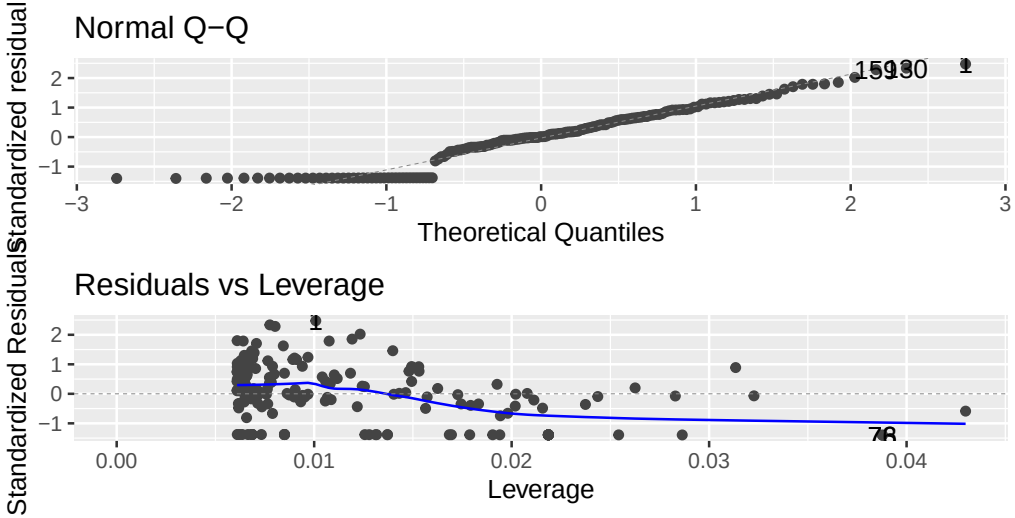




The plots for the model for forwards show similar results to those of hypothesis 1, indicating that many of our assumptions are likely to be satisfied. However, there is one high leverage point with a leverage of about 0.35 as discussed before. Since it did not seem to affect the model and all other assumptions were met, we can safely move forward and still include this point.

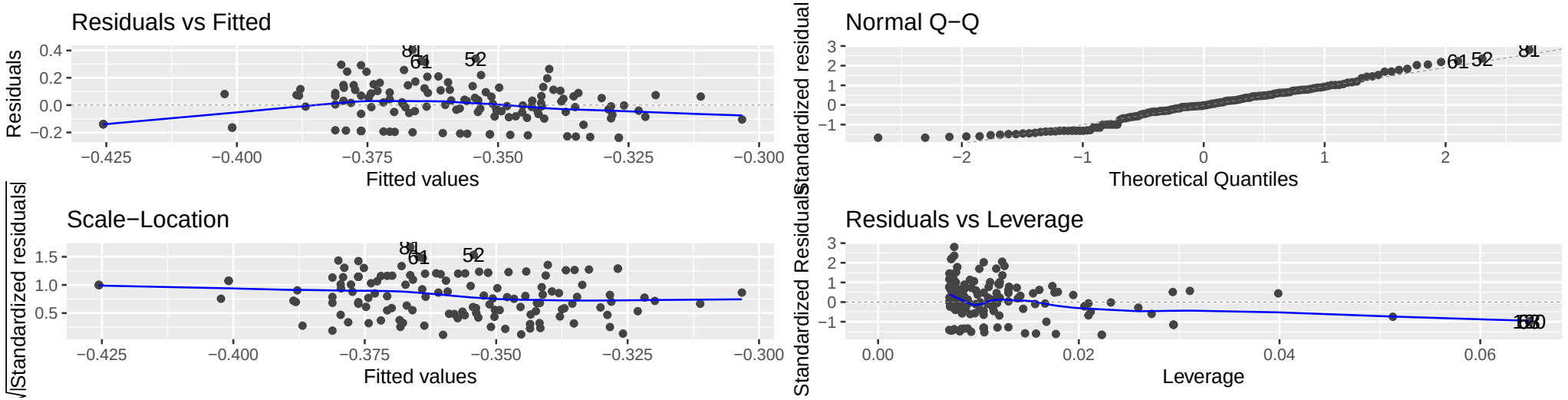
3.4.3 Hypothesis 2 - Midfielders





The plots for the model for midfielders are a little more problematic, as the normal Q-Q plot still does not sufficiently follow the trend line. We saw that the normality assumption was violated as the Shapiro-Test failed. However, all of the other plots seem to be sufficient. Thus, we can move forward with this plot with the assumptions violated knowing that we performed a Box-Cox transformation and other transformations on the model to try to create a better fit, but still could not get it to work.

3.4.4 Hypothesis 2 - Defenders



The plots for the model for defenders makes it seem that all of the assumptions would be satisfied, but we saw that this actually failed the Shapiro-Test at $\alpha = 0.05$ level of significance. Thus, even though the Normal Q-Q plot seems to follow the trend line quite closely by inspection, it is not good enough. However, we would like to note that the Shapiro p-value was much lower for the initial linear model, and we were able to get it much closer to the α level by performing the Box-Cox transformation. Even though it did not pass all of the assumptions, we will still move forward with this model for this project as we have not learned how to fit other models.

4 Results

4.1 Hypothesis 1

Table 3: Linear Model for Loss % by Big Chances Missed, Forwards

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4432	0.02341	18.93	2.008e-32
Chances_Missed_pg	-0.2372	0.1122	-2.115	0.03736

The coefficient for the intercept is 0.4432. This means that for any given forward, we would expect their loss percentage to be 44.3%. The coefficient in the linear model for the big chances missed per game is -0.2372, indicating that for each extra chance missed, on average we would expect any given forward to lose 23.7% fewer games over their career. While this may not seem intuitive at first, we can explain this result by considering that players who miss more chances tend to play on better teams with players that can feed them those chances. Thus, the better the team, the more chances will be created, and more of those may be missed. This is also a statistically significant result as the p-value of 0.037 is less than our α value of 0.05. These results are also viable as we have not transformed the data outside of removing outliers and performing a Box-Cox transformation to pass the normality assumption.

4.2 Hypothesis 2

Table 4: Linear Model for Big Chances Created by INT/G, Forwards

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.6401	0.02419	-26.46	6.021e-44
Interceptions_pg	-0.2794	0.06546	-4.268	4.919e-05

There is no real world interpretation of these coefficients as they are negative, and a player cannot make less than 0 interceptions per game or create less than 0 big chances per game. However, we can still find a clear trend that is confirmed by the strong p-value of $4.92 * 10^{-5}$, which is much less than the α significance level of 0.05. Since both coefficients are pointing in the same direction and the result is statistically significant, we can conclude that if a forward intercepts the ball more often, then they will create more big chances.

Table 5: Linear Model for Big Chances Created by INT/G, Midfielders

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7439	0.0189	39.36	7.755e-85

	Estimate	Std. Error	t value	Pr(>|t|)
Interceptions_pg	-8.871e-05	0.01958	-0.004531	0.9964

There is no real world interpretation of these coefficients due to our transformations. In addition, we do not find a clear trend due to the weak p-value of 0.9964, which is much greater than the α significance level of 0.05. Finally, we can notice that the coefficient for the main effect term is very low and close to 0, meaning that there is probably no relationship between these variables.

Table 6: Linear Model for Big Chances Created by INT/G, Defenders

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.4256	0.03689	-11.54	5.107e-22
Interceptions_pg	0.04935	0.02633	1.874	0.06299

For the model for defenders, we do not find a clear and obvious trend due to the weak p-value of 0.063, which is a little greater than the α significance level of 0.05, along with the fact that the coefficients have opposite signs.

4.3 Model Diagnostics

4.3.1 Hypothesis 1

Table 7: Model Evaluation for the BCM Model

	RMSE	R.Squared	p.Value
value	0.2732188	0.0494242	0.0373594

The RMSE of 0.2732 for the big chances missed model indicates that the model predictions deviate from the actual values by around 27.32% on average. This relatively high RMSE suggests poor predictive accuracy, but it should be noted that the inverse transformation may have affected the model's ability to predict against the real values. The R^2 value of 0.05 is low, suggesting that not much of the variation in the data is explained by the model. The p-value of 0.037 is less than 0.05, meaning that at an $\alpha = 0.05$ level of significance, there is sufficient evidence to conclude that big chances missed has an effect on loss percentage based on this model.

4.3.2 Hypothesis 2

4.3.2.1 RMSE

Table 8: RMSE for the BCC vs INT Model by Position

Forward	Midfielder	Defender
0.8392019	0.6548232	0.4289456

All three models have very high RMSE values, indicating that the model predictions deviate a lot from the actual values. Once again, it should be noted that the inverse transformation may have affected the model's ability to predict against the real values.

4.3.2.2 R Squared

Table 9: R-Squared for BCC vs INT Model by Position

Forward	Midfielder	Defender
0.1698786	1e-07	0.02465

The R^2 values for midfielders and defenders are both very low - both are below 0.03, indicating that these models explain less than 3% of the variation in the data. However, the model for forwards has a relatively high R^2 value of 0.1698, indicating that approximately 17% of the variation in the data is explained by the model. This is highly impressive because our dataset has 59 total variables, so for one variable to predict almost one fifth of all of the variability in the data is exactly what we were looking for.

4.3.2.3 p-value

Table 10: P-Values for BCC vs INT Model by Position

	Forward	Midfielder	Defender
value	4.92e-05	0.9963902	0.0629901

The p-value for forwards is $4.92 * 10^{-5}$, which is less than our chosen α significance level of 0.05. This means that there is strong evidence that interceptions per game have a statistically significant effect on big chances created per game for forwards.

The p-value for midfielders is 0.996, which is much greater than 0.05. This means that there is no statistically significant relationship between interceptions per game and big chances created per game for midfielders. The evidence is insufficient to conclude that interceptions have an effect on big chances for midfielders.

The p-value for defenders is 0.063, which is greater than 0.05. This indicates that there is insufficient evidence that interceptions per game have a statistically significant effect on big chances created per game for defenders.

5 Conclusion and Recommendations

5.1 Summary

After analyzing the data through linear regression and data visualization, there are some key takeaways and findings that can help teams improve their likelihood to win. We found that missed chances had a statistically significant impact on loss percentage, showing a clear negative correlation. We can explain this result by considering that players who miss more chances tend to play on better teams with players that can feed them those chances. Thus, the better the team, the more chances will be created, and more of those may be missed.

In addition, the data for forwards showed a statistically significant relationship between interceptions and big chance creation, meaning that defensive actions by forwards lead to more scoring opportunities (even though we observed negative coefficients, reversing the transformations we performed would give us interpretable output as both coefficients would then be negative and pointing in the same direction) Therefore, it is in a team's best interest to try and create defensive interceptions from these positions.

Surprisingly, midfielders and defenders both showed no significant relationship between interceptions and big chance creation. However, midfielders have the highest baseline big chance creation, showing that they naturally create big chances regardless of defensive actions. As a result of this, it may be better for midfielders to focus on other aspects of the game, while defenders should focus on their defensive contributions rather than trying to create big chances.

When looking at the key takeaways and findings from the data, there are some important recommendations to consider as a coach preparing his team for the season. Defenders need to be proactive on defense and try to get interceptions when they can, transitioning defensive plays into scoring opportunities. However, this does not mean that they need to create big chances because midfielders can handle that role. Midfielders should prioritize positioning and more offensive aspects, rather than defensive interceptions, because they already have a high baseline level of big chance creation.

Finally, forwards should both be able to press high up the field to win the ball through interceptions and create chances, while also being willing and confident in taking shots, not having to worry about missing chances. Our main takeaway from the statistical findings for forwards was that forwards should be defensively minded and relentless in their pursuit of the opposition, rather than being lazy and letting the other players handle defensive duties.

5.2 Limitations

Following from our research questions and dataset, there are a couple of limitations to our data analysis.

First, the dataset is not fully complete and includes lots of missing values. We had to summarize a small section of the data and cut out many observations that did not match our exact criteria.

Next, we only had time and space in our report to analyze a couple of the variables. Not only that, we had preconceived notions about what trends we expected to see, which may have influenced our model creation. For a further look into this dataset, we would try to incorporate more variables into the analysis and try to visualize more of the data at the beginning instead of forming our research questions first and hoping that the data would follow some trends as we expected.

For the future, some possible upgrades to this analysis model could be creating an R Shiny app that allows the user to select which variables from the dataset they would like to compare to each other, along with graphs that display and sort each statistic by position. It could even automatically perform hypothesis tests and create multiple best-fit linear models.

In conclusion, our hope would be that additional exploration into this data could be useful for any football teams looking to learn from the biggest clubs and best players in the world. The more variables that could be analyzed out of this dataset, the more likely it is that we could find some meaningful trends or outliers. Then, teams could look at what exactly makes those outliers - each of which is an individual player - unique or why they stand out. The goal would be to make the data more accessible and definitely more interactive so that people with little knowledge of data science could still access it.