Multiple Regression Analysis: European Soccer League Statistics

Data 603: Statistical Modeling with Data
University of Calgary
Winter 2024

James Ding
Josh Brauner

Introduction

Motivation

Soccer, or football in many parts of the world, is a globally cherished sport, captivating millions with its universal appeal. Beyond its cultural significance, soccer's allure extends into the realms of analytics and sports betting, captivating enthusiasts, and professionals alike.

The surge in soccer analytics is driven by a quest to understand player performance and team success intricacies. Simultaneously, sports betting has transformed into a strategic arena where data-driven insights offer competitive advantages. It is at this intersection of analytics and passion for the sport that our project gains momentum.

Project Context and Applied Domains

This project examines factors influencing players' overall ratings in European soccer leagues from 2008 to 2016, focusing on demographics and performance metrics. By doing so, we aim to uncover key determinants of player success and offer insights into forecasting their trajectory within the sport.

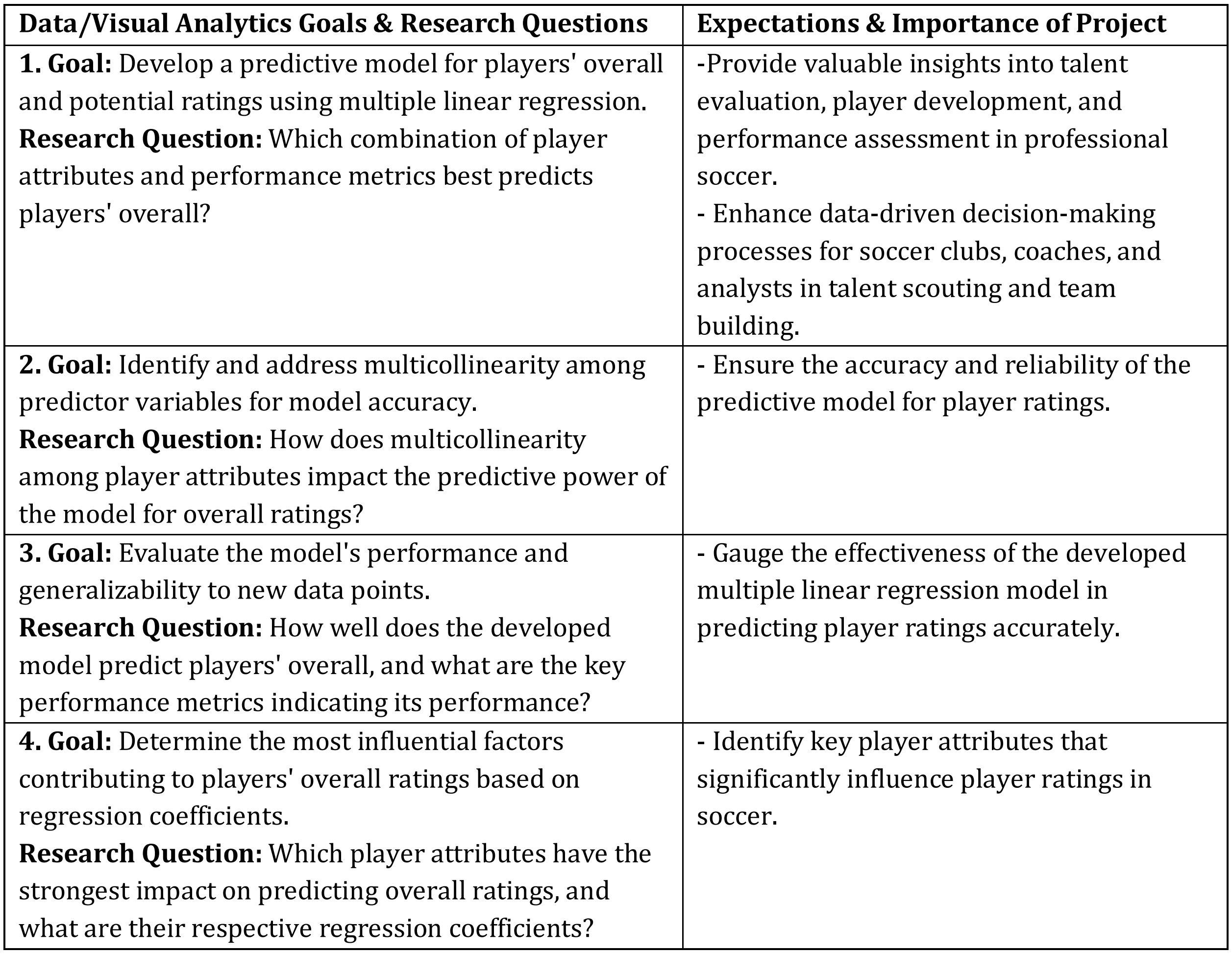
Although many studies were conducted in the past within the industry such as tactical and performance evaluation, financial and market analysis, fan engagement strategies and more, there was not any that solely focused on the overall rating of the players.

Situated at the intersection of sports analytics and predictive modeling, our project addresses critical questions in talent identification and performance evaluation within professional soccer. Through extensive analysis of player attributes, we aim to extract actionable insights, providing strategic guidance to stakeholders across the soccer ecosystem.

Project Focus: Understanding Metrics' Contribution to Player Overall Score

The primary focus of our project is to understand how different metrics contribute to the overall score of a player in European soccer leagues. By analyzing a range of factors such as player attributes, performance metrics, and game intelligence indicators, we aim to unravel the nuances that influence a player's overall rating. This comprehensive analysis will provide valuable insights into talent evaluation, player development strategies, and performance assessment within professional soccer. Having a better understanding of the most important factors that contribute to a player’s overall rating would allow a more thorough understanding of different metrics to focus on when developing players. This impacts both the teams when picking players and fans that are betting on the sport.

Data/Visual Analytics Goal & Research Questions



Data/Visual Analytics Goals & Research Questions	Expectations & Importance of Project
1. Goal: Develop a predictive model for players' overall and potential ratings using multiple linear regression. **Research Question:** Which combination of player attributes and performance metrics best predicts players' overall?	-Provide valuable insights into talent evaluation, player development, and performance assessment in professional soccer. - Enhance data-driven decision-making processes for soccer clubs, coaches, and analysts in talent scouting and team building.
2. Goal: Identify and address multicollinearity among predictor variables for model accuracy. **Research Question:** How does multicollinearity among player attributes impact the predictive power of the model for overall ratings?	- Ensure the accuracy and reliability of the predictive model for player ratings.
3. Goal: Evaluate the model's performance and generalizability to new data points. **Research Question:** How well does the developed model predict players' overall, and what are the key performance metrics indicating its performance?	- Gauge the effectiveness of the developed multiple linear regression model in predicting player ratings accurately.
4. Goal: Determine the most influential factors contributing to players' overall ratings based on regression coefficients. **Research Question:** Which player attributes have the strongest impact on predicting overall ratings, and what are their respective regression coefficients?	- Identify key player attributes that significantly influence player ratings in soccer.

Figure 1: Comparison between data/visual analytics goals and research questions and project's expectations

This project is particularly meaningful to us because we derive great enjoyment from analyzing sports data, and we envision transitioning this passion into a career in sports analytics. By delving into the intricacies of player performance metrics and ratings in soccer, we not only deepen our understanding of the game but also hone our analytical skills crucial for our career aspirations. This project serves as a steppingstone towards our goal of making impactful contributions to the sports industry through data-driven insights, innovative analysis techniques, and strategic decision-making frameworks that can drive success on and off the field. Furthermore, this project aims to be a framework that can be adapted into multiple other sports. This can range from other team sports such as basketball and hockey to individual sports such as figure skating and gymnastics. By identifying and tracking key metrics, we hope that this project allows for accurate prediction of performance for any sport.

Methodology

Data

We acquired our dataset from Kaggle, a renowned platform known for hosting a diverse range of datasets and machine learning competitions. Utilizing SQLite, a lightweight relational database management system, we extracted relevant data from the provided database on Kaggle. Specifically, we focused on two key tables: 'Player' and 'Player Attribute,' combining them based on their primary keys to create a more comprehensive dataset centered around European Soccer League players and their attributes spanning the years 2008-2016.

After extracting the necessary data, we transferred it to Jupyter Notebook, a popular environment for interactive coding and data analysis. Within Jupyter Notebook, we leveraged the powerful data manipulation capabilities of the pandas library in Python to clean and preprocess the dataset. Our cleaning process involved handling missing values by dropping null entries, remove duplicate rows with the same player's name, resulting in only using their latest data taken in 2016. In addition, we removed unnecessary columns such as `player_api_id`, `id_x`, and several more unnecessary columns. Moreover, we decided to remove goalkeeper statistics and goalkeepers due to the uniqueness of this position. We determined that a player is a goalkeeper via their goalkeeping statistics being higher than 60, we determined the threshold through researching player names around that benchmark and learning about their positions via Wikipedia.

As part of our data preparation phase, we also engineered a new feature called 'age' by subtracting players' birthdates from the date the players' data was recorded, providing valuable age-related insights for our analysis. With a refined and structured dataset in hand, we exported it to a CSV format and seamlessly imported it into R for further statistical analysis and modeling, marking the beginning of our in-depth exploration into the dynamics of European soccer player attributes and performance metrics.

In conclusion, our dataset contains one csv file, and it details European Soccer League players' biometrics and various soccer related attributes that will be explained in detail later in the variable section.

Variable Explanations and Data Assumptions

The European Soccer League data we obtained from Kaggle was recorded via live matches. According to Premier League, "Live data is collected by a three-person team covering each match. Two highly trained analysts use a proprietary video-based collection system to gather information on what happens every time a player touches the ball, which player it was and where on the pitch the action occurred." (premierleague.com) In addition, players' biometrics were obtained through routine medical assessments conducted by club medical staff and administrative processes during player registration.

After cleaning the dataset, we have the following variables:

1. **Overall Rating** – A player's overall rating within the league, standardized by FIFA. *Dependent Variable **(1-100)**
2. **Height**: Player's height in **centimeters.** *Independent Variable
3. **Weight**: Player's weight in **kilograms.** *Independent Variable
4. **Preferred Foot**: Player's dominant foot **(left or right).** *Independent Variable
5. **Attacking Work Rate**: Player's level of attacking involvement **(low, medium, high)**. *Independent Variable
6. **Defensive Work Rate**: Player's level of defensive involvement **(low, medium, high)**. *Independent Variable
7. **Crossing**: Player's skill level in crossing the ball. *Independent Variable **(1-100)**
8. **Finishing**: Player's skill level in finishing goal-scoring opportunities. *Independent Variable **(1-100)**
9. **Heading Accuracy**: Player's accuracy in heading the ball. *Independent Variable **(1-100)**
10. **Short Passing**: Player's skill level in short passing. *Independent Variable **(1-100)**
11. **Volleys**: Player's skill level in striking the ball in the air without letting it touch the ground. *Independent Variable **(1-100)**
12. **Dribbling**: Player's skill level in maintaining control of the ball while moving. *Independent Variable **(1-100)**
13. **Curve**: Player's ability to curve the ball during passes or shots. *Independent Variable **(1-100)**
14. **Free Kick Accuracy**: Player's accuracy in free kicks. *Independent Variable **(1-100)**
15. **Long Passing**: Player's skill level in long passing. *Independent Variable **(1-100)**
16. **Ball Control**: Player's ability to control the ball while dribbling. *Independent Variable **(1-100)**
17. **Acceleration**: Player's ability to reach high speeds quickly. *Independent Variable **(1-100)**
18. **Sprint Speed**: Player's maximum running speed. *Independent Variable **(1-100)**
19. **Agility**: Player's agility and ability to change direction quickly. *Independent Variable **(1-100)**
20. **Reactions**: Player's reaction time to events on the field. *Independent Variable **(1-100)**
21. **Balance**: Player's ability to maintain balance while moving. *Independent Variable **(1-100)**
22. **Shot Power**: Player's power in shooting the ball. *Independent Variable **(1-100)**
23. **Jumping**: Player's ability to jump for headers or to reach higher balls. *Independent Variable **(1-100)**
24. **Stamina**: Player's endurance and ability to maintain performance over time. *Independent Variable
25. **Strength**: Player's physical strength. *Independent Variable

26. **Long Shots**: Player's accuracy and power in long-range shots. *Independent Variable
27. **Aggression**: Player's level of aggression in challenges. *Independent Variable
28. **Interceptions**: Player's ability to intercept passes or crosses. *Independent Variable
29. **Positioning**: Player's positional awareness on the field. *Independent Variable
30. **Vision**: Player's ability to see and exploit open spaces or passing opportunities. *Independent Variable
31. **Penalties**: Player's accuracy in penalty kicks. *Independent Variable
32. **Marking**: Player's ability to mark opponents closely. *Independent Variable
33. **Standing Tackle**: Player's ability to perform standing tackles. *Independent Variable
34. **Sliding Tackle**: Player's ability to perform sliding tackles. *Independent Variable
35. **Age**: Player's age in years. *Independent Variable

Modeling Plan

We plan to approach this project using the methods we have learned in Data 603. We will be doing multiple linear regression modelling.

1. Our alpha value is 0.05 for all our hypothesis tests.
2. We run a linear regression model using all predictors.
3. Use stepwise regression to recommend a model.
4. Use partial f test to compare our recommended model and linear model.
5. Depending on the p-value, we will determine whether to use our recommended or linear model. (Not enough evidence to determine that there was a difference between the models)
6. Remove variables with high multicollinearity from our linear model.
7. Use partial f test to compare our recommended model and linear model.
8. Depending on the outcome (Reduced model), We will use individual t-test to check each predictor to obtain best linear model.
9. Start using interaction and higher-order model.
10. Continue doing f-test to find the best model.
11. After finding the best model, we will test for the assumption.
    a. Linearity Assumption – Review residual plots
    b. Independence Assumption – Review residual plots
    c. Normality Assumption – Shapiro Wilk normality test
    d. Equal Variance Assumption (heteroscedasticity) - Breusch Pagan Test
    e. Multicollinearity – Using variance inflation factors (VIF). This test is done after stepwise regression so that interaction terms do not cause an unpredictable effect on the test.
    f. Outliers – Check Cook's distance & leverage
12. If our final perfect model does not satisfy any of these assumptions, we will try our best to transform the data. Once everything is passed, we will use the model to predict future European Soccer League players' overall rating.

Workload Plan:

Role of James Ding: Reports – Strength in writing documents

Role of Josh Brauner: Presentation – Strength in making power points that effectively convey our project.

Collaboration: We went over all the data, methodology, and results together before starting our respective tasks – Strength of both people as it is the most important part of our project, and it is to make sure we are on the same page when we split off

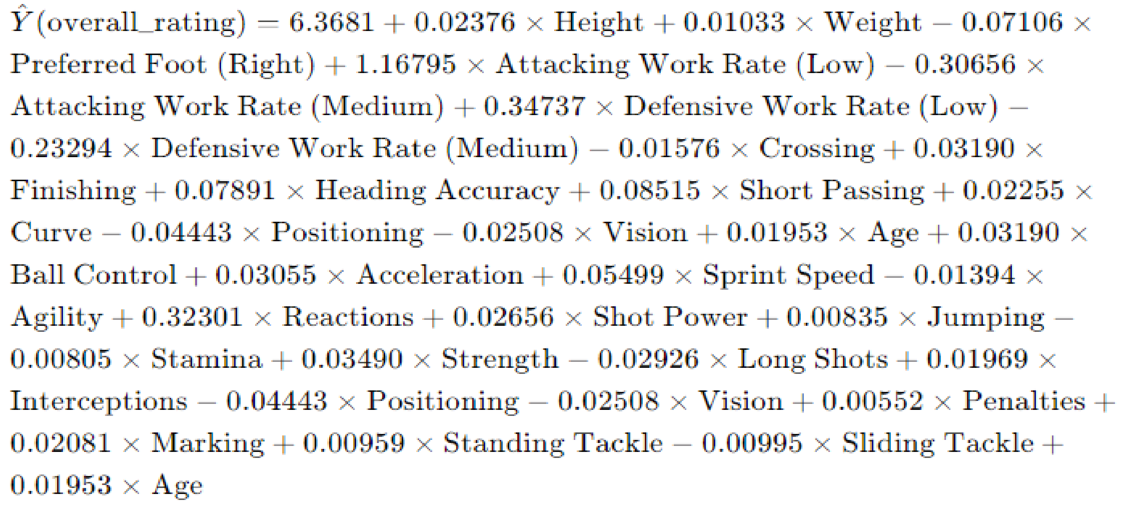
Results

Variable Selection:

First, we built a first-order model that included all the variables.

(linear model: adj r2: 0.787, RSE: 2.895)

First Order Model:



$$\hat{Y}(\text{overall\_rating}) = 6.3681 + 0.02376 \times \text{Height} + 0.01033 \times \text{Weight} - 0.07106 \times \text{Preferred Foot (Right)} + 1.16795 \times \text{Attacking Work Rate (Low)} - 0.30656 \times \text{Attacking Work Rate (Medium)} + 0.34737 \times \text{Defensive Work Rate (Low)} - 0.23294 \times \text{Defensive Work Rate (Medium)} - 0.01576 \times \text{Crossing} + 0.03190 \times \text{Finishing} + 0.07891 \times \text{Heading Accuracy} + 0.08515 \times \text{Short Passing} + 0.02255 \times \text{Curve} - 0.04443 \times \text{Positioning} - 0.02508 \times \text{Vision} + 0.01953 \times \text{Age} + 0.03190 \times \text{Ball Control} + 0.03055 \times \text{Acceleration} + 0.05499 \times \text{Sprint Speed} - 0.01394 \times \text{Agility} + 0.32301 \times \text{Reactions} + 0.02656 \times \text{Shot Power} + 0.00835 \times \text{Jumping} - 0.00805 \times \text{Stamina} + 0.03490 \times \text{Strength} - 0.02926 \times \text{Long Shots} + 0.01969 \times \text{Interceptions} - 0.04443 \times \text{Positioning} - 0.02508 \times \text{Vision} + 0.00552 \times \text{Penalties} + 0.02081 \times \text{Marking} + 0.00959 \times \text{Standing Tackle} - 0.00995 \times \text{Sliding Tackle} + 0.01953 \times \text{Age}$$

Stepwise & Partial F-test:

Null Hypothesis (H0): The recommended model (from stepwise) and the linear model provide an equally good fit to the data.
Alternative Hypothesis (H1): The recommended model provides a significantly better fit to the data than the reduced model.

We explored many regression selection procedures such as R2, adjusted R2, RSE, AIC and BIC. We decided to simply use stepwise regression to obtain a recommended model due to the vast number of predictors we have, we decided on p_enter = 0.05 and p_remove = 0.1 because we want to ensure accuracy in our model. After that, we did the partial f-test between our linear and recommended model. Unfortunately, the p-value turned out to be 0.2185. Therefore, we failed to reject the null hypothesis, indicating that the recommended model and the linear model provide an equally good fit to the data. (recommended model: adjr2: 0.7869, RSE:2.896)

Multicollinearity & Partial F-test:

Null Hypothesis (H0): The reduced model (after removing variables that caused collinearity) and the linear model provide an equally good fit to the data.
Alternative Hypothesis (H1): The reduced model provides a significantly better fit to the data than the linear model.

We moved on and started checking for multicollinearity via VIF in our linear model and removed the variables that caused collinearity which were `marking`, `sliding_tackle`, and `standing_tackle`. After removing the variables and obtaining our reduced model, we decided to run the partial f-test again. This time, our p-value from the f-test was 6.451e-06, which means we can reject the null hypothesis and it indicates that the reduced model provides a significantly better fit to the data than the linear model. (`reduced model`, adj.r2: 0.786, RSE: 2.899)

Individual T-Test:

Null Hypothesis (H0): All coefficients of the predictor variables are equal to zero, indicating that none of the predictors have a significant effect on the response variable.
Alternative Hypothesis (H1): At least one of the coefficients of the predictor variables is not equal to zero, indicating that at least one predictor has a significant effect on the response variable.

From our individual t-test, we would reject the null hypothesis in favor of the alternative, suggesting most of our variables are significant predictors of European Soccer League's players' overall rating. However, we noticed several predictors have p values that are higher than 0.05: `preferred_footright`(0.218), `volleys`(0.326), `dribbling`(0.164), `free_kick_accuracy`(0.122), balance`(0.948), `stamina`(0.523), `aggression`(0.070), and `penalties`(0.173). Therefore, we decided to take them out for our best fit linear model. (`best linear model`, adjr2: 0.786, RSE: 2.9) We then did a partial f-test again to compare our reduced model and best linear model.

Null Hypothesis (H0): The reduced model and the best linear model provide an equally good fit to the data.
Alternative Hypothesis (H1): The reduced model provides a significantly better fit to the data than the best linear model.

Unfortunately, we obtained a p value of 0.087 when comparing the two models. Although there was not enough evidence to reject the null hypothesis, we decided to keep moving on with the best linear model due to the small p-value it has because we believe this model would serve as a great basis for our interaction model.

Interaction Model:

After we have our best linear model, we decided to test for interaction on the model to determine if there are any interaction terms. (Interaction Model, adjR2: 0.929, RSE: 1.674) A strange observation we made while looking at the p-values of the interactions in the model was that `heading_accuracy``s interactions with other variables are all significant while none other variables share the same feat. We will keep that in mind when we do higher-order models.

Null Hypothesis (H0): The best linear model and the interaction model provide an equally good fit to the data.
Alternative Hypothesis (H1): The interaction model provides a significantly better fit to the data than the best linear model.

We compared our interaction model with our best linear model via partial f-test.
The result was expected as the p-value is <2.2e-16, meaning we reject our null hypothesis in favor of our alternative hypothesis.

At this point, we decided to remove all the insignificant predictors to see if it improves the model. After removing the insignificant interactions, we arrived at the adjusted interaction model (adjR2: 0.897, RSE: 2.011). We then do a partial f-test to compare the models, and we got a p-value of <2.2e-16, meaning that we reject the null hypothesis in favor of alternative, that the adjusted interaction model does fare better than the original despite having a lower adjR2 and higher RSE value. We decided to go with the adjusted interaction model because we value the significance of the predictors and the difference between two models’ adjR2 and RSE value is not that big.

Higher Power Model:

We went back to our best linear model in hopes of finding predictors that we can raise to a higher power. After inspection, there were no changes by raising predictors to a higher power. We decided to stick to our adjusted interaction model as our final model.

Final Model (Best fitted model):

$$
\begin{aligned}
&\hat{Y}(\text{overall rating}) = -23.34 + 0.2978 \times \text{Height} + 0.0742 \times \text{Weight} + 21.47 \times \\
&\text{Attacking Work Rate (Low)} + 1.608 \times \text{Attacking Work Rate (Medium)} + 3.624 \times \\
&\text{Defensive Work Rate (Low)} + 2.585 \times \text{Defensive Work Rate (Medium)} + 0.665 \times \\
&\text{Crossing} - 0.113 \times \text{Finishing} - 0.368 \times \text{Heading Accuracy} - 0.0836 \times \\
&\text{Short Passing} + 0.0079 \times \text{Curve} - 0.0713 \times \text{Long Passing} + 0.241 \times \\
&\text{Ball Control} + 0.0932 \times \text{Acceleration} - 0.1897 \times \text{Sprint Speed} + 0.0787 \times \\
&\text{Agility} + 0.0529 \times \text{Reactions} + 0.3998 \times \text{Shot Power} + 0.0513 \times \text{Jumping} - \\
&0.1793 \times \text{Strength} + 0.0657 \times \text{Long Shots} + 0.0262 \times \text{Interceptions} + 0.2473 \times \\
&\text{Positioning} - 0.1488 \times \text{Vision} + 0.00039 \times \text{Age} - 0.1188 \times \text{Height} \times \\
&\text{Attacking Work Rate (Low)} - 0.0026 \times \text{Height} \times \\
&\text{Attacking Work Rate (Medium)} - 0.0043 \times \text{Height} \times \text{Crossing} + 0.00078 \times \\
&\text{Height} \times \text{Finishing} - 0.0029 \times \text{Height} \times \text{Shot Power} + 0.0002 \times \text{Height} \times \\
&\text{Long Shots} + 0.002 \times \text{Height} \times \text{Interceptions} + 0.00076 \times \text{Weight} \times \text{Finishing} - \\
&0.00198 \times \text{Weight} \times \text{Positioning} + 0.0314 \times \text{Attacking Work Rate (Low)} \times \\
&\text{Finishing} + 0.00698 \times \text{Attacking Work Rate (Medium)} \times \text{Finishing} + 0.0646 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Heading Accuracy} - 0.0044 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Heading Accuracy} - 0.0033 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Curve} - 0.0003 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Curve} - 0.0365 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Sprint Speed} - 0.0437 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Sprint Speed} - 0.0313 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Reactions} + 0.0272 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Reactions} + 0.0208 \times \\
&\text{Defensive Work Rate (Low)} \times \text{Agility} - 0.00297 \times \\
&\text{Defensive Work Rate (Medium)} \times \text{Agility} - 0.0546 \times \\
&\text{Defensive Work Rate (Low)} \times \text{Interceptions} - 0.0376 \times \\
&\text{Defensive Work Rate (Medium)} \times \text{Interceptions} - 0.0335 \times \\
&\text{Defensive Work Rate (Low)} \times \text{Positioning} - 0.00302 \times \\
&\text{Defensive Work Rate (Medium)} \times \text{Positioning} - 0.0015 \times \text{Crossing} \times \text{Finishing} - \\
&0.0002 \times \text{Crossing} \times \text{Heading Accuracy} + 0.0029 \times \text{Crossing} \times \text{Sprint Speed} + \\
&0.00013 \times \text{Crossing} \times \text{Shot Power} - 0.00175 \times \text{Crossing} \times \text{Jumping} + 0.0027 \times \\
&\text{Crossing} \times \text{Vision} - 0.0010 \times \text{Finishing} \times \text{Heading Accuracy} + 0.0033 \times \\
&\text{Finishing} \times \text{Short Passing} - 0.005 \times \text{Finishing} \times \text{Interceptions} + 0.0018 \times \\
&\text{Finishing} \times \text{Positioning} + 0.0003 \times \text{Heading Accuracy} \times \text{Short Passing} + 0.0013 \times \\
&\text{Heading Accuracy} \times \text{Long Passing} - 0.0011 \times \text{Heading Accuracy} \times \text{Ball Control} - \\
&0.0012 \times \text{Heading Accuracy} \times \text{Acceleration} + 0.0018 \times \text{Heading Accuracy} \times \\
&\text{Sprint Speed} - 0.0012 \times \text{Heading Accuracy} \times \text{Agility} + 0.0018 \times \\
&\text{Heading Accuracy} \times \text{Reactions} + 0.0025 \times \text{Heading Accuracy} \times \text{Shot Power} + \\
&0.0011 \times \text{Heading Accuracy} \times \text{Jumping} + 0.0039 \times \text{Heading Accuracy} \times \\
\hat{}\quad &\text{Strength} - 0.0018 \times \text{Heading Accuracy} \times \text{Long Shots}
\end{aligned}
$$

Multiple Regression Assumptions:

This section will address how our model meets various assumptions. These assumptions are to ensure that our model result is trustworthy.

Linear Assumption:

Our model assumes linearity in nature. Using the residuals vs fitted plot, we check to see if there are any discernable patterns that are non-linear. From the plot, we see that there are no apparent patterns showing in the trend of our data, suggesting that it passes the linearity assumption.

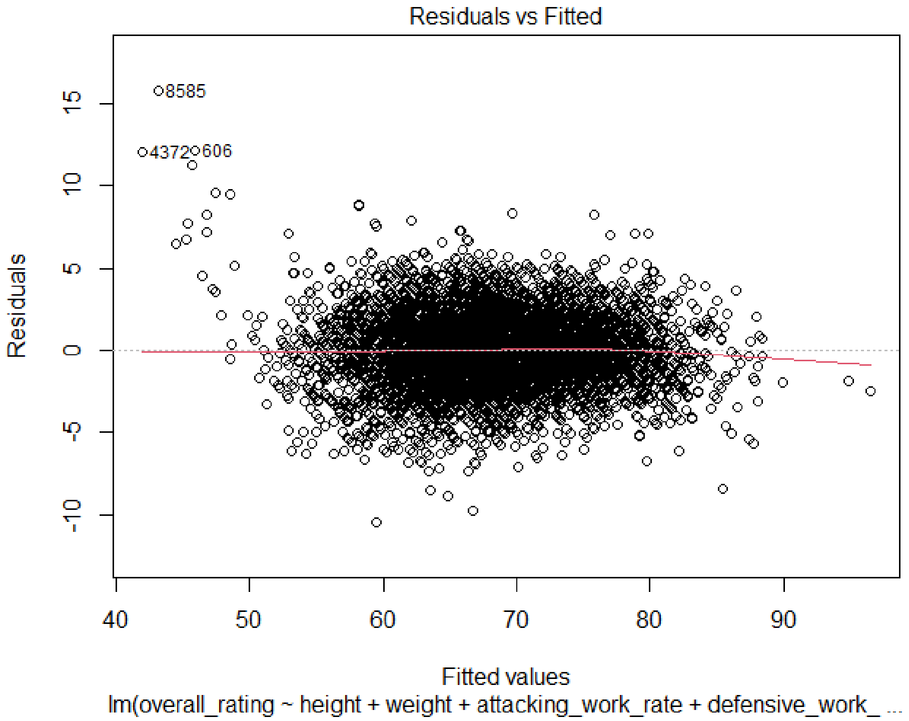


Figure 2: Residuals vs Fitted plot for determining linear assumption.

Independence Test:

Our model assumes independence of error terms. Using the residuals vs observations plot, we check to see if there are any discernable patterns. From the plot, we see that there are no apparent patterns showing in the trend of our data, suggesting that it passes the independence assumption.

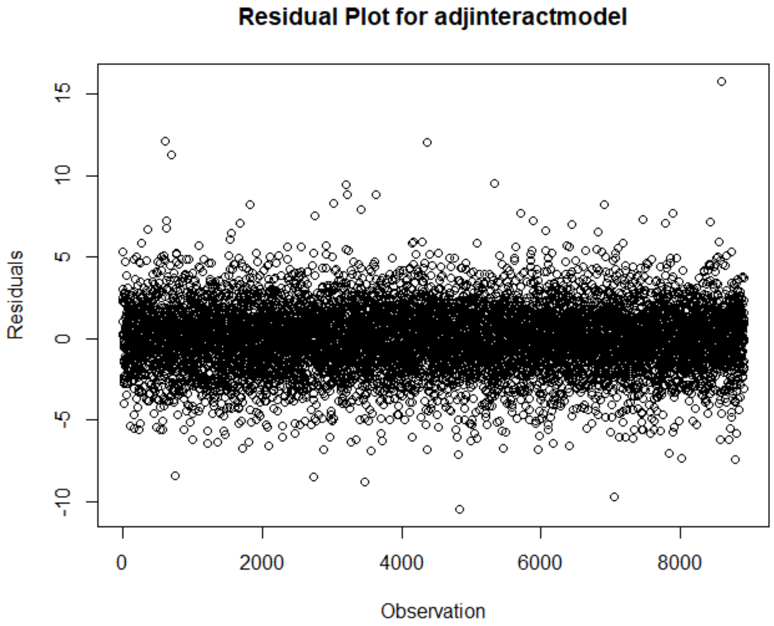


Figure 3: Residuals vs Observations plot for determining independence assumption.

Normality Test:

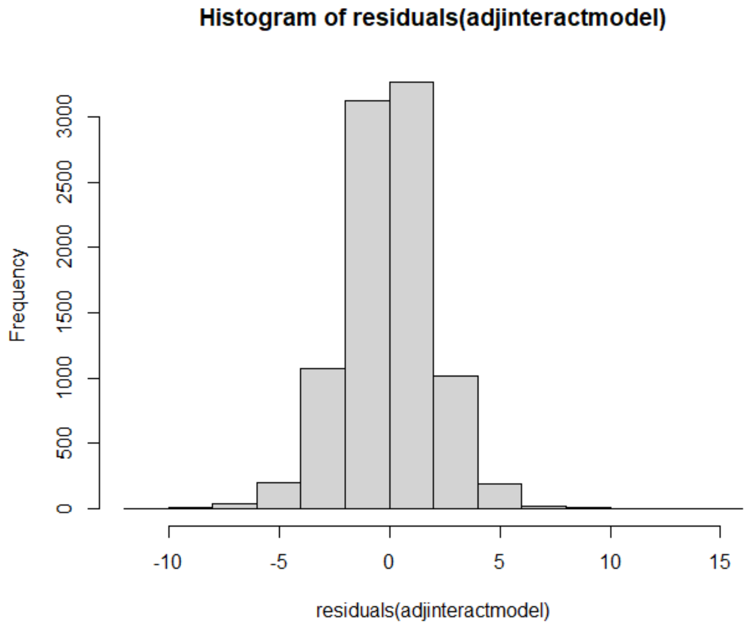


Figure 4: Histogram to check normality of the model.

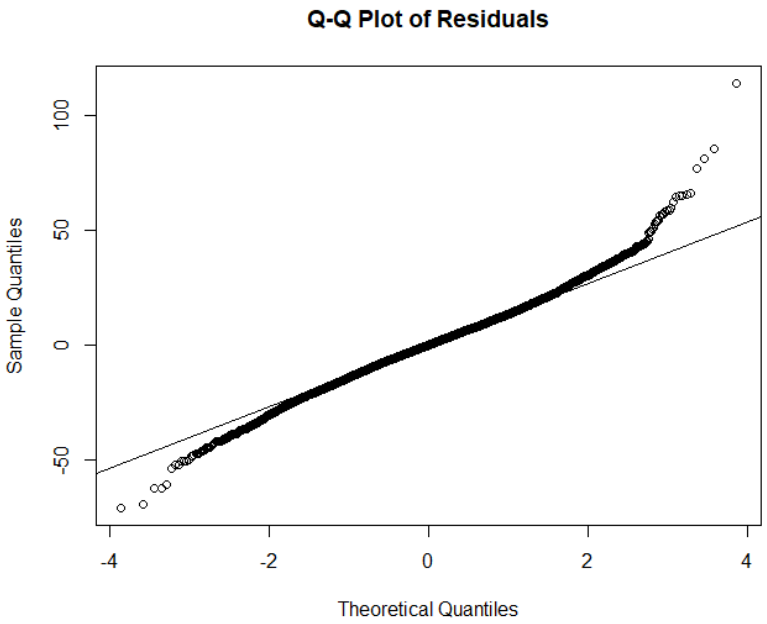


Figure 4: Q- Q plot to check the normality of the model.

Shapiro-Wilk Normality Test:

- Null Hypothesis (H0): The residuals are normally distributed.
- Alternative Hypothesis (H1): The residuals are not normally distributed.

Based on the histogram of residuals and Q-Q plot of residuals, we originally thought that the data was going pass the normality test.

However, when we implemented the Shapiro-Wilk Normality Test for normality assumption. Our p-value was 3.67e-17, which was lower than the provides evidence to reject the null hypothesis, suggesting that the residuals are not normally distributed. Quick sidenote that R only allowed (3 to 5000 sample size, so we had to randomized 5000 out of my data for the test) we tried this a couple more times, they all ended up being lower than 0.05 p-value. Then we performed the box-cox transformation, obtained a lambda 1.475, applied to the mode, p-value (4.562e-15) still came out lower than 0.05. After that, logarithmic transformation, resulting in p-value (< 2.2e-16) being lower than 0.05, we think at this point we just must accept the fact that our model will not pass the normality test.

In conclusion, the p-value of the Shapiro-Wilk Normality Test was lower than 0.05 despite our attempts to normalize our data via box-cox and logarithmic transformation. Therefore, we reject our null hypothesis, indicating that the residuals are not normally distributed in our model. (W = 0.988, p-value < 2.2e-16)

Equal Variance Assumption:

Ho: Heteroscedasticity is not present

H1: Heteroscedasticity is present.

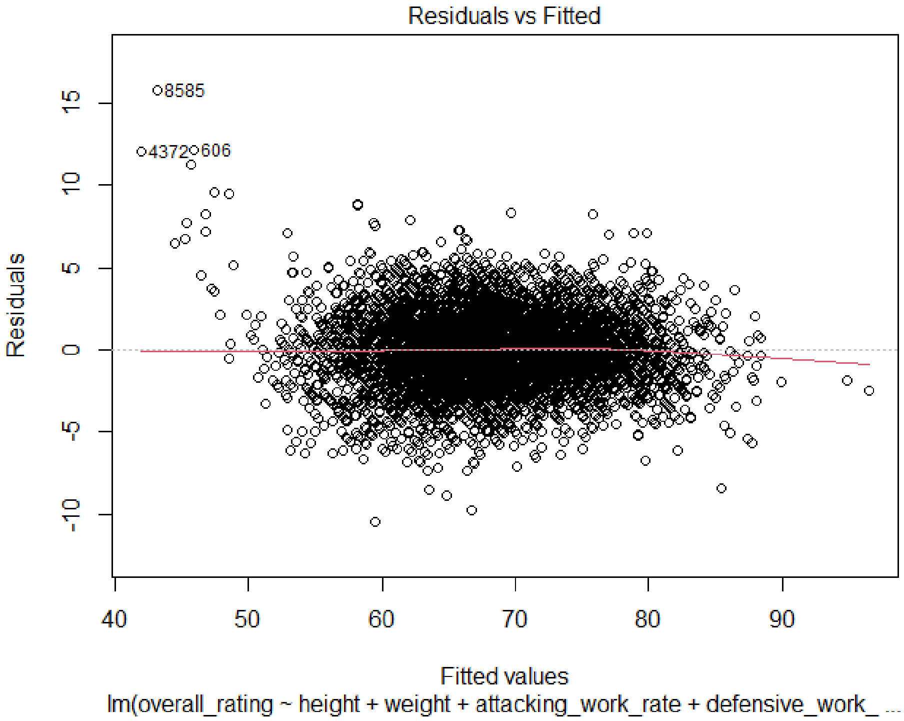


Figure 5: Residual vs Fitted plot for checking model's equal variance assumption.

For the equal variance assumption, we went straight to the Studentized Breusch-Pagan test after looking at the residual vs fitted plot. Again, we failed the equal variance assumption. According to our Studentized Breusch-Pagan test, we received values that BP = 1458.2 (1235 with box-cox transformed model), and p-value < 2.2e-16. Therefore, we reject our null hypothesis indicating that our model is heteroscedastic.

Influential Points & Outliers

Outliers can have a large effect on the model, we plot residuals vs leverage plot as well as cook's line to spot outliers.

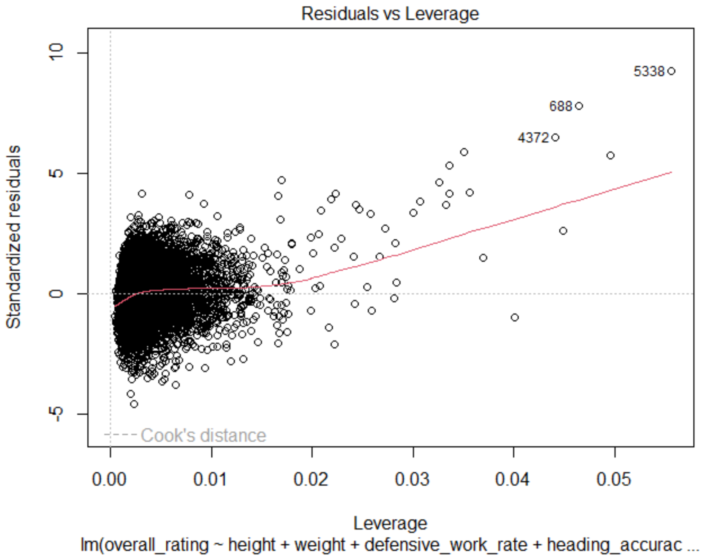


Figure 6: Residual vs leverage plot to determine if high leverage points are within Cook's Line

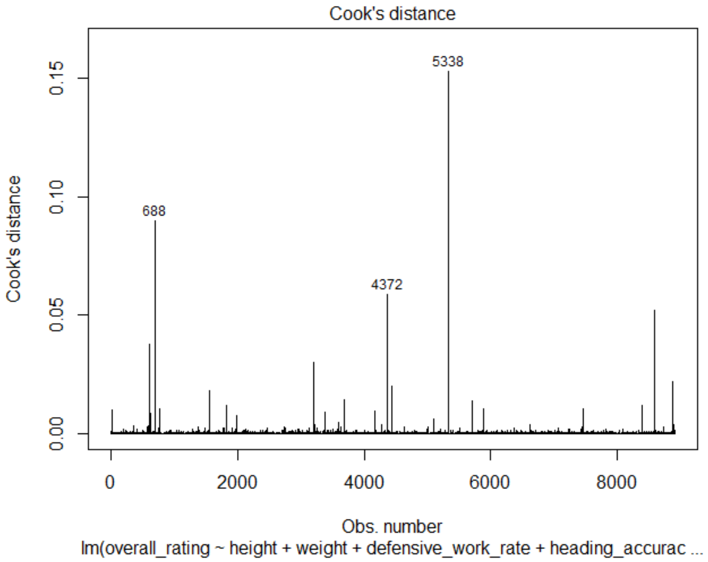


Figure 7: Cook's Distance portraying potential outliers.

As we can see from the graphs above, point 688, 4372, and 5338 are high on the leverage. However, there are not any points beyond Cook's Distance. This signifies that there are no influential points that have a disproportional effect on our regression result.

Final Model:

$$
\begin{aligned}
&\hat{Y}(\text{overall rating}) = -23.34 + 0.2978 \times \text{Height} + 0.0742 \times \text{Weight} + 21.47 \times \\
&\text{Attacking Work Rate (Low)} + 1.608 \times \text{Attacking Work Rate (Medium)} + 3.624 \times \\
&\text{Defensive Work Rate (Low)} + 2.585 \times \text{Defensive Work Rate (Medium)} + 0.665 \times \\
&\text{Crossing} - 0.113 \times \text{Finishing} - 0.368 \times \text{Heading Accuracy} - 0.0836 \times \\
&\text{Short Passing} + 0.0079 \times \text{Curve} - 0.0713 \times \text{Long Passing} + 0.241 \times \\
&\text{Ball Control} + 0.0932 \times \text{Acceleration} - 0.1897 \times \text{Sprint Speed} + 0.0787 \times \\
&\text{Agility} + 0.0529 \times \text{Reactions} + 0.3998 \times \text{Shot Power} + 0.0513 \times \text{Jumping} - \\
&0.1793 \times \text{Strength} + 0.0657 \times \text{Long Shots} + 0.0262 \times \text{Interceptions} + 0.2473 \times \\
&\text{Positioning} - 0.1488 \times \text{Vision} + 0.00039 \times \text{Age} - 0.1188 \times \text{Height} \times \\
&\text{Attacking Work Rate (Low)} - 0.0026 \times \text{Height} \times \\
&\text{Attacking Work Rate (Medium)} - 0.0043 \times \text{Height} \times \text{Crossing} + 0.00078 \times \\
&\text{Height} \times \text{Finishing} - 0.0029 \times \text{Height} \times \text{Shot Power} + 0.0002 \times \text{Height} \times \\
&\text{Long Shots} + 0.002 \times \text{Height} \times \text{Interceptions} + 0.00076 \times \text{Weight} \times \text{Finishing} - \\
&0.00198 \times \text{Weight} \times \text{Positioning} + 0.0314 \times \text{Attacking Work Rate (Low)} \times \\
&\text{Finishing} + 0.00698 \times \text{Attacking Work Rate (Medium)} \times \text{Finishing} + 0.0646 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Heading Accuracy} - 0.0044 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Heading Accuracy} - 0.0033 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Curve} - 0.0003 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Curve} - 0.0365 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Sprint Speed} - 0.0437 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Sprint Speed} - 0.0313 \times \\
&\text{Attacking Work Rate (Low)} \times \text{Reactions} + 0.0272 \times \\
&\text{Attacking Work Rate (Medium)} \times \text{Reactions} + 0.0208 \times \\
&\text{Defensive Work Rate (Low)} \times \text{Agility} - 0.00297 \times \\
&\text{Defensive Work Rate (Medium)} \times \text{Agility} - 0.0546 \times \\
&\text{Defensive Work Rate (Low)} \times \text{Interceptions} - 0.0376 \times \\
&\text{Defensive Work Rate (Medium)} \times \text{Interceptions} - 0.0335 \times \\
&\text{Defensive Work Rate (Low)} \times \text{Positioning} - 0.00302 \times \\
&\text{Defensive Work Rate (Medium)} \times \text{Positioning} - 0.0015 \times \text{Crossing} \times \text{Finishing} - \\
&0.0002 \times \text{Crossing} \times \text{Heading Accuracy} + 0.0029 \times \text{Crossing} \times \text{Sprint Speed} + \\
&0.00013 \times \text{Crossing} \times \text{Shot Power} - 0.00175 \times \text{Crossing} \times \text{Jumping} + 0.0027 \times \\
&\text{Crossing} \times \text{Vision} - 0.0010 \times \text{Finishing} \times \text{Heading Accuracy} + 0.0033 \times \\
&\text{Finishing} \times \text{Short Passing} - 0.005 \times \text{Finishing} \times \text{Interceptions} + 0.0018 \times \\
&\text{Finishing} \times \text{Positioning} + 0.0003 \times \text{Heading Accuracy} \times \text{Short Passing} + 0.0013 \times \\
&\text{Heading Accuracy} \times \text{Long Passing} - 0.0011 \times \text{Heading Accuracy} \times \text{Ball Control} - \\
&0.0012 \times \text{Heading Accuracy} \times \text{Acceleration} + 0.0018 \times \text{Heading Accuracy} \times \\
&\text{Sprint Speed} - 0.0012 \times \text{Heading Accuracy} \times \text{Agility} + 0.0018 \times \\
&\text{Heading Accuracy} \times \text{Reactions} + 0.0025 \times \text{Heading Accuracy} \times \text{Shot Power} + \\
&0.0011 \times \text{Heading Accuracy} \times \text{Jumping} + 0.0039 \times \text{Heading Accuracy} \times \\
&\text{Strength} - 0.0018 \times \text{Heading Accuracy} \times \text{Long Shots}
\end{aligned}
$$

R2adj and RMSE of Best Fitted Model:

R2adj = 0.897, this value indicates that 89.7 percent of the variation of the response variable overall rating is explained by the final model containing the height, weight, defensive work ratee, heading accuracy, short passing, curve, long passing, ball control, acceleration, sprint speed, agility, reactions, jumping, strength, vision, age, interceptions, positioning, shot power, as well as the interactions between defensive work rate and agility, interceptions, positioning, and heading accuracy's interactions with long passing, agility, reactions, shot power, jumping, and strength.

RSE = 2.011, this value indicates that the standard deviation of the unexplained variation in estimation of response variable overall rating is 2.011.

Interpreting Coefficients:

Referring to our final model:

- **Height (0.2978)**: For each unit increase in height, the overall rating increases by 0.2978 points, holding all other variables constant.
- **Weight (0.0742)**: For each unit increase in weight, the overall rating increases by 0.0742 points, holding all other variables constant.
- **Attacking Work Rate (Low) (21**.47): Players with a low attacking work rate have an overall rating that is 21.47 points higher, compared with players with a high attacking work rate, holding all other variables constant.
- **Attacking Work Rate (Medium) (1.608**): Players with a low attacking work rate have an overall rating that is 1.608 points higher, compared with players with a high attacking work rate, holding all other variables constant.
- **Defensive Work Rate (Low) (3.624)**: Players with a low defensive work rate have an overall rating that is 3.624 points higher, compared to players with a high defensive work rate, holding all other variables constant.
- **Defensive Work Rate (Medium) (2.585)**: Players with a medium defensive work rate have an overall rating that is 2.585 points higher, compared to players with a high defensive work rate, holding all other variables constant.
- **Crossing (0.655)**: For each unit increase in crossing, the overall rating increases by 0.655 points, holding all other variables constant.
- **Finishing (-0.133)**: For each unit increase in finishing, the overall rating decreases by 0.133 points, holding all other variables constant.
- **Heading Accuracy (-0.368)**: For each unit increase in heading accuracy, the overall rating decreases by 0.368 points, holding all other variables constant.
- **Short Passing (-0.0836)**: For each unit increase in short passing ability, the overall rating decreases by 0.0836 points, holding all other variables constant.
- **Curve (0.0079)**: For each unit increase in curve ability, the overall rating increases by 0.0079 points, holding all other variables constant.
- **Long Passing (-0.0713)**: For each unit increase in long passing ability, the overall rating decreases by 0.0713 points, holding all other variables constant.
- **Ball Control (0.241)**: For each unit increase in ball control ability, the overall rating increases by 0.241 points, holding all other variables constant.
- **Acceleration (0.0932)**: For each unit increase in acceleration, the overall rating increases by 0.0932 points, holding all other variables constant.

- **Sprint Speed (-0.1897)**: For each unit increase in sprint speed, the overall rating decreases by 0.1897 points, holding all other variables constant.
- **Agility (0.0787)**: For each unit increase in agility, the overall rating increases by 0.0787 points, holding all other variables constant.
- **Reactions (0.0529)**: For each unit increase in reactions, the overall rating increases by 0.0529 points, holding all other variables constant.
- **Shot Power (0.3998)**: For each unit increase in shot power, the overall rating increases by 0.3998 points, holding all other variables constant.
- **Jumping (0.0513)**: For each unit increase in jumping ability, the overall rating increases by 0.0513 points, holding all other variables constant.
- **Strength (-0.1793)**: For each unit increase in strength, the overall rating decreases by 0.1793 points, holding all other variables constant.
- **Long Shots (0.0657)**: For each unit increase in long shots, the overall rating increases by 0.0657 points, holding all other variables constant.
- **Interceptions (0.0262)**: For each unit increase in interceptions, the overall rating increases by 0.0262 points, holding all other variables constant.
- **Positioning (0.2473)**: For each unit increase in positioning, the overall rating increases by 0.2473 points, holding all other variables constant.
- **Vision (-0.1488)**: For each unit increase in vision, the overall rating decreases by 0.1488 points, holding all other variables constant.
- **Age (0.00039)**: For each year increase in age, the overall rating increases by 0.00039 points, holding all other variables constant.
- **Height (-0.1188)**: For each unit increase in height, the overall rating decreases by 0.1188 points, holding all other variables constant.

Once the interaction terms are accounted for, many of these variables are likely to switch signs and thus must be taken together to determine real effects on overall rating.

Conclusion

To summarize our findings from the analysis, we first started with a first order linear model, we removed collinearity within the model, and implemented individual t-test to achieved best linear model. Then we added interaction to the model, we also checked for higher power but that was to no avail. We did individual t-test to our interaction model and finally, we have our adjusted interaction model.

Afterwards, we decided to run tests to pass assumptions. We passed the linearity and independence assumptions but failed the normality and homoscedasticity assumptions despite transforming our data via the box-cox and logarithmic method.

Discussion:

The conducted analysis yielded several key findings regarding the relationship between various player attributes and overall rating in soccer. The regression model revealed significant coefficients for attributes such as short passing, ball control, reactions, and defensive work rate, among others. These results provide valuable insights into the factors influencing overall player performance and rating.

The success of the modeling exercise in addressing the research questions and objectives can be evaluated positively. The model effectively captures the complex relationship between player attributes and overall rating, providing a quantitative understanding of the factors contributing to player performance. However, it is essential to acknowledge the limitations of the model, such as potential multicollinearity issues and the need for further validation using additional data sets.

To enhance the achievement of the research objectives in future studies, several suggestions can be considered. Firstly, exploring alternative statistical approaches, such as machine learning algorithms, could provide additional insights and improve the predictive accuracy of the models. Additionally, incorporating more comprehensive datasets that include a wider range of player attributes and match statistics may offer a more nuanced understanding of player performance.

In conclusion, the findings of this study underscore the importance of various player attributes in determining overall rating in soccer. By leveraging statistical modeling techniques, we have gained valuable insights into the factors influencing player performance. Moving forward, further research efforts should aim to refine and expand upon these findings to enhance our understanding of player evaluation in the sport.

Citations:

1. Kaggle Dataset:

- Hugo Mathien. "European Soccer Database." Kaggle, 2016, https://www.kaggle.com/hugomathien/soccer

2. Football-Data.co.uk:

- "Football Data." Football-Data.co.uk, https://www.football-data.co.uk/

3. Sofifa:

- "Sofifa." Sofifa.com, https://sofifa.com/

4. Premier League:

- "Premier League Stats." Premierleague.com, https://www.premierleague.com/stats/

5. Football Data API:

- "Football Data API." Football-data.mx-api.enetscores.com, http://football-data.mx-api.enetscores.com/