Examining New Age Analytics of European soccer

STAT 420: Methods of Applied Statistics

Spring 2021

Group 51

Jefferson Mathews, jrm10

David Lin, yiyangl7

Vidushi Somani, vsomani3

Benny Zhao, bzhao22

Contents

Introduction

European soccer is increasingly popular in the recent few years, which inspired us to use it as the topic for this statistical project. This project is focused on the analysis of statistical data about European soccer, specifically, the statistics on four European Leagues, EPL, La Liga, Serie A, and Ligue 1, during the 17-18, 18-19, and 19-20 season. We are mainly interested in predicting success with various predictors, including Expected Goals (xG), Expected Goals Allowed (xGA), Possession (Poss), Team Assignment, and Average Age (Age). We use the proportion of wins as our response variable, which represents how successful a team is objectively and accurately. To find the best statistical model, we utilize a series of methods, including collinearity analysis, simple relationship analysis, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), residual diagnostics, and outlier diagnostics. Through the exploration, the optimal model is found.

Methods

Response variable & predictor variable selection

First, we must pick a metric that best represents the success of a team for any given season; the metric we chose would be the proportion of games won by a team out of all the games played by the team in the season.

Next, we picked a selection of predictor variables from the data set that we think would have prediction power for predicting the proporition of the games won by a team. We decided on the following predictor variables to consider:

- Expected goals (`xG`), which measures the quality of the shots made by the team throughout the season, taking into account factors such as assist type, shot angle, distance from the goal, whether it was a headed shot, and whether it was defined as a big chance.
  - Essentially, it is the offensive power of a team, measured as a probability out of 100 of scoring, so a higher value means a more offensively capable team.
- Expected goals allowed (`xGA`), which measures a team's ability to prevent scoring chances
  - Essentially, it is the defensive power of a team, measured as a probability out of 100 for failing to prevent the opponent from scoring, so a lower value is a more defensively capable team.
- Amount of possession (`Poss`), which measures the proportion of passes attempted
- The average player ages of a team (`Age`)
- Team assignment (`Notes`), which indicates how a team was placed onto the league

Collinearity Analysis

First, we chose to examine whether there was collinearity witihin the full model that included all of the aforementioned predictors by looking at the values of the variance inflation factors of each of the predictors.

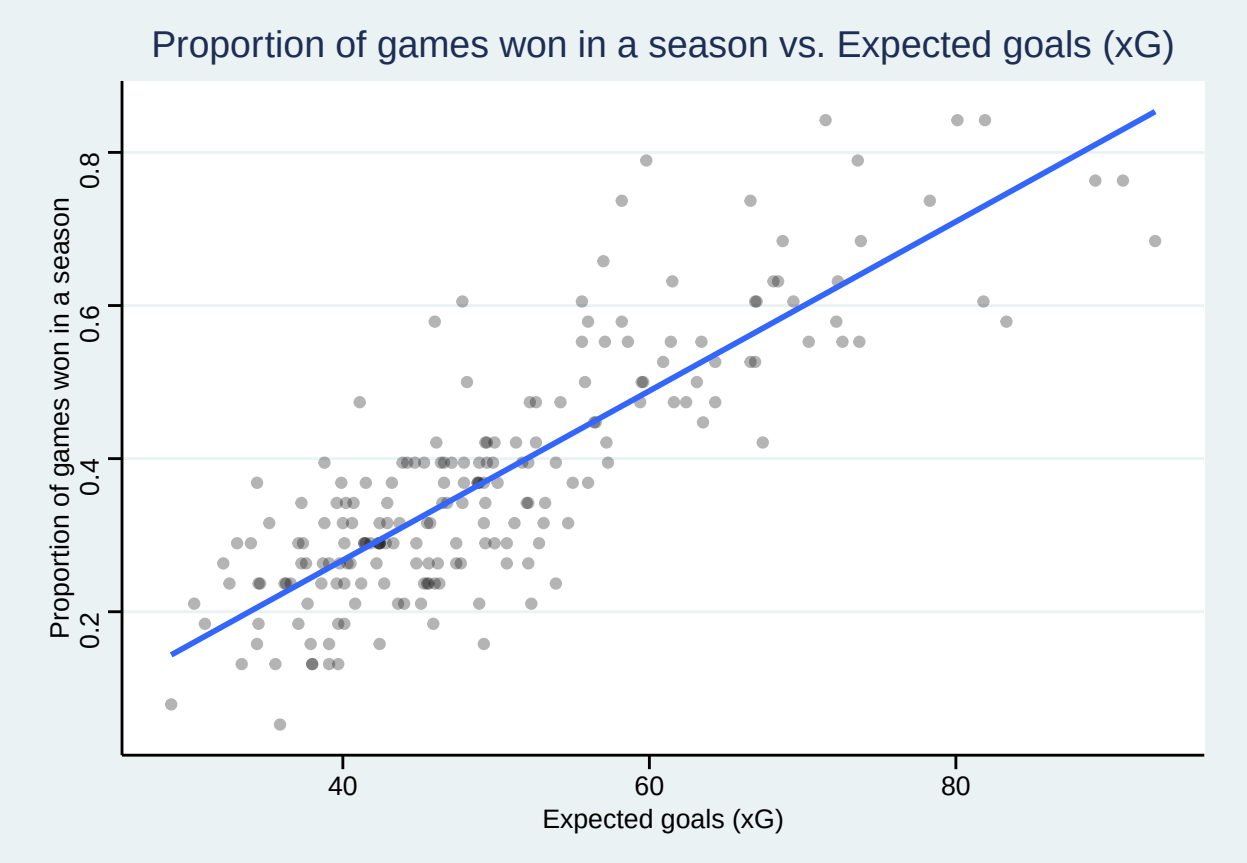
Predictor variable	VIF value
xG	3.205039
xGA	1.966579
Poss	3.030101
Age	1.030270
Team assignment: Neither	2.068080
Team assignment: Relegation Spot	2.266972

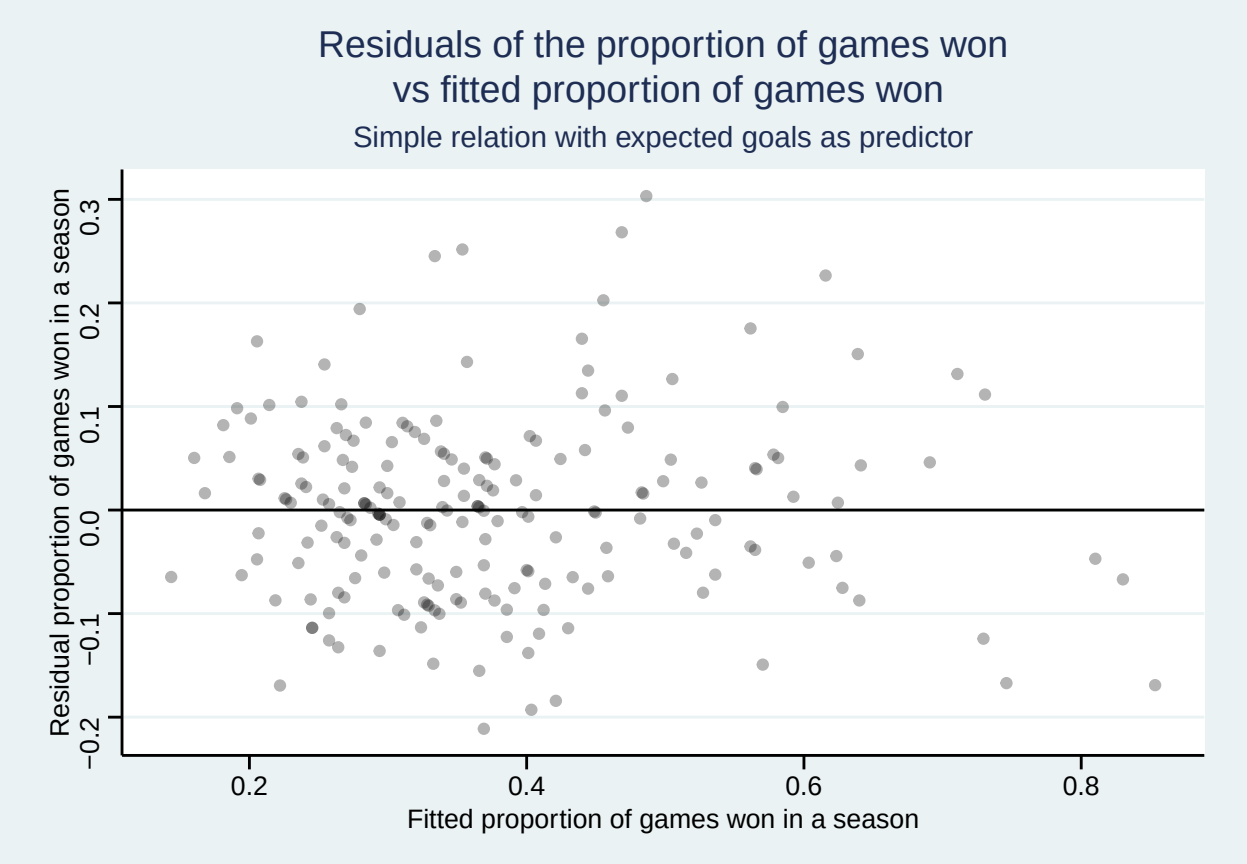
Since none of the VIF values of the predictors were greater than 5, we concluded that there was no need to exclude any variables from our final model.

Simple relationship analysis

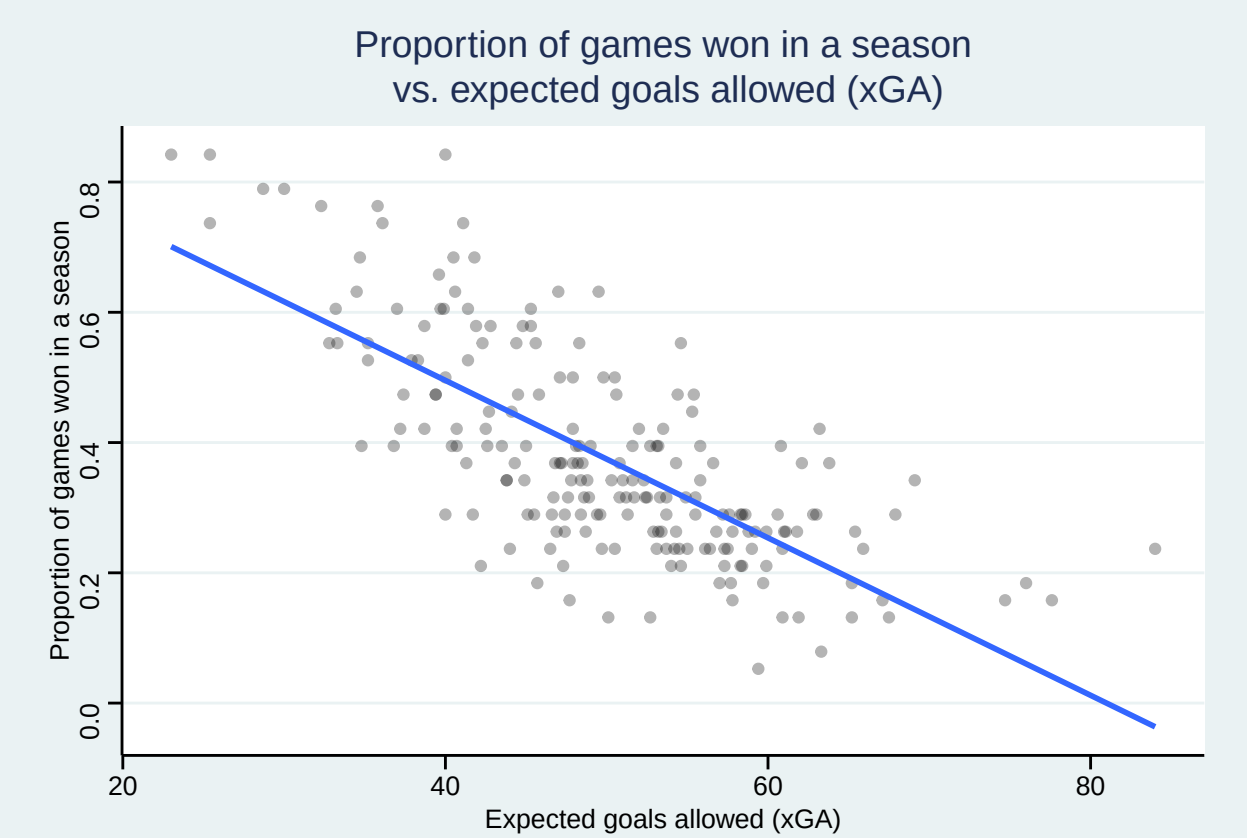
Next, we decided to look for any evidence of a relationship, linear or otherwise, between the proportion of games won by a team in a season and each of those predictor variables, individually.

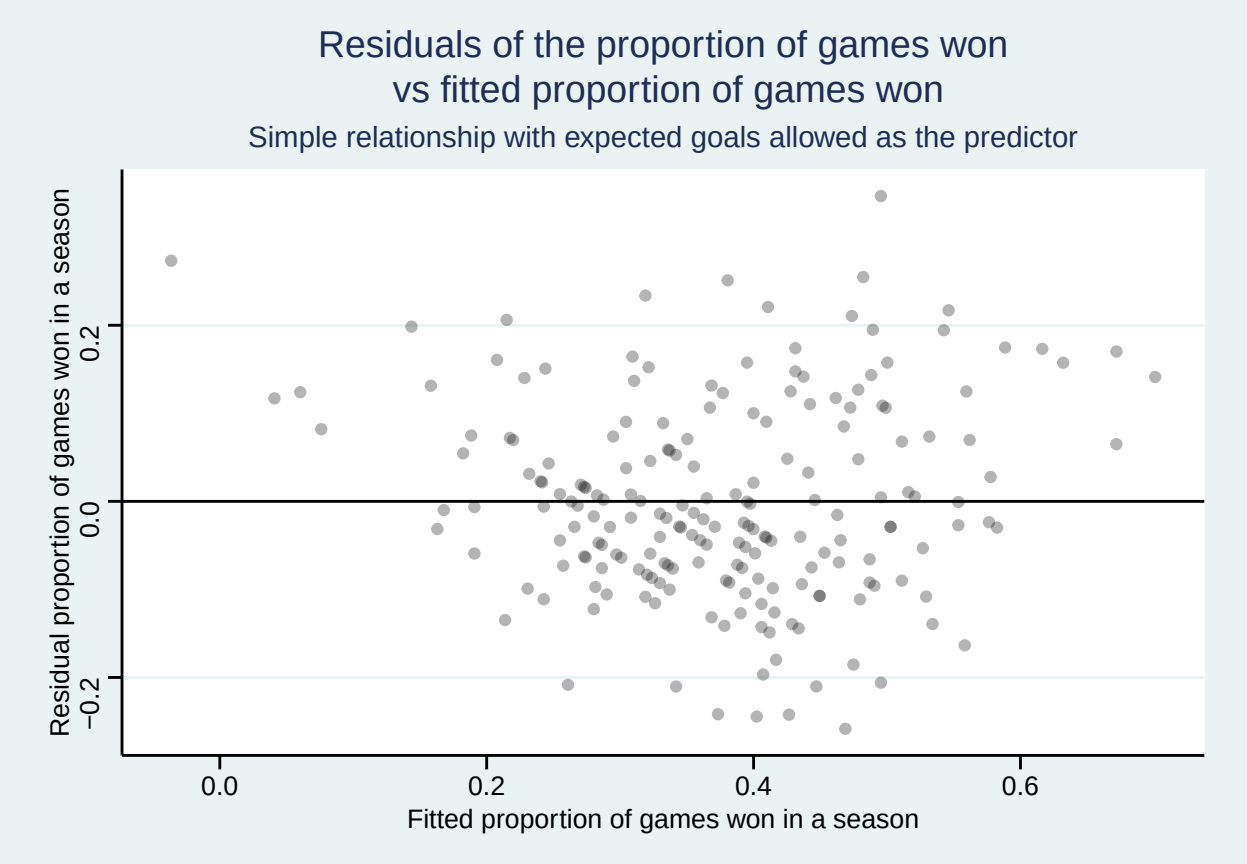
The first simple relationship we examined was the proportion of games won by a team in a season and the expected goals.



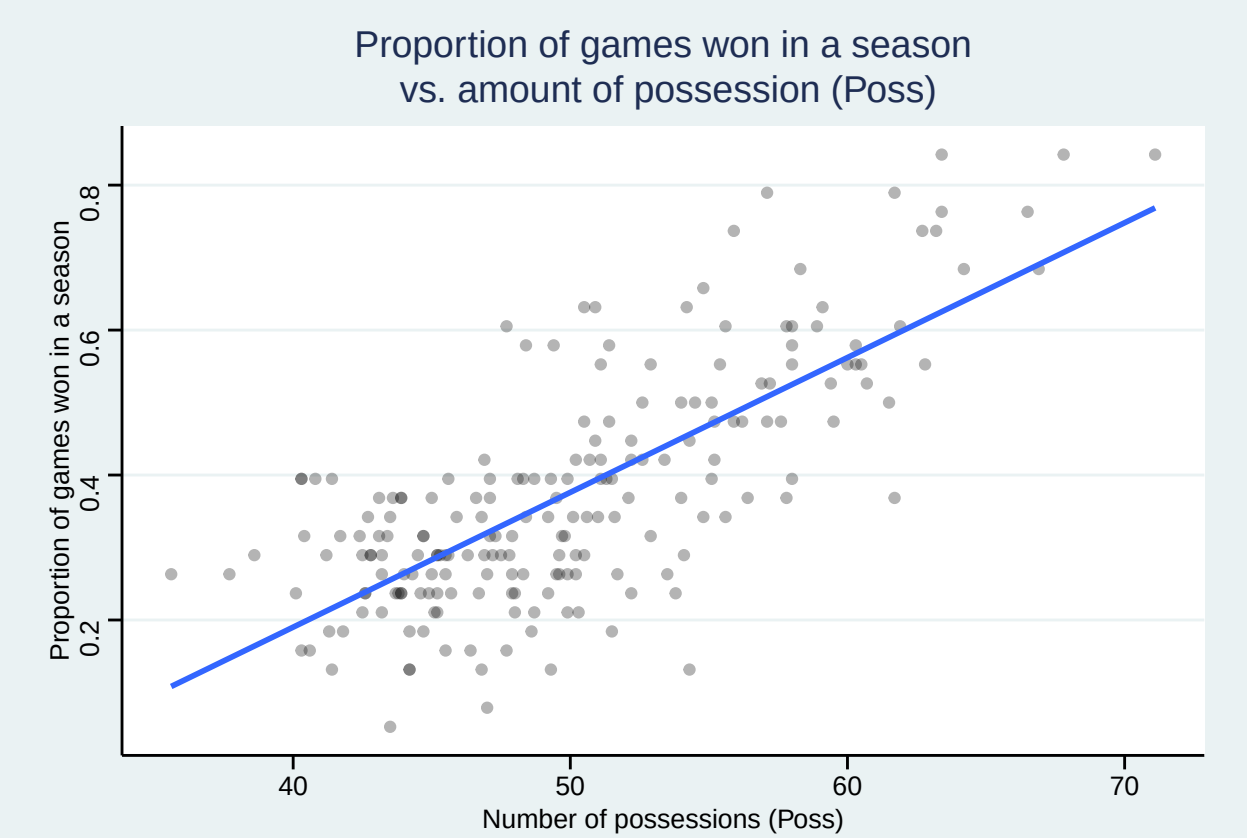


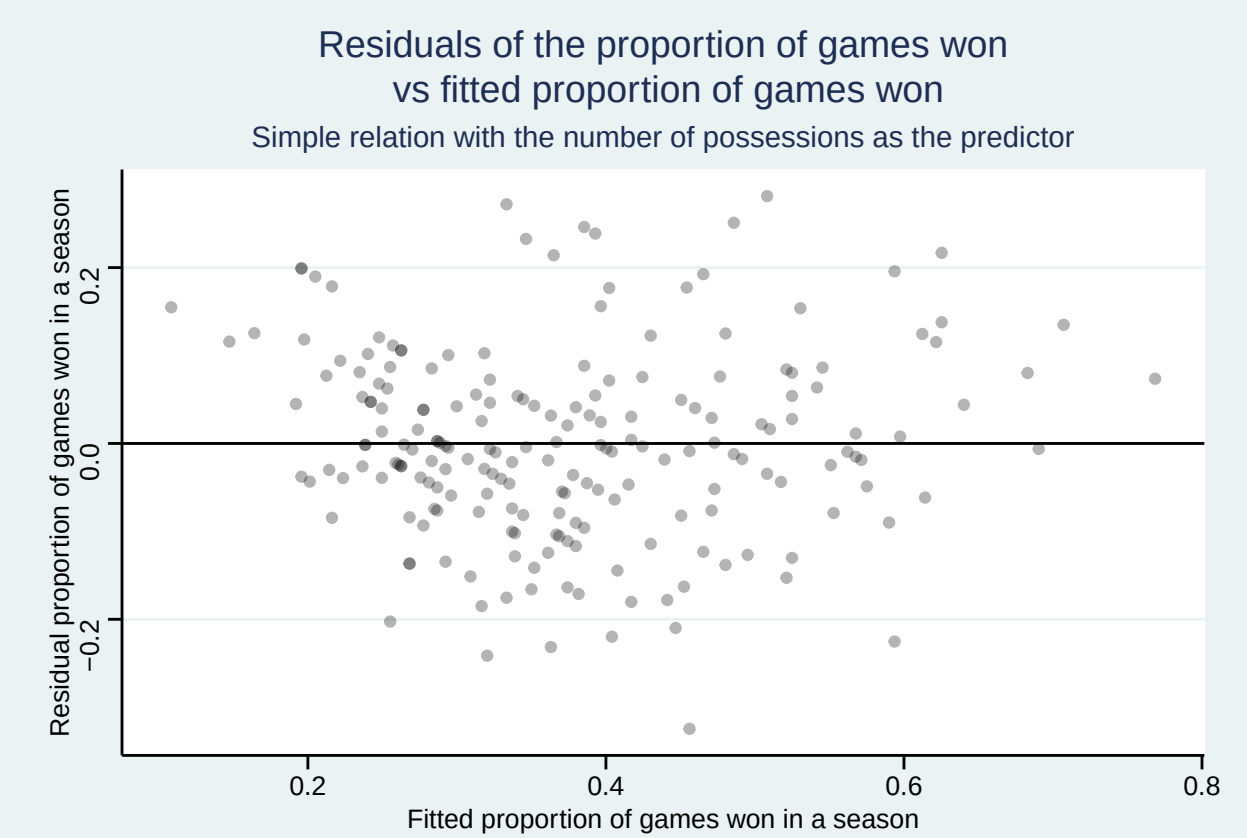
Next, we performed residual analysis by plotting the residual plot of the model between the proportion of games won and the expected goals against.


Proportion of games won in a season
vs. expected goals allowed (xGA)
Proportion of games won in a season
0.0
0.2
0.4
0.6
0.8
20
40
60
80
Expected goals allowed (xGA)



The next variable to consider in a simple relationship with the proportion of games won in a season was the number of posessions.


Proportion of games won in a season
vs. amount of possession (Poss)
Proportion of games won in a season
0.2
0.4
0.6
0.8
40
50
60
70
Number of possessions (Poss)



The observations we made on each of the three previous simple relationships were similar, in that none of the relationships look curved, so a polynomial transformation will not be necessary. However, there are some issues regarding the distribution of the residuals for the predicted proportion of games won in a season, where the variance is non-constant and not-normally distributed about the line of best fit in the proportion of games won in a different season for different values of the expected goals, expected goals allowed, and the number of possessions; therefore, a log transformation would be the way to remedy this.

We summarize the issues with the distribution of the residuals in the following table, which displays the p-values that were produced from running the Shapiro-Wilk and the Breusch-Pagan tests on the distributions of the residuals from the three simple models.

Predictor variable	Shapiro-Wilk test p-value	Breusch-Pagan test p-value
Expected goals	0.0127497	0.0217262
Expected goals allowed	0.0496668	0.1040044
Amount of possession	0.5846183	0.2331615

However, we believed it was more important for our model to be interpretable, and since the flaws of the distribution of the variance of the residuals were not too severe, we decided not to transform those three variables.

Additionally, we looked at the individual p-values of the t-tests that were run on the coefficient of the predictor, with the null hypothesis that there was no linear relationship between the predictor (expected

goals, expected goals against, number of possessions) and the proportion of games won in a season, which is summarized in the following:

Predictor variable	t-test p-value
Expected goals	0
Expected goals allowed	0
Amount of possession	0

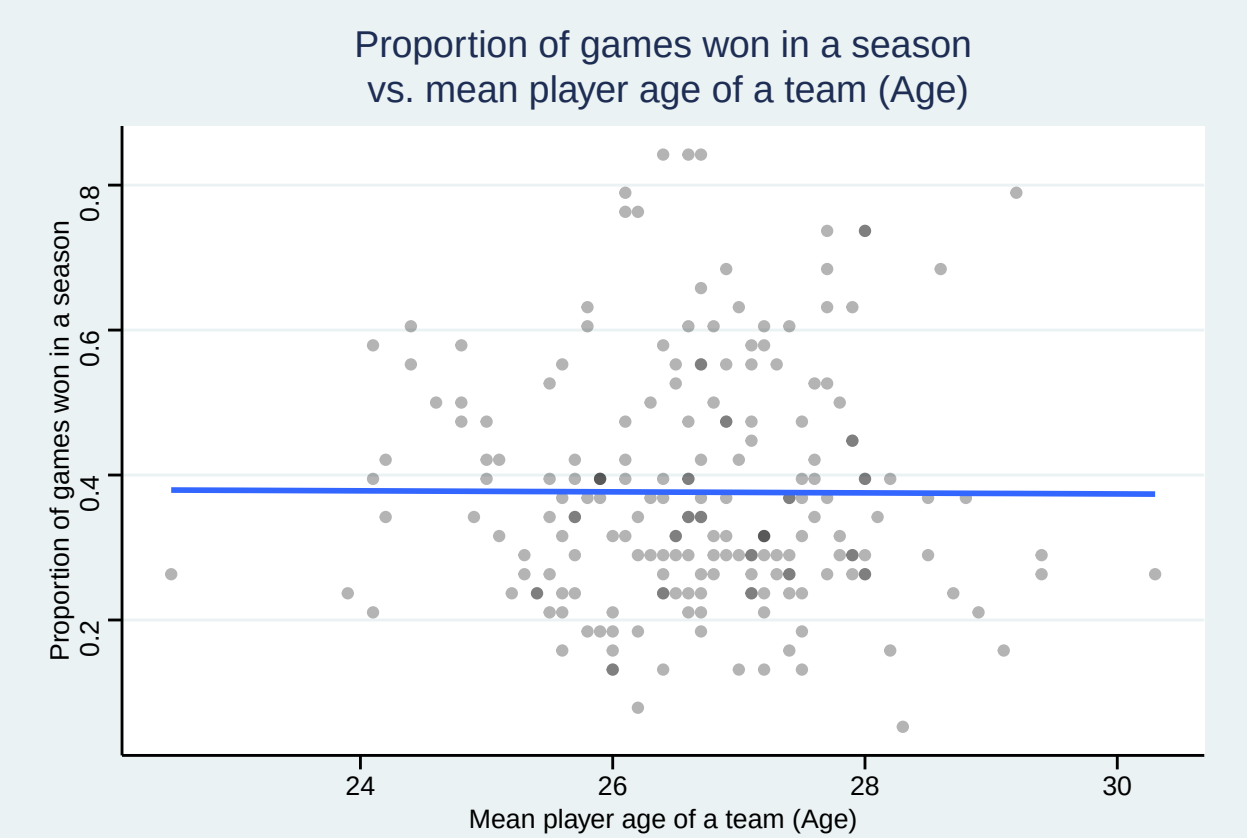
Since the p-values were very low, we concluded that were was strong evidence that indicates the statement of there not being a relationship between the proportion of games won, and the expected goals, expected goals against, number of possessions, individually, is false. Hence, these predictor variables are viable to include in our final model.

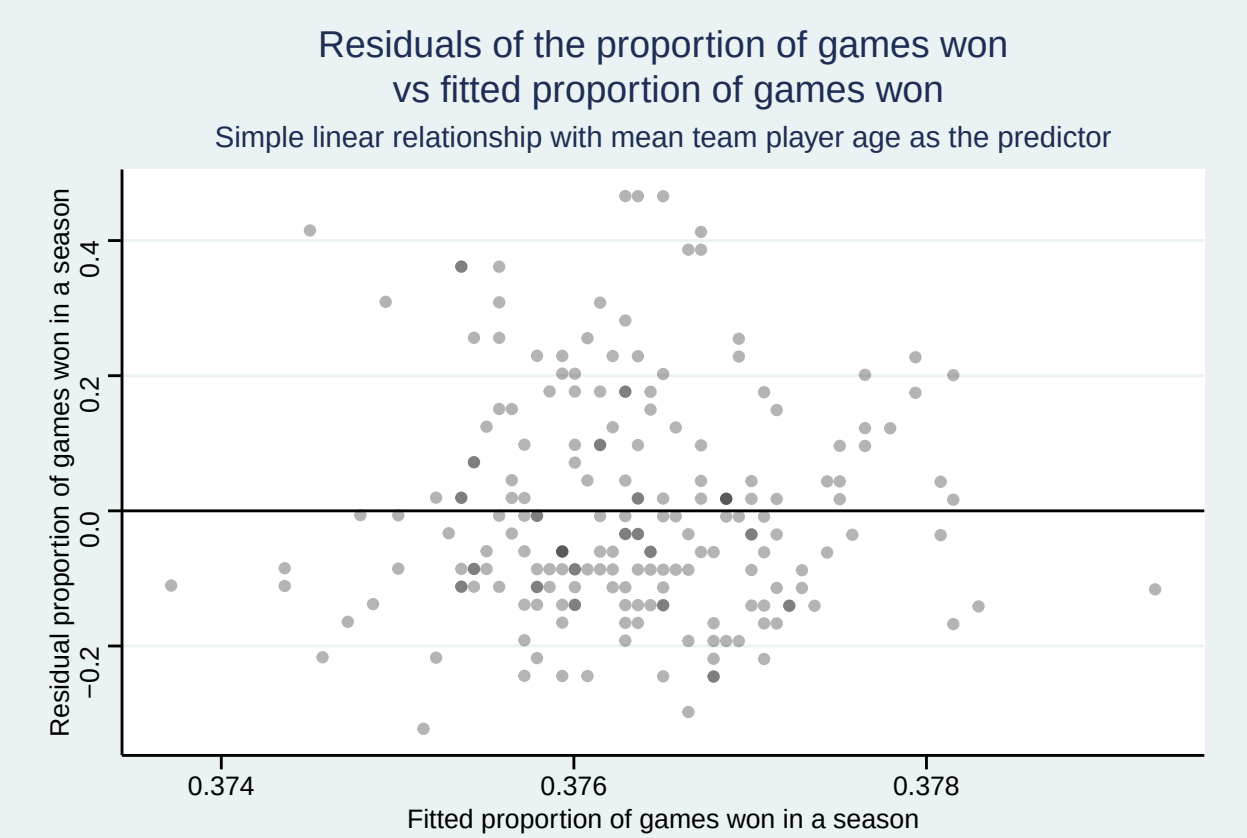
We also created prediction intervals to extrapolate values for certain response variables.

```
##          fit        lwr       upr
## 1 0.04632222 -0.1902233 0.2828677
## 2 0.93068265  0.6881729 1.1731924
```

```
##           fit        lwr       upr
## 1 0.004120338 -0.2863685 0.2946092
## 2 0.934074243  0.6357252 1.2324233
```

The next variable to consider in a simple relationship with the proportion of games won in a season was the age of the players on a team.


Proportion of games won in a season
vs. mean player age of a team (Age)
Proportion of games won in a season
0.2
0.4
0.6
0.8
24
26
28
30
Mean player age of a team (Age)



Given the small value of the coefficient for the mean player age of a team in the simple model, we suspected that there would be, at best, a very weak relationship between the proportion of games won and the mean player age of a team. We viewed the p-value of the t-test for the coefficient of the predictor, and given the large p-value of 0.943597, we concluded that using the mean player age as an additive predictor would not add any predictive power into our final model, but we may consider it for an interaction term.

In order to quantify how much predictive power would increase by adding the mean player age as a predictor, we decided to add the mean player age as predictor to the model that had the highest R^2 value out of the simple models that had expected goals, expected goals allowed, and number of possessions, which are summarized in the following:

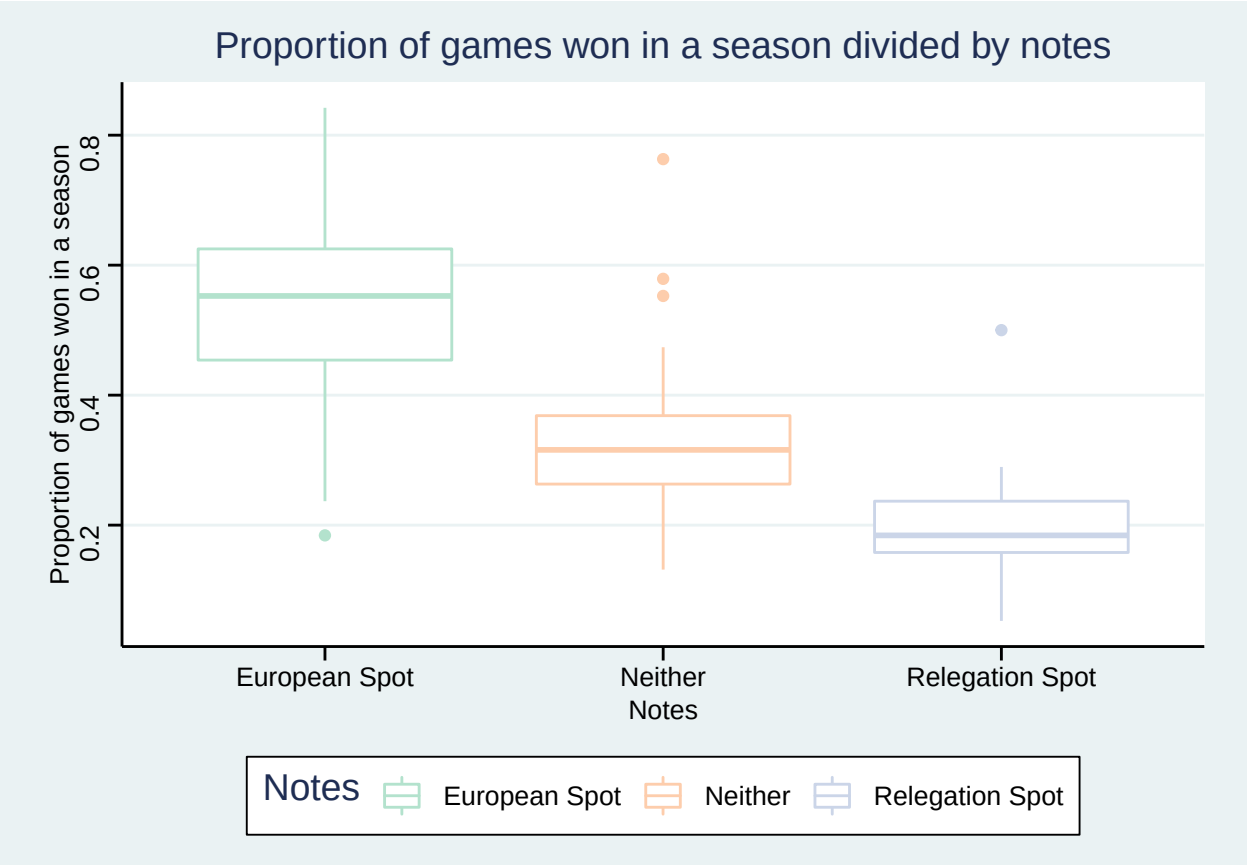
Predictor variable	R^2
Expected goals	0.7034973
Expected goals allowed	0.5239362
Amount of possession	0.5602088

We ran an F-test to determine the additional predictive power by adding the mean player age as a predictor to the expected goals model, which returned the p-value of 0.19402. Given the magnitude of this p-value, we had more evidence to not include the mean player age as an additive predictor into our model.

The final variable to consider in a relationship with the proportion of games won in a season was the `Notes` value of the team, which was the only categorical variable out of the predictors we chose to potentially

include in our final model.

By plotting the proportion of games won and dividing the teams by `Notes`, we can see there is a visible difference in the distribution of the proportion of games won.



In order to quantify how much adding `Notes` as a predictor would be, we decided to add `Notes` as predictor to the model that had the highest R^2 value out of the simple models, which was the expected goals model.

After running an F-test to compare the simple model using only the expected games as a predictor and using expected games and the `Notes` as predictors, the p-value returned was $4.7635797 \times 10^{-16}$. Given the small p-value, we concluded that adding the `Notes` as a categorical dummy variable would greatly improve the predictive power of our model.

We decided on one of the candidate models using expected goals (xG), expected goals allowed (xGA), number of possessions (Poss), and the `Notes` as predictors for the proportion of games won in a season ($\mathrm{W}_{\mathrm{prop}}$), which can be written as the following:

$$y_{\mathrm{W}_{\mathrm{prop}}} = 0.2120 + 0.006140x_{\mathrm{xG}} - 0.003698x_{\mathrm{xGA}} + 0.002031x_{\mathrm{Poss}} - 0.07553x_{\mathrm{Neither}} - 0.1309x_{\mathrm{Relegation\ Spot}},$$

where $x_{\mathrm{Neither}} = 1$ if the `Notes` of the team is neither or equal to 0 otherwise, and $x_{\mathrm{Relegation\ Spot}} = 1$ if the `Notes` of the team is Relegation spot or equal to 0 otherwise.

Interaction terms

We also chose to examin two more models produced from backwards variable selection with both AIC and BIC as the criterion and using two-way interaction terms between the expected goals (xG), expected goals allowed (xGA), number of possessions (Poss), average player age in a team (Age), and `Notes`. We began with the full model using the following variables:

- Expected goals
- Expected goals allowed
- Amount of possession
- `Notes`
- All possible two-way interactions between the above four variables along with the average player age

The model produced by backwards variable selection using AIC is the following:

$$\begin{gathered} y_{\mathrm{W_{prop}}} = -1.4737 + 0.005644x_{\mathrm{xG}} + 0.02699x_{\mathrm{xGA}} + 0.01318x_{\mathrm{Poss}} - 0.07311x_{\mathrm{Neither}} + \\ -0.1448x_{\mathrm{Relegation\ Spot}} + 0.04263x_{\mathrm{Age}} - 0.0002303x_{\mathrm{xGA}}x_{\mathrm{Poss}} - 0.0007121x_{\mathrm{xGA}}x_{\mathrm{Age}} \end{gathered},$$

where $x_{\mathrm{Neither}} = 1$ if the `Notes` of the team is neither or equal to 0 otherwise, and $x_{\mathrm{Relegation\ Spot}} = 1$ if the `Notes` of the team is Relegation spot or equal to 0 otherwise.

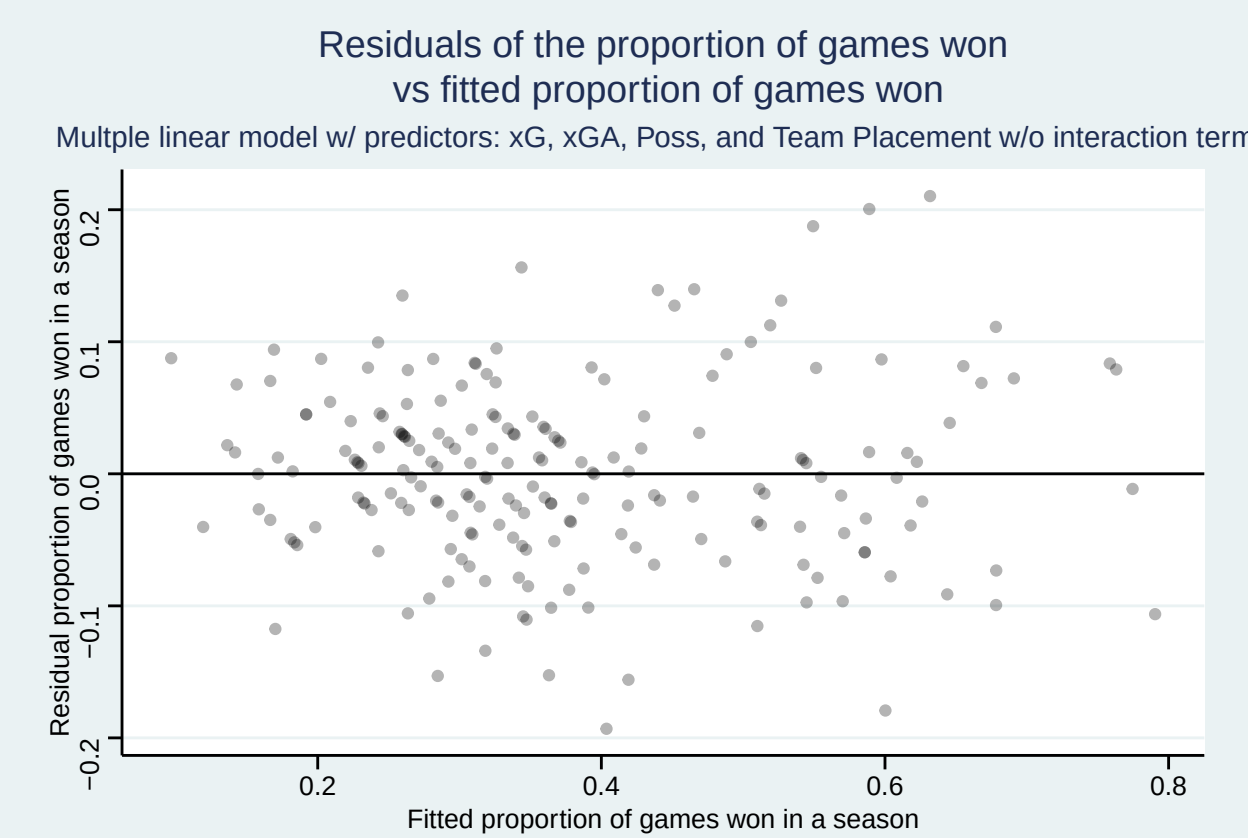
The model produced by backwards variable section using BIC is the following:

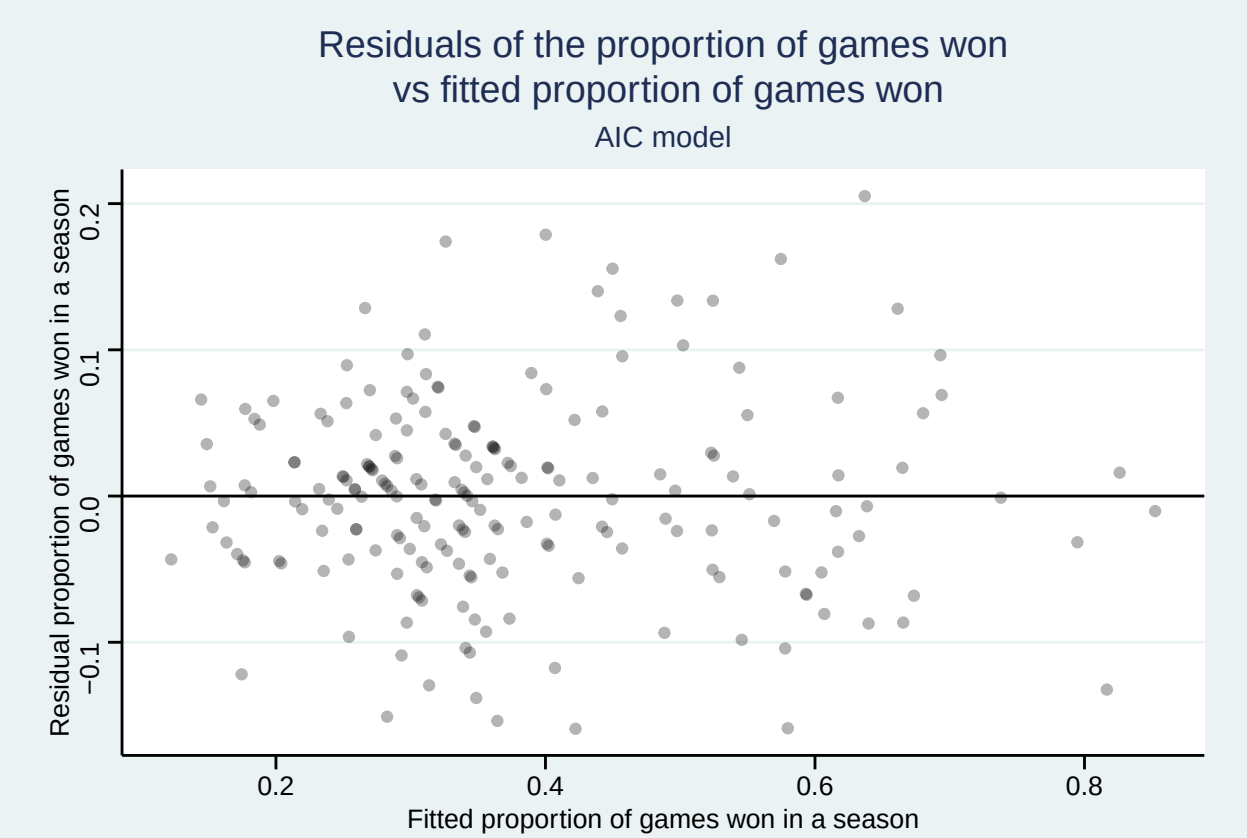
$$\begin{gathered} y_{\mathrm{W_{prop}}} = -0.3313 + 0.005501x_{\mathrm{xG}} + 0.008182x_{\mathrm{xGA}} + 0.01336x_{\mathrm{Poss}} - 0.07613x_{\mathrm{Neither}} + \\ -0.1466x_{\mathrm{Relegation\ Spot}} + 0.04263x_{\mathrm{Age}} - 0.0002366991x_{\mathrm{xGA}}x_{\mathrm{Poss}} \end{gathered},$$

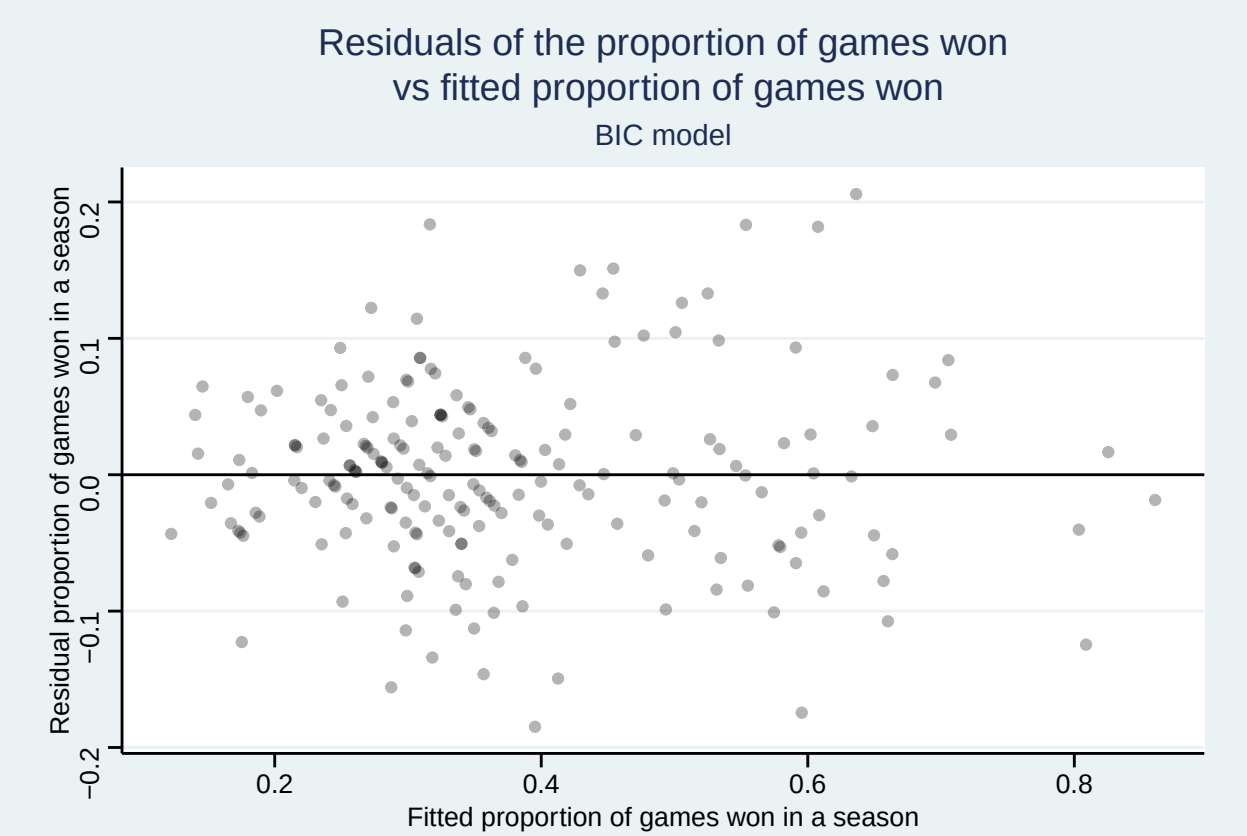
where $x_{\mathrm{Neither}} = 1$ if the `Notes` of the team is neither or equal to 0 otherwise, and $x_{\mathrm{Relegation\ Spot}} = 1$ if the `Notes` of the team is Relegation spot or equal to 0 otherwise.

Residual diagnostics

Next, we examined the distribution of residuals from the three aforementioned models.


Residuals of the proportion of games won
vs fitted proportion of games won
Multple linear model w/ predictors: xG, xGA, Poss, and Team Placement w/o interaction term
Residual proportion of games won in a season
−0.2
−0.1
0.0
0.1
0.2
0.2
0.4
0.6
0.8
Fitted proportion of games won in a season


Residuals of the proportion of games won
vs fitted proportion of games won
AIC model
Residual proportion of games won in a season
0.2
0.1
0.0
−0.1
0.2
0.4
0.6
0.8
Fitted proportion of games won in a season



We observed that from all three distributions of residuals from the three models, the variance of the residuals tends to increase in the range of fitted proportion of games won in a season between 0.3 and 0.6. The first multiple linear model and the BIC model appears to have residuals that look normally distributed along the $y = 0$ line, but the AIC model's residuals tend to have positive skew.

To quantify how large of an issue the nonconstant residual variance and skewed residual distribution, we performed Shapiro-Wilk and Breush-Pagan tests on the residuals for all three models. The p-values from the tests performed is summarized in the following:

Model	Shapiro-Wilk test p-value	Breusch-Pagan test p-value
Multiple linear model (w/o interaction terms)	0.4227991	0.0033629
AIC model	0.0780570	0.0317978
BIC model	0.0721157	0.0064687

Outlier diagnostics

In order to determine which teams were considered outliers, we used the Cook's distance to determine the influence that each team had on the resulting model. We decided that teams that had a Cook's distance greater than $\frac{4}{n}$, where n is the number of teams in the data set.

From the multiple linear model without interaction terms, there were 15 outliers; from the AIC model, there

were 14 outliers. From the BIC model, there were 13 outliers.

From the prior residual diagnostics, we hope that by removing the outliers, the distribution of the residuals resembles more of a normal distribution about the line of best fit and that the variance of the residuals becomes more consistent.

After removing the respective outliers from the data set, (15 for the multiple linear model without interaction terms, 14 for the AIC model, and 13 for the BIC model), we compared the adjusted R^2 value before and after removing the outliers and summarized our findings:

Model	Adj. R-Squared from Model w/ High Inf.	Adj. R-Squared from Model w/o High Inf.
Multiple linear model (w/o interaction terms)	0.8188286	0.8554364
AIC model	0.8311125	0.8699546
BIC model	0.8280759	0.8659728

Given the increase in the value of the adjusted R^2 for all three models after removing the outliers for each individual model, we decided to replace the former three models that included the high influence teams in the data set.

We also decided to perform residual diagnostics once again on the new models that only accounts for low influence teams:

Model	Shapiro-Wilk test p-value	Breusch-Pagan test p-value
Multiple linear model (w/o interaction terms)	0.8038321	0.2043908
AIC model	0.9609045	0.1037089
BIC model	0.9825724	0.3566308

We observed that the p-values of the Shapiro-Wilk test and the Breusch-Pagan test performed on the distributions of the residuals are much larger than when they were performed on the residuals of the models that accounted for high influence teams. Hence, we decided to keep these three models as our candidate models.

Comparison of final candidate models.

We arrived with three candidate models to best predict the proportion of games won by a team in a season:

- Multiple linear model that has the following predictors: expected goals (xG), expected goals allowed (xGA), number of possessions (Poss), and `Notes`.
- Backwards varaible selection model using AIC as the criterion, with the same additive predictors as the above multiple linear model along with average team player age (Age) and two interaction terms:
  - The interaction between expected goals and number of possessions, and the interaction between the expected goals against and the number of possessions.

- Backwards variable selection model using BIC as the criterion, with the same additive predictor as the first multiple linear model above, along with a single interaction term:
  - The interaction between expected goals against and the number of possessions

In order to compare the models, we summarized the relevant details of all three models as a table:

Selection criteria	Multiple linear model	AIC model	BIC model
No. of parameters (inc. intercept)	6.0000000	8.0000000	7.0000000
Adj. R-squared	0.8554364	0.8699546	0.8659728
RMSE	0.0571525	0.0541925	0.0549465
LOOCV-RMSE	0.0590768	0.0565712	0.0568372
Residual Shapiro-Wilk test p-value	0.8038321	0.9609045	0.9825724
Residual Breusch-Pagan test p-value	0.2043908	0.1037089	0.3566308

Results

Most favorable model

When examining the table displaying the selection critera of the three candidate models, we decided on the BIC model to be the most effective model out of the three. The coefficients of the parameters, the standard error, along with the t-value and the p-value of the t-test performed on each individual predictor are summarized below:

Parameter	Coefficient	Standard error	t-value	p-value (two-sided)
Intercept	-0.3565199	0.1496288	-2.382696	0.0182282
Expected Goals (xG)	0.0054742	0.0006564	8.340061	0.0000000
Expected goals allowed (xGA)	0.0096279	0.0030237	3.184132	0.0017111
Amount of possession (Poss)	0.0133974	0.0030235	4.431088	0.0000163
Team Placement: Neither	-0.0835819	0.0130401	-6.409602	0.0000000
Team Placement: Relegation Spot	-0.1702663	0.0190380	-8.943484	0.0000000
Interaction between expected goal allowed & amt. of possession	-0.0002535	0.0000590	-4.299144	0.0000280

Predictor analysis

Since the intercept begins at a negative number, and the response variable is a proportion, any of the predictors that have positive coefficients are those that help a team in being more capable of winning games in the season. This is obvious with expected goals and number of possessions, since a higher expected goal value means that a team is stronger offensively and capable of scoring goals, and the greater the number of possessions means that a team has more opportunities to score a goal.

However, what was surprising was the value of the expected goals allowed, which measures the number of goals the team allowed the enemy to score. Intuitively, there should be negative relationship between the expected goals allowed and the proportion of games won, since a smaller expected goals allowed would mean that the team is stronger defensively and capable of preventing their opponent from scoring.

Furthermore, for the categorical dummy variable `Notes`, the model predicts that teams that have a `Notes` of neither or relegation spot instead of European spot will win a lesser proportion of games in a season on average.

As for the interaction term between the expected goals allowed and the number of possessions, we conjectured that a team that had many possessions would in theory perform well, however, if they had poor defense and keep allowing the opponent to steal the ball from, and allowing them to score, then the team would likely perform worse.

We can represent our final model using mathematical symbols as follows:

$$y_{\text{W}_{\text{Prop}}} = -0.3565 + 0.005474x_{\text{xG}} + 0.009628x_{\text{xGA}} + 0.01340x_{\text{Poss}}$$
$$-0.08358x_{\text{Neither}} - 0.1703x_{\text{Relegation Spot}} - 0.0002535x_{\text{xGA}}x_{\text{Poss}}$$

Where variables are defined as follows:

Variable	Meaning	Additional notes
$y_{\text{W}_{\text{Prop}}}$	Proportion of games won in a season	Measured out of 1
x_{xG}	Expected goals	Measured out of 1
x_{xGA}	Expected goals allowed	Measured out of 1
x_{Poss}	Amount of possession	
x_{Neither}	`Notes`	1 if `Notes` is Neither; otherwise, it is 0
$x_{\text{Relegation Spot}}$	`Notes`	1 if `Notes` is Relegation Spot; otherwise, it is 0

Reasoning behind the choice of final model

Compared to the other two candidate models that we ended up with, differences in the RMSE, LOOCV-RMSE, adjusted R^2 were marginal. Hence, we looked at the remaining criteria, which were the number of parameters and the p-values produced from running the Shapiro-Wilk and Bruesch-Pagan tests on the distribution of the residuals from each of the models. We concluded that the differences in the number of parameters were also marginal, with the least number of parameters being 6, and the most being 8.

However, the differences in the p-values from the Shapiro-Wilk and the Breusch-Pagan were quite substantial. None of the p-values produced were small enough to completely disqualify using one of the models to predict the proportion of games won in a season as the smallest p-value was the Breusch-Pagan test on the AIC model with a magnitude of 0.1037089. However, compared to the other two models, the p-values of those tests on the BIC model's residuals were the biggest, with the largest difference of 0.1787403 from the smallest p-value of the Shapiro-Wilk test on the multiple linear model that did not include interaction terms. Since we did not use any transformations of the variables, we picked the BIC model to ensure that any flaws in the distribution of the residuals is minimized and that we are able to use prediction intervals since prediction intervals require the distribution of the response (the proportion of games won in a season in this case) given a predictor to be normally distributed.

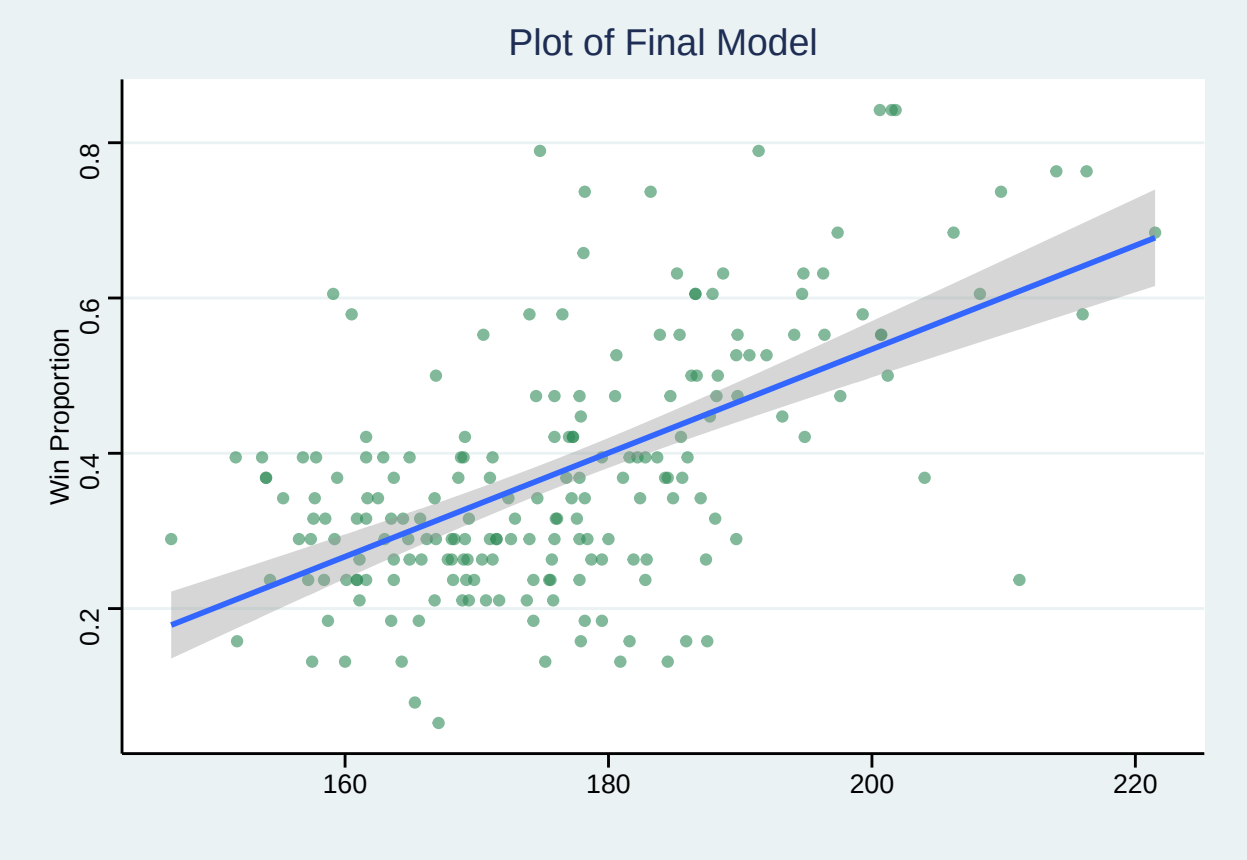
Given the context of the data set and what we hope to achieve with this study, being able to create prediction intervals for the success of team in a season would be very beneficial.

Graphical visualizations of our model

```
ggplot(data = european_soccer, mapping = aes(x = xG + xGA + Poss + Age, y = W_prop))+
  geom_point(col = "Seagreen", alpha = 0.6)+
  geom_smooth(method = "lm")+
```

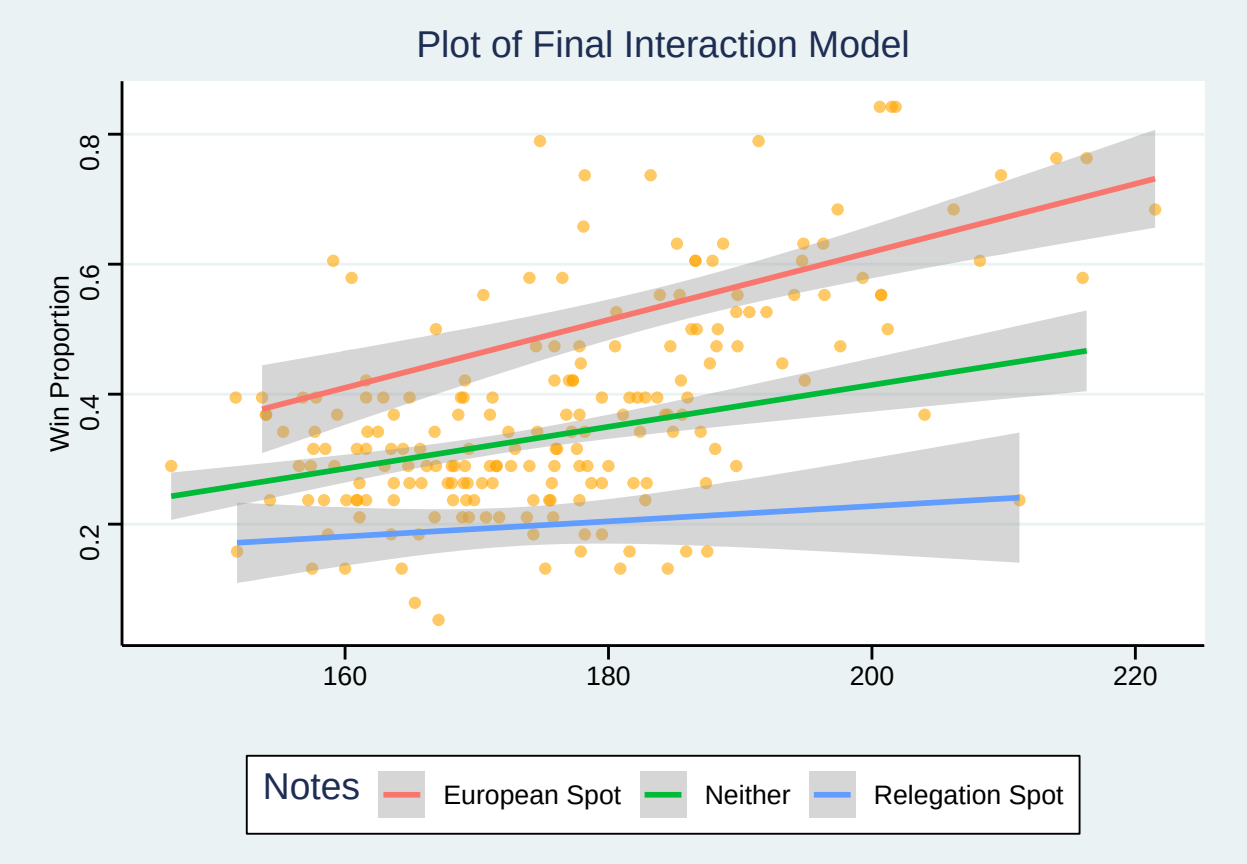
```
  theme_stata()+
  labs(title = "Plot of Final Model", x = "", y = "Win Proportion")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(data = european_soccer, mapping = aes(x = xG + xGA + Poss + Age, y = W_prop, color = Notes))+
  geom_point(col = "Orange", alpha = 0.6)+
  geom_smooth(method = "lm")+
  theme_stata()+
  labs(title = "Plot of Final Interaction Model", x = "", y = "Win Proportion")
```

```
## `geom_smooth()` using formula 'y ~ x'
```


Plot of Final Interaction Model
Win Proportion
0.2
0.4
0.6
0.8
160
180
200
220
Notes
European Spot
Neither
Relegation Spot

Discussion

Based on the coefficients of the model, we would predict that the single action that would increase the proportion of games won would be for a team to increase the number of possessions they obtain during a game; specifically, by we predict that the proportion of games won to increase by 0.0133974 on average for a single increase in the number of possessions in a game, assuming that all other factors are kept the same. Additionally, the single factor that decreases the proporition of games won would be the `Names` label assigned to team, where being a Relegation Spot instead of European spot decreases the proporition of games on average by 2.5352636×10^{-4}, assuming that all other factors like expected goals are kept the same.

Additionally, from during our method of finding candidate models, we concluded that the average player age has little impact on the proportion of games won in a season. We conjecture that this may be due to younger players having greater fluid intellience, where they are in their physical and cognitive peak, but they have less experience compared to the older players that have the greater crystallized intelligence that have the intuition of the game.

When assessing the viability of our model we concluded that after adjusting for correlation due to random change, approximately 86.5972832% of the variance in the proportion of games in a season is explained by our model.

Moving on, during our method of finding candidate models, we had droppped about 13 to 15 outliers depending on the candidate model. We are interested in which variables that these outliers deviated so far from the herd that resulted in them being considered high influence. For example, they could be potential powerhouse teams that dwarf the other teams.

One of the flaws would be how the teams were recorded multiple times in the data set as we aggregated data sets from different seasons, so the same teams would be multiple samples within our data set, but just be from different years. This brings up the query of conducting a study on aggregated data of a team, so that we would be able to predict the success of a team by comparing the change in the variables like expected goals from season to season as a way of determining what a team should focus their efforts on improving in order to maximize the proportion of the games won.

In general, we need to acknowledge that the data collection may not be entirely accurate and could be incomplete. It's usability is not quantifiable and there may be some bias in the logging of the data.

Continuing with this dataset, some areas that we would have liked to explore include creating a training and testing model to use the attendance of teams to predict wins, losses, and goals. There are also many different areas we could explore given this data and the results we were able to produce in this study.

Appendix

List of R librarys used

- ggplot2: visualize the data
- ggthemes: add extra geoms, scales, and themes for ggplot2
- gridExtra: help arranging plots on the page and drawing table
- lmtest: add functions for Shapiro-Wilk and Breusch-Pagan test
- faraway: provide the vif() function for calculating variance inflation factor
- knitr: enable knitting the rmd file as a pdf file, with latex and Markdown support
- readr: read and load the csv file

Code Snippet worth including

We used the following function when performing the residual diagnostics, with `plotit` set to `FALSE`, as we used `ggplot2` to display the residual plots.

```
diagnostics = function(model, pcol = "grey", lcol = "dodgerblue",
                       plotit = TRUE, testit = TRUE) {
  if (plotit == TRUE) {
    par(mfrow = c(1, 2))
    plot(x = fitted(model), y = resid(model), col = pcol,
         xlab = "Fitted", ylab = "Residuals", main = "Residual plot")
    abline(h = 0,  col = lcol)

    qqnorm(resid(model), col = pcol,
           main = c("Normal Q-Q Plot of Inputted Model"))
    qqline(resid(model), col = lcol)
  }
  if (testit == TRUE) {
    library(lmtest)
    p_sw <- shapiro.test(resid(model))$p.value
    p_bp <- bptest(model)$p.value[[1]]
    list(p_sw = p_sw, p_bp = p_bp)
  }
}
```

```
loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}
```