An Analysis of the Influence of Goalkeepers in Soccer

AT743 Final Project

Alan Zhou

2024-12-10

1. Introduction

In soccer, offenses and defenses have traditionally been assessed by the number of goals scored and goals allowed. While these two values fully predict the outcome of an individual game, they can be somewhat volatile over the course of a few games and do not lend themselves as well to coaching and personnel decisions.

Over the past few decades, advanced statistics have been devised to be more predictive of long-term success and to be more actionable in terms of training and scouting. The classic example of an advanced statistic in soccer is expected goals (xG), which assigns to each shot a probability of scoring a goal based on various positional factors independent of the individual players involved and then sums these probabilities.

For this project, we focus on goalkeeping and its impacts on three areas:

- How much does shot-stopping ability contribute to the team's overall point total?
- How much do non shot-stopping goalkeeper actions contribute to the team's offensive output?
- How much do non shot-stopping goalkeeper actions contribute to the team's defensive stability?

To do this, we look at the 98 teams in the 2023-24 season in the "Big 5" European leagues (England's Premier League, Spain's La Liga, Germany's Bundesliga, Italy's Serie A, and France's Ligue 1). The data for this project is sourced from the site FBref, which in turn sources its advanced statistics from Opta. We use `rvest` to scrape data from three tables:

- https://fbref.com/en/comps/Big5/2023-2024/stats/squads/2023-2024-Big-5-European-Leagues-Stats
- https://fbref.com/en/comps/Big5/2023-2024/keepers/squads/2023-2024-Big-5-European-Leagues-Stats
- https://fbref.com/en/comps/Big5/2023-2024/keepersadv/squads/2023-2024-Big-5-European-Leagues-Stats

Since the Bundesliga and Ligue 1 play fewer games than the other three leagues, the number of games played would be a lurking variable for any cumulative statistics. As such, we replace them with their "per 90" versions in the analysis that follows by dividing by matches played (MP).

2. Shot-Stopping and Point Total

The "post-shot expected goals" statistic (PSxG) measures, for each shot on target allowed (SoTA), the probability that the shot would go in given its trajectory. The difference between the PSxG and the actual number of goals allowed (GA) is often used as a measure of a goalkeeper's shot-stopping ability, with a higher $PSxG - GA$ being better. In order to normalize for volume and to account for some other technical details with PSxG, we define "shot-stopping" (ShotStop) by

$$ShotStop = 100 \cdot \frac{PSxG - (GA - pGA - OG)}{SoTA},$$

where pGA and OG are the number of penalty goals allowed and the number of own goals. This can be interpreted as the number of percentage points by which the goalkeeper's shot-stopping ability decreases the chances that a shot on target results in a goal.

In league play in soccer, teams are awarded 3 points for a win (W), 1 point for a draw (D), and 0 points for a loss (L). We start our analysis by building linear regression two models for points per game (PPG) over the course of the season, one where we only account for offense and defense and the other where we account for shot-stopping as well. As metrics for offensive and defensive capability, we use non-penalty expected goals per 90 (npxG90) and PSxG per 90 (PSxG90), respectively.

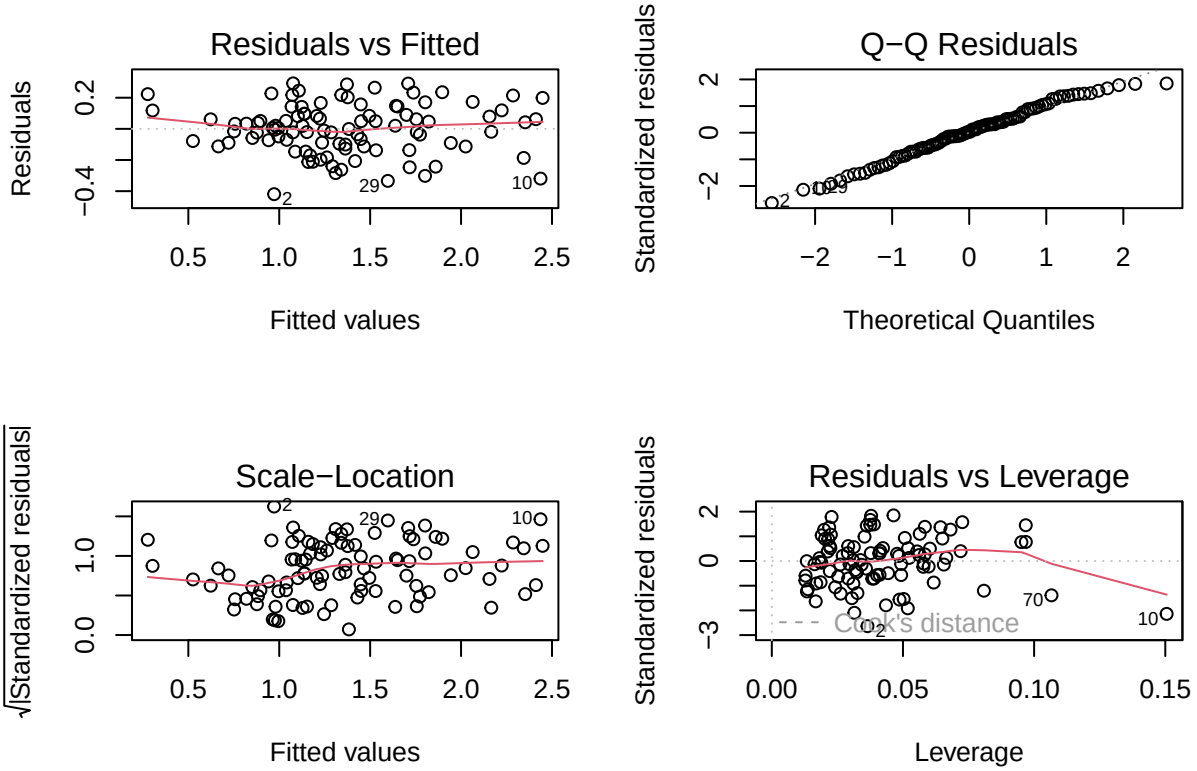
The latter model turns out to be (with some rounding)

$$\widehat{PPG} = 1.025 + 0.936 \times npxG90 - 0.666 \times PSxG90 + 0.029 \times ShotStop.$$

What this tells us is that for each game, starting from a base value of 1.025 (slightly more than the number of points for a 0-0 draw), each non-penalty expected goal generated by the offense adds 0.94 to the number of points earned from a game while each non-penalty expected goal allowed by the defense subtracts 0.666 from the number of points earned. Each percentage point by which the goalkeeper is able to reduce the chances of a shot on target resulting a goal also adds 0.029 to the points per game.

The overall F-test for this model gives us a p-value below 2.2×10^{-16}, so the dependencies captured by the model are statistically significant at the $\alpha = 0.05$ level. Also, the adjusted R-squared for this model is 89.4%, while the adjusted R-squared for the model without shot-stopping is 84.8%, so the inclusion of shot-stopping does appear to add to the effectiveness of the model. We can also verify this with a partial F-test, for which the resulting p-value is 4.63×10^{-9}. This low p-value means that the additional variable is a significant inclusion to the model.

2.1 Validity of model assumptions



In order for us to be able to use this model with some degree of confidence, we need to check the underlying assumptions of the linear regression model, which we do with the LINE criteria.

In the Residuals vs Fitted plot, we see that the residuals generally lie in a horizontal band centered at 0 with no obvious trend, so linearity and homoscedasticity appear to be met; this is also backed up by the Scale-Location plot. That said, there is a little bit of tapering for low fitted values that may indicate some loss of homoscedasticity. The Q-Q Residuals plot generally is consistent with normality of residuals since most points lie along the diagonal reference line.

These statements can be backed up with the use of some analytic tests. The Shapiro-Wilk test checks whether values fit a normal distribution, with the null hypothesis being that a normal distribution is an appropriate fit and the alternative (low p-value) being that a normal distribution is not appropriate. When we apply the Shapiro-Wilk test to the residuals, we get $p = 0.368$, consistent with normality.

For heteroscedasticity, we can use the Breusch-Pagan test, for which the null hypothesis is that the residuals are heteroscedastic. This test gives us $p = 0.094$, which is consistent with heteroscedasticity at the $\alpha = 0.05$ significance level, but low enough that some further investigation may be warranted.

From the Residuals vs Leverage, we see that there are no points which have large Cook's distance, or even large leverage. This indicates that the data does not appear to have any outliers, so there are no points we need to throw out to redo the model.

3. Goalkeeper Actions and Offensive Output

While the goalkeeper is traditionally thought of as a purely defensive role, they can also play a role in the offense through ball retention (which doubles as a defensive tactic) and through initiation of offensive transitions. While this contribution generally pales in comparison to the roles of the outfield players, it is nonetheless of interest to see what actions that a goalkeeper performs can influence the offense of the team, as this allows the goalkeeper to contribute more to a team's success than shot-stopping alone suggests.

For this analysis, we look at passing and goal kick statistics. One categorization that appears is that of whether a pass or goal kick is a *launch*, which is a kick that sends the ball forward at least 40 yards. There is much debate amongst both fans and analysts about how often goalkeepers should launch the ball compared to playing the ball short, and incorporating this factor through both frequency and accuracy is of interest. The statistics we consider are

- passing attempts per 90, pass launch percent, and pass average length;
- goal kick launch percent and goal kick average length;
- launch attempts per 90, launch completions per 90, and launch completion percent.

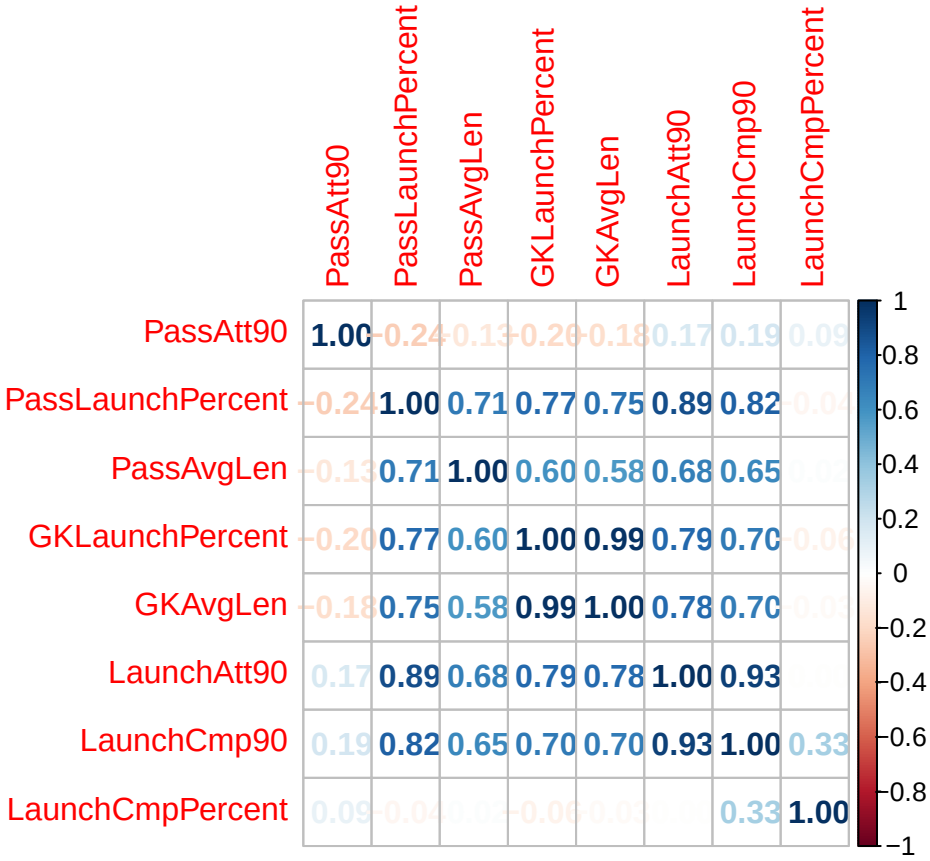
We omit goal kick attempts per 90, as the goalkeeper has very little influence in whether a passage of play results in a goal kick.

In a linear regression model for npxG90 using these variables, the adjusted R-squared is 41.6% which indicates a decently predictive model and the p-value from the F-test is 2.54×10^{-9}, so the model is statistically significant in the sense that we reject the null hypothesis of no linear dependence on any of the variables.

Somewhat notable is that in this model, none of the individual variables are found to be statistically significant given the other variables in place, with p-values far above the $\alpha = 0.05$ threshold. Given that the model overall was found to be statistically significant, and there are some variables that are defined in terms of others in the predictor set, we expect multicollinearities to be contributing to this apparent discrepancy.

3.1 Multicollinearities

To investigate the multicollinearities, we look at a correlation matrix.



Here we see that PassAtt90 and LaunchCmpPercent are very minimally correlated with the other variables, but the remaining variables are somewhat or very strongly correlated with each other, in particular LaunchAtt90 with LaunchCmp90 and GKLaunchPercent with GKAvgLen. We can quantify exactly which variables are more correlated with the others using the VIF.

Variable	VIF
PassAtt90	7.864
PassLaunchPercent	27.562
PassAvgLen	2.151
GKLaunchPercent	52.591
GKAvgLen	47.675
LaunchAtt90	114.767
LaunchCmp90	63.063
LaunchCmpPercent	7.960

Here we see several predictors above the rule-of-thumb $VIF = 5$ threshold for multicollinearity, and so as a heuristic approach to removing some redundancies, we remove predictors one at a time greedily.

Variable	VIF
PassAtt90	1.073
PassLaunchPercent	3.149
PassAvgLen	2.049
GKAvgLen	2.276
LaunchCmpPercent	1.014

After this process, we have kept passing attempts per 90, pass launch percent, pass average length, goal kick average length, and launch completion percent. This new model has a slightly improved adjusted R-squared of 42.5%, and the PassLaunchPercent and LaunchCmpPercent variables are individually significant according to the t-test.

3.2 Model selection using AIC

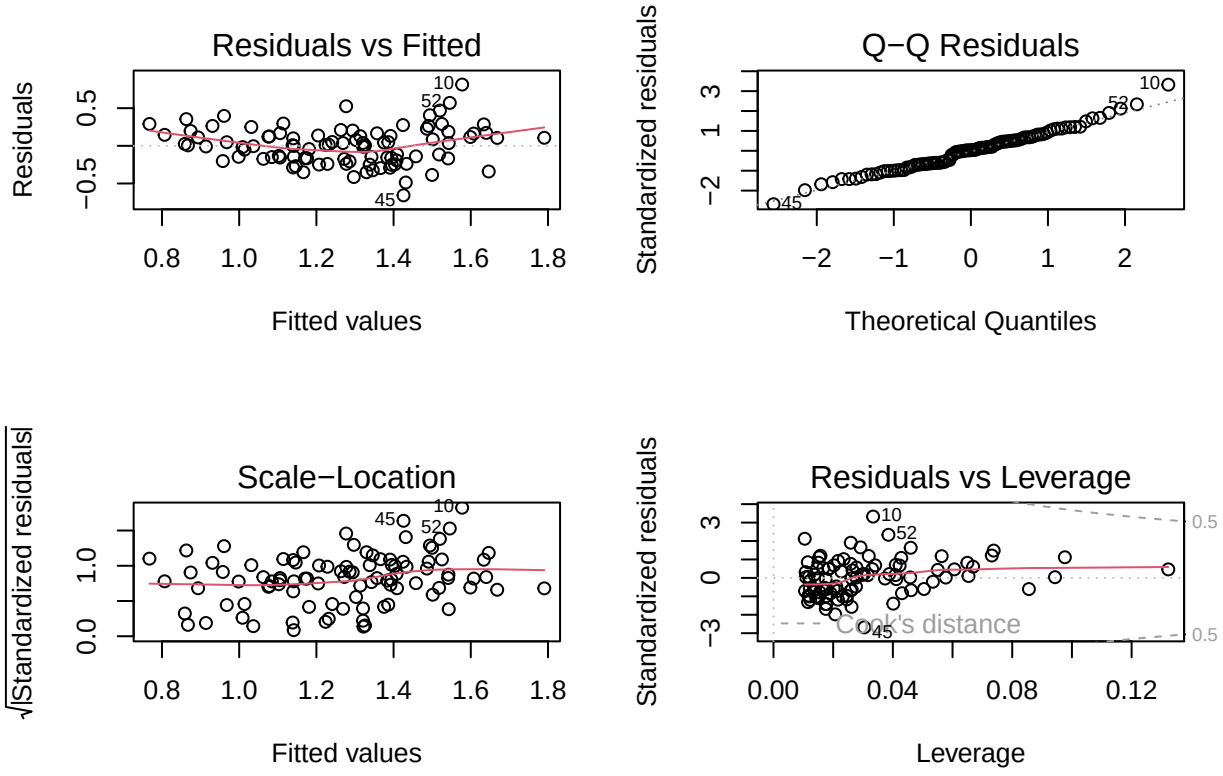
From this reduced set of variables to consider, we now look for the best model in accordance with a stepwise algorithm based on the Akaike Information Criterion (AIC).

Direction	Variables	AIC
Backward	PassLaunchPercent + LaunchCmpPercent	-264.15
Forward	PassLaunchPercent + LaunchCmpPercent	-264.15
Both	PassLaunchPercent + LaunchCmpPercent	-264.15

All three algorithm variants produce the same model, using PassLaunchPercent and LaunchCmpPercent, with an AIC of -264.15 and an adjusted R-squared of 44.1%. Thus we take this as our final model.

3.3 Model adequacy

We quickly check whether this regression model is adequate by a similar approach to what we did before.



From these plots, as well as the Shapiro-Wilk and Breusch-Pagan tests, we can see that the normality and heteroscedasticity assumptions are met. However, the Residuals vs Fitted plot indicates a weak non-linear trend in the residuals, suggest that some variable transformations are warranted. These may be investigated in the future.

3.4 Conclusions for this model

The final model produced is

$$\widehat{npxG90} = 1.498 - 0.017 \times PassLaunchPercent + 0.010 \times LaunchCmpPercent.$$

This means that from a baseline npxG90 of 1.498, each additional percentage points of pass attempts which are launched forward decreases the (non-penalty) expected goals per 90 by 0.017, while each additional percentage point of launches completed (combining passes and goal kicks) increases the expected goals per 90 by 0.010.

4. Goalkeeper Actions and Defensive Stability

While one of the ways a goalkeeper can prevent more goals is to be better at shot-stopping, another way they can do this is to reduce the number of shots and the average shot quality by being proactively involved in defensive actions. For simplicity, we take these as being captured by the PSxG90 statistic.

There are other ways a goalkeeper can be involved defensively, including via ball retention, but for our purposes we focus on the following:

- the proportion of opposition crosses that the goalkeeper stops (whether by clearing or by catching);
- defensive actions the goalkeeper takes outside of the penalty area per 90;
- average distance of the goalkeeper from the goal for any defensive action (in or out of the penalty area).

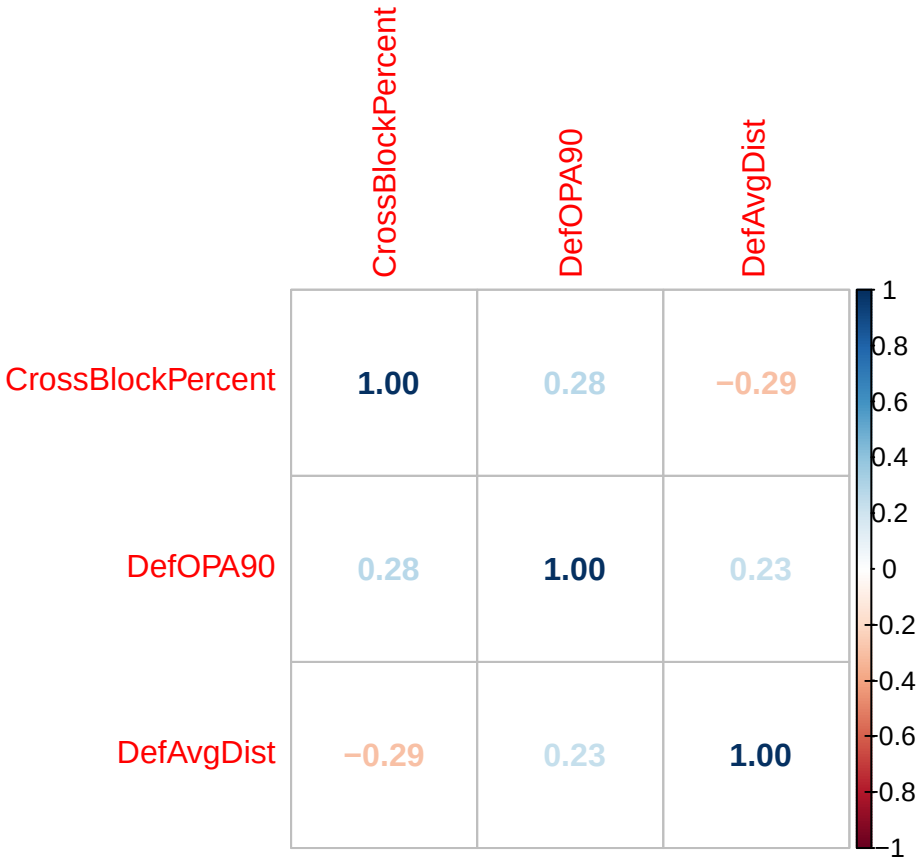
The model produced here is

$$\widehat{PSxG90} = 1.691 - 0.043 \times CrossBlockPercent + 0.239 \times DefOPA90 - 0.022 \times DefAvgDist.$$

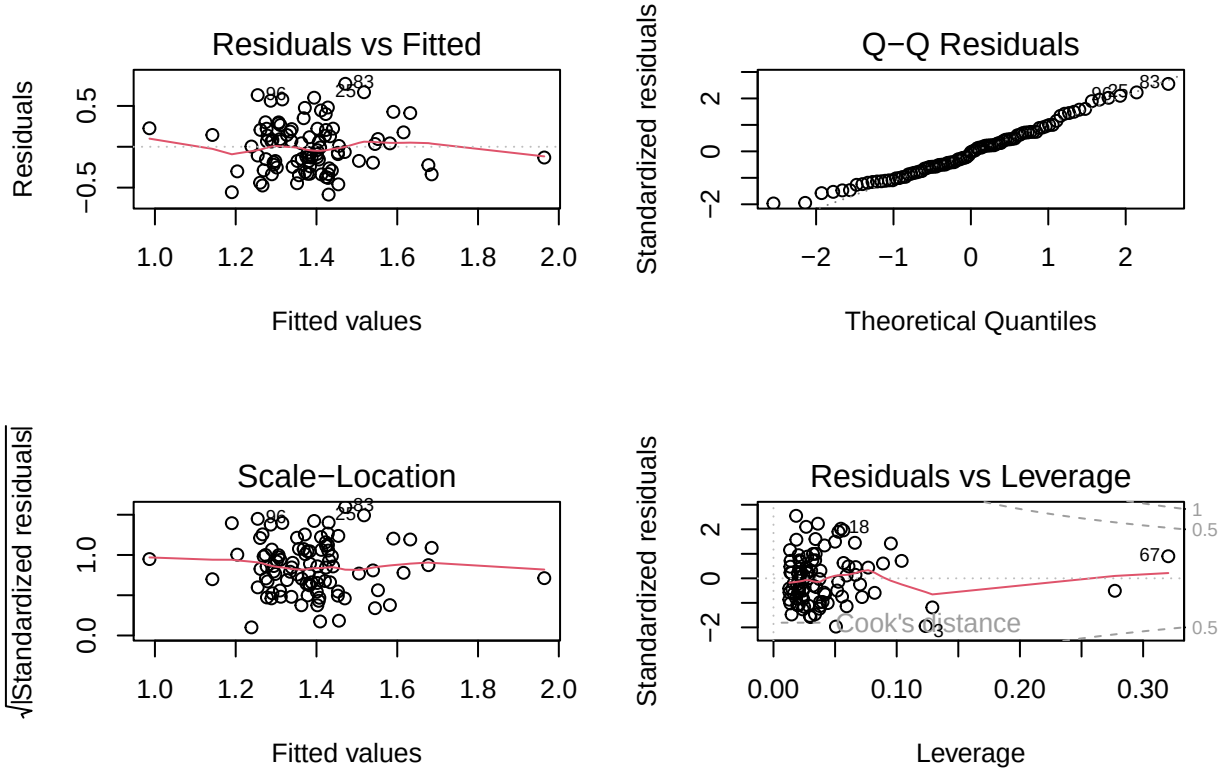
Starting from a base value of 1.691, each percentage point increase in the proportion of opposition crosses stopped by the goalkeeper decreases the PSxG90 by 0.043, each defensive action outside the penalty area (per 90 minutes) increases the PSxG90 by 0.239, and each additional yard of distance for the average defensive action decreases the PSxG90 by 0.022. The overall model has an adjusted R-squared of 12.1%, which indicates that most of the variability in PSxG90 is explained by other factors. That said, the p-value found by the F-test is 0.0023, so the linear dependence found by the model is still significant even if it leaves much to be explained. Each individual variable is also found by the t-test to be significant at the $\alpha = 0.05$ level.

4.1 Model adequacy checks

As in previous models, we look at multicollinearity and at various residual plots.



From the correlation matrix, none of the three variables correlate strongly with each other, so we can safely proceed with all three variables. The VIF values also turn out to be between 1 and 2, backing up this claim.



The Q-Q Residuals plot is consistent with normality aside from perhaps some deviation at the lower tail, while the Residuals vs Fitted and Scale-Location plots are both consistent with linearity and heteroscedasticity. These claims are also consistent with the outputs of the Shapiro-Wilk and Breusch-Pagan tests. The Residuals vs Leverage plot indicates that there are no outliers, though there are a couple of points with unusually high leverage.

5. Final Models and Concluding Remarks

In summary, the models we have developed are

$$\widehat{PPG} = 1.025 + 0.936 \times npxG90 - 0.666 \times PSxG90 + 0.029 \times ShotStop,$$

$$\widehat{npxG90} = 1.498 - 0.017 \times PassLaunchPercent + 0.010 \times LaunchCmpPercent,$$

$$\widehat{PSxG90} = 1.691 - 0.043 \times CrossBlockPercent + 0.239 \times DefOPA90 - 0.022 \times DefAvgDist.$$

Each of these have been found to be statistically significant at the $\alpha = 0.05$ level and meet the assumptions needed for linear regression, although as noted in Subsection 3.3, we may still get some improvement with a transformed model for npxG90. The adjusted R-squared values for these three models are 89.4%, 44.1%, and 12.1% respectively, while the multiple R-squared values are 89.8%, 45.2%, and 15.0%, reflecting the variability explained by the predictors we considered.

Through either training or the transfer market, if a team is able to improve upon the goalkeeper's PassLaunchPercent by 10 points (lower) and the LaunchCmpPercent by 5 points (higher), both of which are within one standard deviation, the predicted improvement in npxG90 is roughly 0.22, which in turn leads to an improvement of about 0.2 PPG. This is roughly 7 or 8 points in a full season, and while this does not turn a relegation team into contenders, it is a very meaningful amount of points for teams looking to survive relegation, looking to move up from midtable to a spot in European competition, or looking to secure a championship. Changes on the defensive end are less pronounced and harder to improve upon (thinking in terms of standard deviations of improvement), and analysis of other team defensive statistics would likely need to be incorporated.

While shot-stopping also has positive impacts on PPG, one deficiency in the analysis is that it is not clear how consistent of a measure of goalkeeping quality it is. While it is often counterintuitive that goalkeeper quality might not be reflected in outperformance of PSxG, this is a phenomenon already observed in soccer analytics on the offensive end. In particular, while xG does not take into account the identity of the player taking a shot, it is observed that very few players significantly deviate from their xG in the long run, so "goals minus expected goals" would not be a useful metric for signing players. (Famously, Cristiano Ronaldo is regarded as one of the greatest strikers of all time, yet his goal tally is at or slightly below what xG predicts. His goal count is because he is incredibly proficient at generating higher-xG shots.) A similar phenomenon with goalkeepers would indicate that goalkeepers would be better assessed by their ability to contribute to xG (or npxG) and PSxG rather than "pure shot-stopping." To fill in the gap in the analysis, we would need to collect data from multiple years, split by individual goalkeepers, and track performance over time.

Putting this together, our overall recommendation based on these models is to prioritize training goalkeepers to be comfortable on the ball to play out of the back with short passes and to be able to launch passes to a deliberate target space as often as possible (rather than aimless clearances). In addition, the overall team movement should facilitate playing in this manner so as to improve offensive outcomes. On the defensive end, a goalkeeper who is more comfortable claiming crosses and defending away from goal will contribute to reduced PSxG90 even if they do not directly improve their shot-stopping. The variables considered in these models can largely be controlled by the team itself, without regard to what the opposition does, and hence are more actionable in training and may see greater or more consistent improvements.