Phenomenon of Robert Lewandowski. Analysis of strikers' performance in the best European soccer clubs.

Sebastian Brylka

Introduction

In the world of soccer, 2020 was the year of Robert Lewandowski. His club, Bayern Munich, won all competitions they participated in, including the most prestigious soccer tournament in the world, UEFA Champions League. Individually, Lewandowski had the phenomenal year. It is enough to say that he was the top scorer in every competition he played in, outdistancing every other striker in the world. Fully deserved, Lewandowski was chosen The Best FIFA Men's Player of the year, as the first Polish player in history. I had the pleasure to observe his success and experience this Lewandowski-mania while being in Poland, which made me wonder – why Lewandowski is so good? Unfortunately, back then I did not have enough tools to perform statistical analysis of his performances, but after taking STAT 206, I am finally ready to answer my question. I decided to separate my research into 3 parts – in the first one, I will compare different statistics of the best strikers to create a model predicting how many goals a player can score in one season. In the second one, I will compare the striker's performances in different settings, for example at which age do they perform the best or in which league they play in. Of course, every part of my analysis will refer to Robert Lewandowski and his performances. In the last part, I will focus solely on Lewandowski and calculate how probable he is to score 20 goals this season. I hope that at the end of my research, I will be able to answer the fundamental question of my study – what makes Lewandowski such a good soccer player?

Data

All data are observational and come from the website https://www.sports-reference.com/. I gathered the statistics of subjectively chosen 27 best strikers of the last 10 years, where each season of a striker is treated as one unit, summing up to 386 units. Unfortunately, more detailed statistics started being recorded only 4 years ago and only in Top 5 Leagues (England, Spain, Germany, Italy, France), which decreases number of units with more detailed stats to 103. The statistics come only from the league games. The database was obtained on 28th of April 2021.

What makes a striker successful? Finding a regression model to predict number of goals scored by a striker

Model 1

The first question that comes to mind is – how do we know the striker is successful? The answer is very simple – by the number of goals he scores. While for any other position we cannot really choose one parameter that tells how good a player is, with strikers the case is simple – he must score goals. One can argue that more and more teams decide to play without a classical striker or decide to play with a striker

whose main task is not to score goals, but to absorb the defenders of the opposing team and create the space for the other players, but in my study, I will take a classical approach and judge the striker by the number of goals he scores. To account for different sizes of each leagues and different number of games played by each player, the response variable will be number of goals scored per 90 minutes of game time of a player. To obtain reliable results, I have decided to subset my database to players who played over 900 minutes in a season (10 full games), which left me with 338 units.

As I mentioned in the description of the data, the dataset does not include detailed statistics for most of my units, so the first model I want to derive is the one including only most basic statistics, available for all units.

Response variable	Possible Predictors
Goals per 90 minutes	Games played, Games started, Minutes played, 90s played (Minutes played/90), Goals, Assists, Non-penalty goals, Yellow cards, Red cards, Shots on target, Shots on target per 90 minutes played, Goals per shot on target, …

Using best subsets method, I find out that best models with 2 and more predictors have very high values of adjusted r-squared (over .93). However, when I want to account for what information might be available in potential predictors that are not in the model and look for the models with small values of Mallows's Cp, I conclude that the model with 7 predictors is the best. In this model, adjusted r-squared of .97 is almost the same as in other models, but Mallows's Cp is the smallest, smaller than the number of predictors in the model + 1.

	model	p	rsq	rss	adjr2	cp	bic	stderr
1	Gl	2	0.740	6.347	0.740	2594.47	-441	0.1378
2	P-Gl	3	0.932	1.656	0.932	433.74	-887	0.0705
3	Gl-SoT-ST9	4	0.948	1.282	0.947	263.08	-967	0.0621
4	Gl-SoT-ST9-G.	5	0.963	0.914	0.962	95.48	-1075	0.0526
5	MP-Gl-SoT-ST9-G.	6	0.967	0.797	0.967	43.38	-1115	0.0491
6	P-Gl-GP-SoT-ST9-G.	7	0.970	0.727	0.970	13.15	-1140	0.0470
7	MP-P-Gl-GP-SoT-ST9-G.	8	0.971	0.706	0.970	5.58	-1144	0.0464
8	MP-St-P-Gl-GP-SoT-ST9-G.	9	0.971	0.704	0.971	6.45	-1140	0.0464

Figure 1. R-output for best subsets method

The predictors used in the best model are Matches Played, Played 90s, Goals, Non-penalty goals, Shots On Target, Shots On Target per 90 Minutes and Goals per Shot On Target. However, using this set of predictors leads to big issues with multicollinearity, which is represented by the VIF values for predictors (for most predictors, VIF in Figure 2 is much higher than 5).

MP	Played90s	Gls	GminusPK	SoT	SoTper90	GperSoT
3.978707	12.381586	17.996665	27.306358	36.017167	15.159976	4.638062

Figure 2. VIF values for predictors

After dropping 2 predictors with highest VIF, Non-Penalty Goals and Shots On Target, the value of adjusted r-squared of the model decreases to 0.96, which is not a significant drop. At the same time, situation with multicollinearity looks much better, as the VIF in Figure 3 values are much lower.

MP	Played90s	Gls	SoTper90	GperSoT
3.813794	6.678802	7.534036	3.763809	2.418380

Figure 3. VIF values for predictors

However, I am still worried about the multicollinearity in the model, so I keep experimenting with dropping some predictors. It turns out that after dropping only Goals predictor from the model, adjusted r-squared dropped to 0.86, while after dropping only Played 90s, adjusted r-squared changed to 0.94, which is still an acceptable value, but resolves the problems with multicollinearity (VIF in Figure 4 for all predictors is lower than 5).

MP	Gls	SoTper90	GperSoT
1.781150	4.326117	2.988652	2.022221

Figure 4. VIF values for predictors

Now, the predictors in my model are Matches Played, Goals, Shots on Target per 90 Minutes and Goals per Shot On Target.

Following a four-step process, my model is

$$GlsPer90 = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

The fitted model is

$$\widehat{GlsPer90} = -1.721 - 0.014 * MP + 0.025 * Gls + 0.333 * SoTper90 + 1.832 * GperSot$$

Adjusted r-squared for this model is 0.946 and all predictors are significant with 95% confidence level as all p-values for coefficients are below 0.05.

Assessing the model

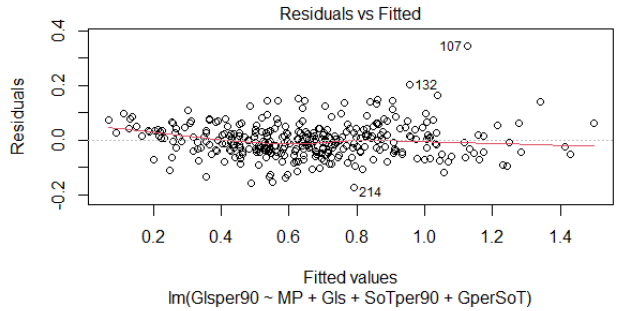

Figure 5a

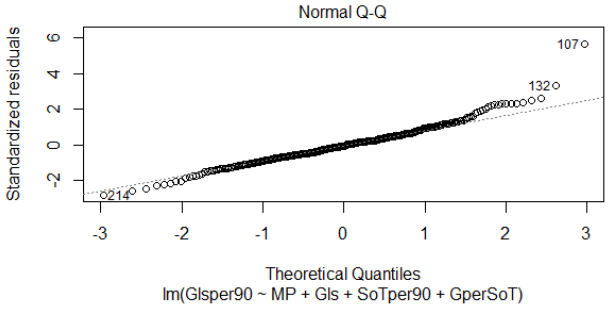

Figure 5b

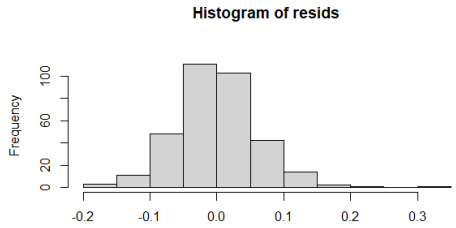

Figure 5c

All observations are random and independent. The normal quantile plot in Figure 5b looks fairly straight, with one huge outlier at the high end. The value of the Cook's distance of this outlier is 0.37, which means it is not influential. The residuals versus fitted plot in Figure 5a shows a consistent band of residuals on either side of the zero line. Histogram of residuals is also relatively normal and centered around zero. There are no concerns with the conditions.

How can I use my model? I want to see how well the model predicts number of goals per 90 minutes Robert Lewandowski would have scored in 2019/2020 season. Using Lewandowski's stats from that season, I am 95% confident that Robert Lewandowski would have had between 0.993 and 1.242 goals per 90 minutes that season. In reality, he had 1.11 goals per 90 minutes, which is somewhere in the middle of the prediction interval, so the model does a good job predicting this value.

As we can see, my model does a good job in predicting number of goals per 90 minutes of game time of a striker, but some takeaways may be surprising. The coefficient of matches played is negative is negative, which means that, allowing for simultaneous change in other predictors, when the striker plays one more game, the predicted number of goals per 90 minutes decreases. It may be caused by the fact that strikers are very often substituted before the end of the game if they do not play well, so if they do not score and get subbed off before playing full 90 minutes, we do not add full 90 minutes to minutes played, but we add the full game to matches played, so it may lower the value of goals per 90 minutes played. The other coefficients are positive, which also makes sense – the more goals they score, the more shots on target per 90 minutes they have and the more goals they score from one shot on target, the higher value of goals per 90 minutes they have (again, allowing for simultaneous change in other predictors).

Model 2

I have found the model predicting number of goals the striker scores per 90 minutes played using the most basic explanatory variables. Now, I will try to create the model predicting the same response but using more advanced statistics. As I explained in my data description section, the detailed data are available only for a few leagues and only for the last 4 seasons, which will drastically decrease observational units I can use. Again, I will use only the data for seasons when a player played more than 900 minutes. All these restrictions decrease my sample size to 98 observations.

The response variable stays the same, number of goals scored by the striker per 90 minutes played, but I have 46 potential explanatory variables, which are, for example, player's expected goals, successful passes, number of players dribbled or touches in the opponent's penalty area. Because of the high number of predictors, I will choose the predictors for my model by stepwise regression. After running the automated process for stepwise regression in R, I get the model with 10 predictors with adjusted r-

squared of 0.85. After checking VIF for all predictors, I can see that multicollinearity is not the issue here (all values lower than 5), so I will use backward elimination with the significance level of 5% to make my model simpler without decreasing adjusted r-squared significantly.

After running the backward elimination process, I end up with a model with 7 predictors, all of which are significant under 95% confidence level because of low p-values for each coefficient. The value of adjusted r-squared of this model is 0.84, which is a minimal decrease from 0.85 in 10 predictors model. The predictors in this model are

xG	Expected goals[1]
npxGperSh	Non-penalty expected goals per one shot
TouchesAttPenper90	Touches of the ball in the opponent's penalty area per 90 minutes
Dist	Average distance, in yards, from goal of all shots taken
npGminusxG	Non-penalty goals minus expected goals
PKatt	Attempted penalty kicks
Passper90	Passes per 90 minutes

Using the traditional four-step process, the model is

$$GlsPer90 = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * X_5 + \beta_6 * X_6 + \beta_7 * X_7 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Fitting the model, I get

$$\widehat{GlsPer}90 = -0.554 + 0.008 * xG + 2.83 * npxGperSh + 0.048 * TouchesAttPenper90 + 0.021 * Dist + 0.038 * npGminusxG + 0.014 * PKatt - 0.118 * Passper90$$

Assessing the model

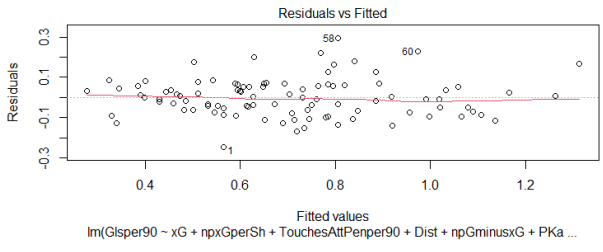

Figure 6a

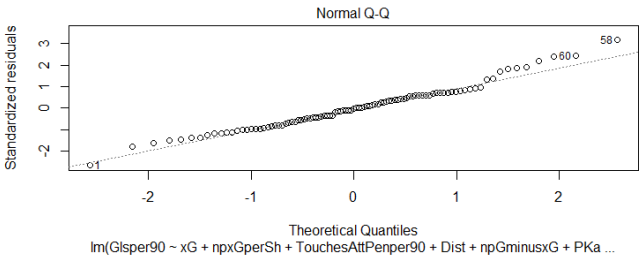

Figure 6b

[1] https://www.americansocceranalysis.com/explanation

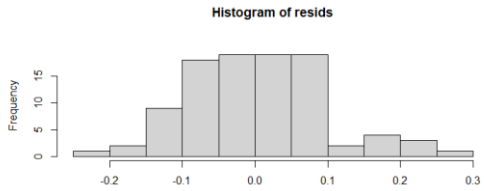

Figure 6c

All observations are random and independent. The normal quantile plot in Figure 6b looks fairly straight. The value of the Cook's distance of the biggest outlier is 0.12, so there are no influential outliers. The residuals versus fitted plot in Figure 5a shows somewhat consistent band of residuals on either side of the zero line, but we can notice cluster of points to the left side of the plot. Also, the histogram of residuals is skewed to the left, so I am a little worried about the conditions. To be sure that the relationships in my model are significant, I will run a bootstrap test for each coefficient.

After generating a bootstrap sample of 5000 repeats and getting confidence intervals based on bootstrap distribution for each of my coefficients, I notice that none of the 95% confidence intervals included 0, which makes me confident that all my coefficients are significant.

To use my model, again I will try to predict the number of goals per 90 minutes Robert Lewandowski should have scored last season. Using my model and Lewandowski's stats from 2019/2020 season, I am 95% confident that he should have scored between 0.857 and 1.26 goals per 90 minutes, and in fact he had 1.11 goals per 90 minutes, so we can see that the model does a good job again.

Model 2 explains a little less variability in the response than model 1, having the adjusted r-squared of 0.84 compared to 0.94, but using less obvious set of predictors makes it much more interesting from the soccer point of view. When analyzing predictors one by one, we can notice some interesting things. Firstly, the predictor of xG with positive coefficient is pretty obvious – the more goals the striker is expected to score from all shots he makes, the more goals per 90 minutes he has. The other predictor, non-penalty expected goals from per shot with positive coefficient, is much more interesting. It tells us that the striker is better off not shooting from every position, even when the probability of scoring is relatively low, but instead he should focus on finding opportunity to shoot from better and easier positions, when xG is higher. It may be opposed to a popular belief that the strikers should shoot even in hard situations and "let the goalie make a mistake". The predictor of non-penalty goals minus expected goals also gives an interesting conclusion – to score more goals, the striker should overperform and score more than he is expected to from all his shots. It may come down to the player's quality as the value of the xG in a certain situation does not depend on the shooter. Better strikers can score in harder situations (situations with lower xG), so overperforming xG definitely helps them perform better. The predictor of distance is a little surprising. Intuitively, the closer to the goal, the easier it is to score, but the model tells us that as the average distance of all shots increases, the ratio of goals per 90 minutes also increases (allowing for simultaneous change in other predictors). I think it may be caused by the fact that the closer to the goal, the more defenders are on the way, so it is easier to try to shoot between the defenders as the ball is always faster than the players, but more research is needed to explain it. The predictor of Penalty Attempts with positive coefficient is obvious – the more penalties the striker takes, the more goals he scores. Last 2 predictors can be explained together. In soccer, most of the action takes play in the middle of the field, that is outside the penalty box. Players playing in this

area usually have the most passes, but they also do not have as many shots as the players waiting for the ball in the penalty box, that is strikers. Negative coefficient of the predictor of passes per 90 makes sense because the more passes the player makes, the more he is involved in the middle of the field, where the chances of scoring are much lower than from the penalty area. On the other hand, the more touches inside the opponent's penalty area he makes, the more chances to score he has, which of course also means he scores more goals.

Summing up the regression models, both models do a good job in predicting the striker's goals per 90 minutes. The first model explains more variability in the response, but the second model gives much more insight into the way the striker has to play to score more goals. If I were to recommend how to use my models, I would say that the first model may be used by a fan who is interested if the striker of his team does as well as he is predicted to, while the second model could be used by the analysts of the professional soccer team who analyze the performance of their strikers and then tell them what to change in their game.

Now, the question is: do my models tell me anything about Robert Lewandowski and his impressive last season? What changed in his game that made him the best striker in the world? I will compare his performance from the season 2018/2019, when he had only 0.67 goals per 90 minutes, to the following season, when he scored impressive 1.11 goals per 90 minutes. First, I will compare my predictors from the first model.

	2018/2019	2019/2020
Matches played	33	31
Goals	22	34
Shots on Target per 90	1.49	1.92
Goals per Shot on Target	0.39	0.49

We can see that he played 2 less games in the 2019/2020 season but scored 12 more goals. This comparison shows us that he significantly improved the number of shots on target per 90 minutes and improved the efficiency of these shots, converting more shots on target into goals. In general, his finishing significantly improved, which could happen because of his hard individual work after the disappointing 2018/2019 season. Also, in the summer of 2019 there was no big tournament for national teams, so maybe getting these extra few weeks of rest played a role. Some people also say that the more experienced you are, the better finisher you become, so maybe it also mattered.

We can come to more interesting conclusions from the predictors of the second model

	2018/2019	2019/2020
Expected goals	28.4	27.5
Non-penalty expected goals per one shot	0.18	0.18
Touches of the ball in the opponent's penalty area per 90 minutes	8.39	8.37
Average distance, in yards, from goal of all shots taken	12.3	12.7

Non-penalty goals minus expected goals	-6.4	5.3
Attempted penalty kicks	4	5
Passes per 90 minutes	0.3	0.42

From this comparison we can see that most statistics did not change by much, which means Lewandowski did not change the way he plays. The only thing that really changed was the difference between non-penalty goals and expected goals. In 2018/2019, he missed a lot of good chances and scored over 6 less goals than he was expected to from all his shots. Next season, his finishing drastically improved. Despite having a very similar number of expected goals, this time he overperformed his xG and scored over 5 more goals than he was expected to. Again, we can conclude that Lewandowski's finishing got much better between the seasons and that is how he became the best striker in the world.

Does it matter for a good striker where he plays? Finding differences between different factors using ANOVA analysis

Model 3

The next question I want to answer is: is Lewandowski just very lucky because he was in a good league to score a lot of goals? I will perform the ANOVA analysis for the variation of number of goals per 90 minutes depending on the league the striker plays in. I will subset my data to the strikers who played over 900 minutes in a season in one of the top 5 leagues, which leaves me with 285 observations. Because all data is observational, we cannot generalize the results if the study to the whole population. Every conclusion applies only to the particular group of strikers and the seasons I chose.

Again, the response is the number of goals scored by a striker per 90 minutes, but the explanatory factor is the league he plays in. League has 5 levels: English, German, Spanish, Italian and French.

The ANOVA model where α is a league effect:

$$Glsper90 = \mu + \alpha_i + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$i = 1,2,3,4,5$$

I will check the conditions by looking at the boxplot of the number of goals per 90 minutes in different leagues and the normal qq-plot of residuals

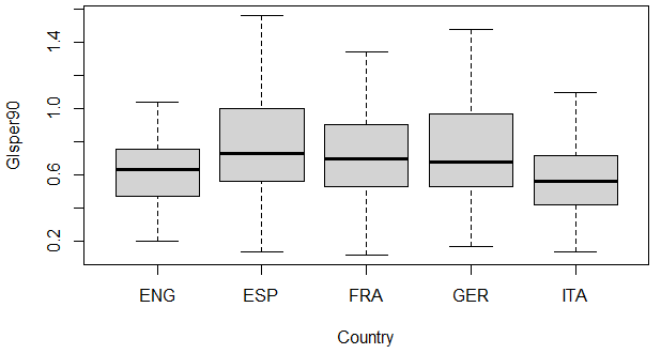

Figure 7a

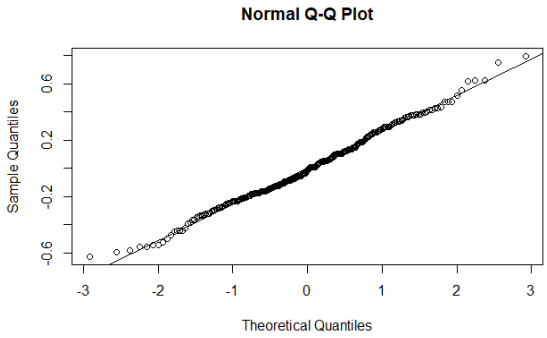

Figure 7b

In the Figure 7a, I can see there are no outliers in the data and the boxplots are relatively symmetric. The Levene's test for the homogeneity of variances gave me a p-value of 0.007, which implies that not all variances are equal, but the ratio of the highest and lowest standard deviation is less than 2, so I decide to stick with the ANOVA model and consider the condition of constant variance satisfied. Also, the normal quantile plot of residuals in Figure 7b shows that the distribution of residuals is reasonably normal. All residuals are random and independent, and all effects are additive.

The hypothesis test:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$$

$$H_a: Some\ \alpha_i \neq 0$$

When I fit the model in R, I find out that the F-statistic for the test is 5.508 and the p-value is .000279, which is way below the significance level of 0.05. I reject the null hypothesis. I am 95% confident that there is a difference in scored goals per 90 minutes between at least 2 leagues. I want to know which leagues differ, so I use the Tukey's HSD to get the intervals for the difference in the response.

	diff	lwr	upr	p adj
ESP-ENG	0.14138889	0.02251424	0.26026354	0.0106980
FRA-ENG	0.08879274	-0.05301615	0.23060162	0.4238159
GER-ENG	0.10430556	-0.05068802	0.25929914	0.3483544
ITA-ENG	-0.03867567	-0.15796815	0.08061682	0.9003856
FRA-ESP	-0.05259615	-0.19440504	0.08921273	0.8467859
GER-ESP	-0.03708333	-0.19207691	0.11791025	0.9652354
ITA-ESP	-0.18006455	-0.29935704	-0.06077207	0.0004323
GER-FRA	0.01551282	-0.15769684	0.18872248	0.9991891
ITA-FRA	-0.12746840	-0.26962773	0.01469093	0.1024312
ITA-GER	-0.14298122	-0.29829550	0.01233306	0.0874678

Figure 8. Tukey's HSD C. I. for differences

From the confidence intervals, we can see that the only pairs where the response differs in significant level are Spain-England and Italy-Spain. In both cases, I am 95% confident that the strikers score more in Spanish league because of the confidence intervals.

As we can see, there is no difference between German league and any other league in terms of scored goals per 90 minutes, so we cannot tell Lewandowski had any handicap because of the league he played in.

Model 4

Maybe there are some other factors that helped him? What about the age? There are many theories when the strikers perform the best. Some people say that they score more goals when they are young, strong, and full of youth fantasy, but when they get older, the defenders know how to stop them. The other theory says that they play better as they get older and gain more experience.

For the sake of simplicity, I decided not to treat every age as a separate group, but to create the age groups such that

Number of the group	Age limits
1	Under 22
2	22-25
3	26-29
4	Over 29

The ANOVA model where α is an age group effect:

$$Glsper90 = \mu + \alpha_i + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$i = 1,2,3,4$$

Diagnostic plots

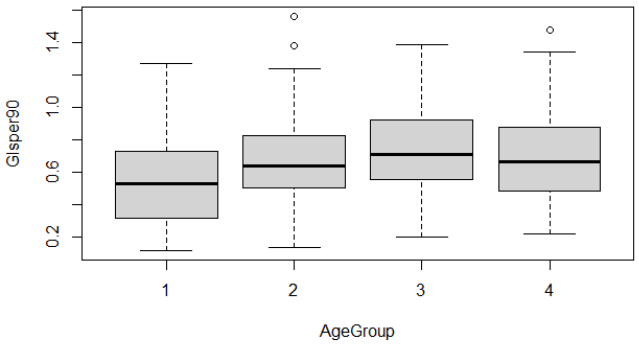

Figure 9a

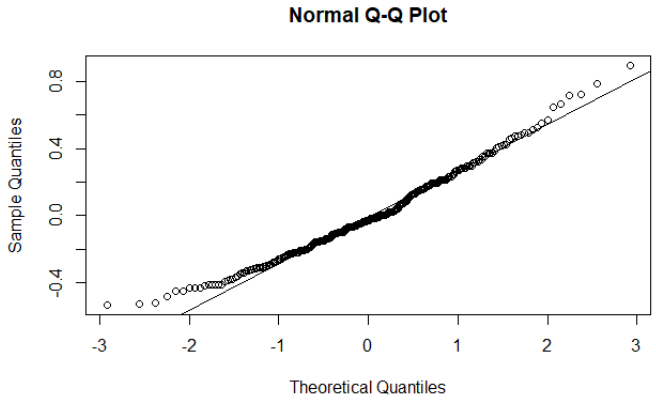

Figure 9b

The effects are additive and errors are random and independent. I can see on the boxplot in Figure 9a that the boxplots are relatively symmetric with few outliers. Also, the spreads seem to be very similar, but to be sure I perform the Levene's test. The high p-value of 0.67 confirms that the equality of variances. Looking at the normal qq-plot, I may be worried about the normality of the residuals as the data is skewed on both sides. To check if there are any better transformations of my data, I will generate a diagnostic plot.

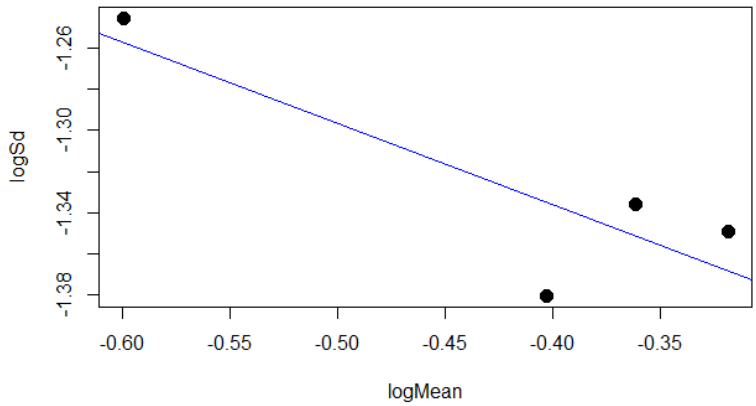

Figure 10. Diagnostic plot

The line fits the plot pretty well and the slope is about -0.39, but the transformation does not improve the normality of the data, so I will stay with untransformed model and assume the conditions are satisfied.

The hypothesis test:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$

$$H_a: Some\ \alpha_i \neq 0$$

When I fit the model, the F-statistic is 4.843 and p-value is .00226, much lower than the significance level of .05. I reject the null hypothesis again and conclude there is a difference for some age groups. To see which age groups differ, again, I will use the Tukey's HSD method

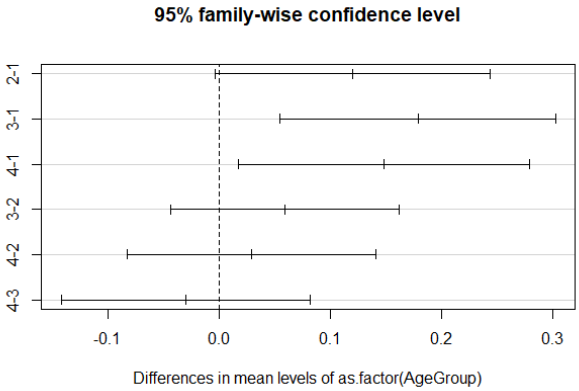

Figure 11

As we can see on the Figure 11, the age groups that differ are 3-1 and 4-1, and for both the confidence intervals are positive, so I am 95 % confident that players aged between 26 and 29 and players over 30 score more goals per 90 minutes than the players under 22. It tells us that usually the youngest players perform worse than the oldest players, but once they get some experience and are aged 22 or older, we cannot tell the difference between their performance and the performance of the most experienced players. Does that help us explain the phenomenon of Lewandowski? Since he reached his prime when he was 31, we can tell that he had some advantage over the youngest players, but there is not a significant difference between his age group and groups 2 and 3, so his age does not help us explain why his last season was so good.

Model 5

Previous ANOVA analysis showed some differences between the response in different leagues and between different age groups, but still, there was no difference between most treatments in both models under 95% confidence level. In this model, I will account for both league and age group in one model predicting number of goals per 90 minutes. First, I will look at the possible interaction between terms.

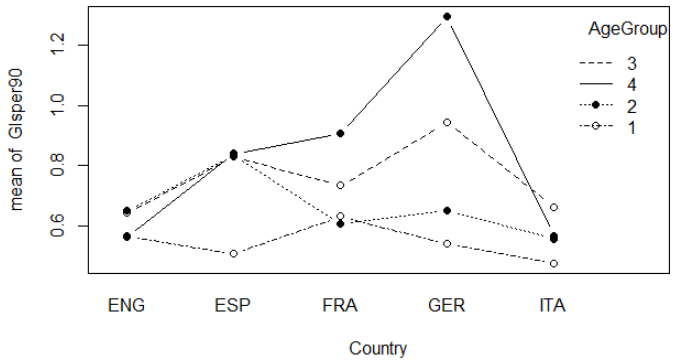

Figure 12. Interaction plot

In Figure 12, we can clearly see that the lines connecting the mean response on different combinations of levels of the factors are very different, which indicates the presence of the interaction between league and age group.

The model where α is a league effect and β is an age group effect is

$$Glsper90 = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$i = 1,2,3,4,5$$

$$j = 1,2,3,4$$

The hypothesis test

$$H_0: \gamma_{ij} = 0 \; for \; all \; i \; and \; j$$

$$H_a: Some \; \gamma_{ij} \neq 0$$

When I fit the model, I find out that F-value is 4.45 and p-value is very small, much lower than the significance level of .05. I reject the null hypothesis and I am 95% confident that there is an interaction between league and age group.

Diagnostic plots

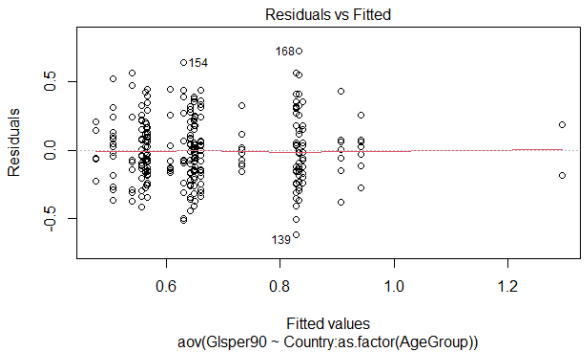

Figure 13a

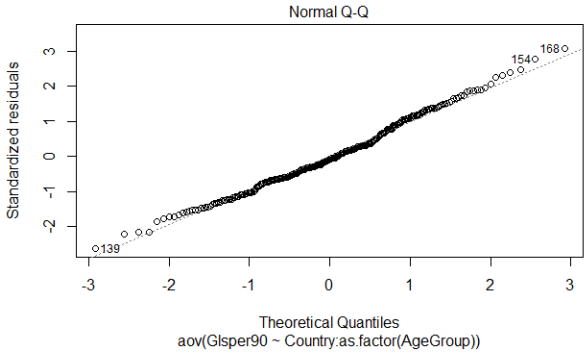

Figure 13b

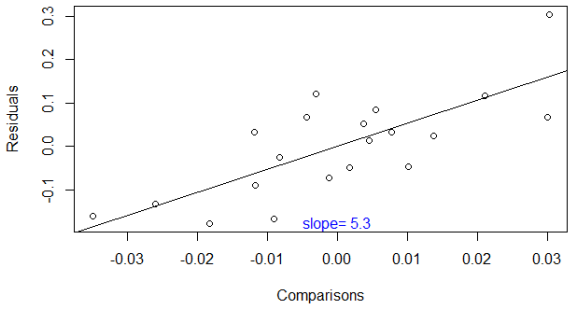

Figure 13c

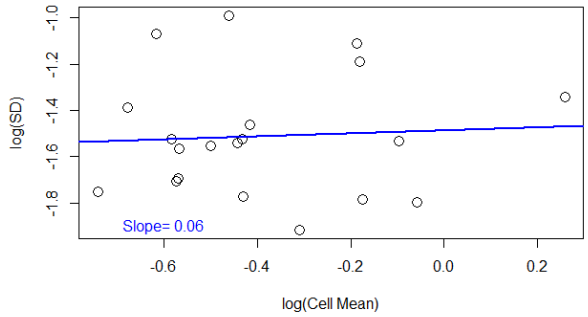

Figure 13d

In the figure 13a, we can see roughly equal distribution below and above zero, but points are more spread on the left, which may suggest transforming “down”. In the Figure 13b, there are no problems with normality. In the figure 13c, on Tukey additivity plot, we can see that the points follow some positive trend, which may mean that there is a scale for which the main effects are additive, but since we do not have many points on the plot and there is no obvious best-fitting line, we cannot tell that there is a scale for which the main effects are additive. In the figure 13d, the slope is almost equal to 0 tells that there is no better scale for which the standard deviations would be roughly equal.

After trying different "down" transformations, I found out that none of them gave a better fit to the model. Also, after trying the additive model with the transformation suggested by the Tukey's additivity plot, we see no improvement in the fit of the model. To ensure that the model with interaction is correct, I will use the randomization F-test.

The estimate of the p-value for the interaction term from the randomization test is 0.0052, which is below the significance level of 0.05. It gives me strong evidence that there is an interaction between the league and age group in the model. Using Tukey's HSD, we generate the confidence intervals for the differences. The differences of the combination in league and age group that are significant under the significance level of 0.05 (confidence intervals of the difference are positive) are

GER:4-ITA:1, ESP:3-ESP:1, ESP:2-ESP:1, ESP:4-ESP:1, FRA:4-ESP:1, GER:3-ESP:1, **GER:4-ESP:1**, **GER:4-GER:1**, ESP:3-ITA:2, ESP:2-ITA:2, FRA:4-ITA:2, GER:3-ITA:2, **GER:4-ITA:2**, **GER:4-ENG:4**, **GER:4-ENG:1**, ESP:3-ITA:4, ESP:2-ITA:4, FRA:4-ITA:4, GER:3-ITA:4, **GER:4-ITA:4**, **GER:4-FRA:2**, **GER:4-FRA:1**, **GER:4-ENG:3**, **GER:4-ENG:2**.

What these results give us in terms of analysis of Robert Lewandowski and his last season? Let's look how his group, GER:4, compares to the other groups (these confidence intervals are bolded). These results give us a strong hint that Lewandowski may have had some advantage over many other strikers from different leagues and different age groups and we can come to some interesting conclusions. We can say that we are 95% confident that the strikers over the age of 29 in German League score more goals per 90 minutes than all strikers in English league or that they score more goals per 90 than the strikers under 22 in all leagues. From this study we can conclude that German league is a good league for older strikers and Lewandowski benefited from this fact.

I want to emphasize the fact that all my data is observational. My conclusions apply only to the chosen group of strikers in the chosen period and the association between the response and explanatory factors is possible, but not sure. I do not generalize my results to all strikers in the world and all seasons. My model does not address the effects of the other factors that affect the number of goals scored by a striker per 90 minutes.

Will Robert Lewandowski score 20 goals this season? Using logistic regression to find the probability of Lewandowski scoring at 20 goals in one season

Model 6

Now, I want to use the logistic regression to calculate the probability of Robert Lewandowski scoring more than 20 goals in the current season, which is not over yet, so we will use the values of predictors up to date to calculate the probabilities. I decided to subset my database to the player who played at least 10 full games and scored at least 1 goal in a season, which left me with 372 observations.

I will start with a simple logistic regression. The response variable is whether a player scores more than 20 goals in one season. The response variable I chose is the number of goals per 90 minutes of playing time as I know that the total number of goals and goals per 90 are strongly correlated from the correlation matrix.

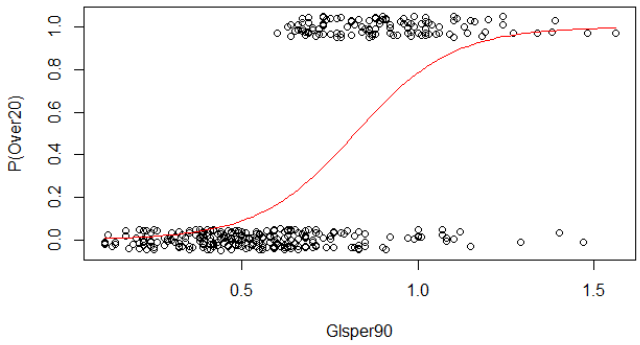

Figure 14

The model in logit and probability forms

$$\log(odds) = \beta_0 + \beta_1 * X$$

$$\pi = \frac{e^{\beta_0+\beta_1 X}}{1 - e^{\beta_0+\beta_1 X}}$$

Fitted models

$$\widehat{\log(odds)} = -5.9495 + 7.2421 * Glsper90$$

$$\hat{\pi} = \frac{e^{-5.9495+7.2421*Glsper90}}{1 - e^{-5.9495+7.2421*Glsper90}}$$

Assessing the model

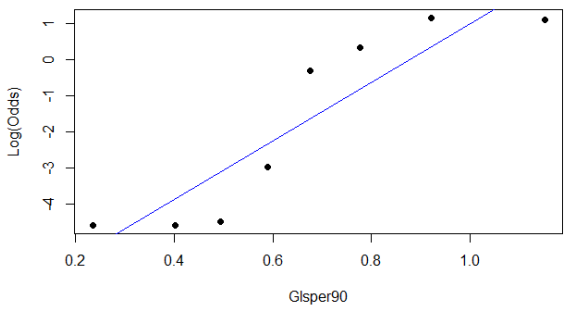

Figure 15

From the empirical logit plot in Figure 15, linearity seems reasonable. The outcome of number of goals in the season is not random but is a result of so many factors that we can apply a probability model. Also, all strikers are independent from each other, so the condition of independence is satisfied. However, we should remember that the results should serve as the tentative guidelines only as the data is not random.

To test the overall effectiveness of my logistic model, I conduct the hypothesis test

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

The p-value for the slope coefficient is much smaller than the significance level of 0.05, so I reject the null hypothesis. The 95% confidence interval for the slope is (5.732 to 8.752), which is consistent with the hypothesis test results as the interval does not include 0. To make sure that my conclusions are correct, I run the likelihood ratio test for utility of a simple regression model. P-value of 0 confirms the results from the first test. I have strong evidence that whether the striker scores at least 20 goals in a season depends on number of goals per 90 minutes.

Using the model, in the season 2020/2021 so far, Lewandowski scores 1.48 goals per 90 minutes, so, assuming that he keeps scoring with the same frequency for the rest of the season, the actual fitted values are

$$\log(odds) = -5.9495 + 7.2421 * 1.48 = 4.768808$$

$$odds(Over\ 20) = e^{0.1330524} = 117.778765965$$

$$\pi(Over\ 20) = 0.9915813$$

According to this model, the chance of Lewandowski scoring at least 20 goals this season is about 99%.

Model 7

According to the previous model, the probability of Lewandowski scoring at least 20 goals this season is around 99%, which makes it almost sure that he will score them. However, I used the number of goals per 90 minutes to predict total number of goals, so the strong correlation was obvious. In the next model, I will use more non-obvious predictors and implement them into multiple linear regression model. First, I will look at the possible logistic regression plots of the response against 3 chosen predictors – Matches Played, Shots on Target per 90 minutes and Goals per Shot on Target.

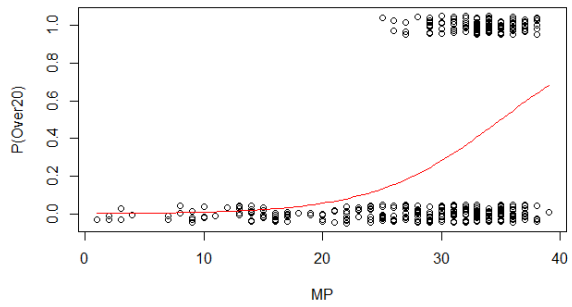

Figure 16a

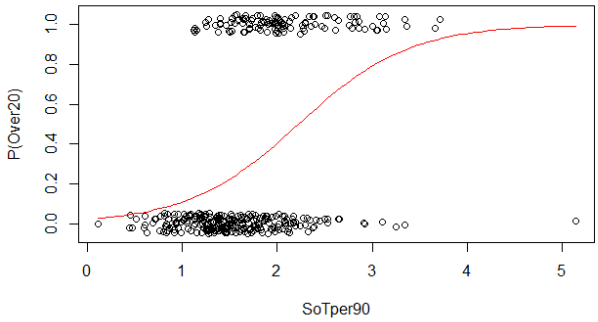

Figure 16b

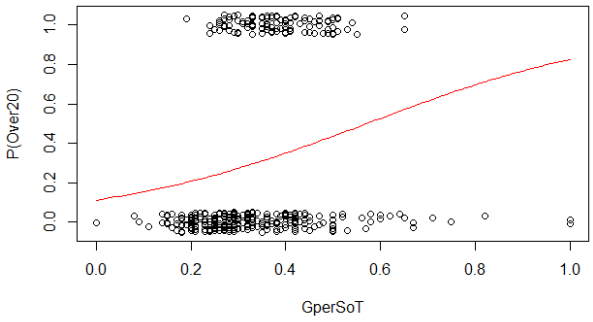

Figure 16c

On three plots in Figure 16 we can see that if we consider any of the predictors Matches Played, Shots on Target per 90 and Goals per Shots on Target alone, we can see a strong association with the response of whether a player scores 20 goals in a season.

The logistic model in logit form and probability form

$$\log(odds) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3$$

$$\hat{\pi} = \frac{e^{\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_3}}{1 - e^{\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_3}}$$

Fitted models

$$\widehat{\log(odds)} = -1.350538 + 0.0220868 * Matches\ Played + 0.3547686 * Shot\ on\ target\ per\ 90 + 1.24273 * Goals\ per\ shot\ on\ target$$

$$\hat{\pi} = \frac{e^{-1.350538+0.0220868*Matches\ Played+0.3547686*Shot\ on\ target\ per\ 90+1.24273*Goals\ per\ shot\ on\ target}}{1 - e^{-1.350538+0.0220868*Matches\ Played+0.3547686*Shot\ on\ target\ per\ 90+1.24273*Goals\ per\ shot\ on\ target}}$$

Assessing the model

I do not have enough statistical knowledge to assess the conditions of the multiple logistic regression model. From the empirical logit plots in Figure 17, I can see that the distribution of each predictor individually is roughly linear, so I will assume that the linearity condition for the model is satisfied. I will also assume that all other conditions are satisfied and see how I can use the model. Because of these assumptions, I have to remember that the results should be regarded as tentative guidelines only.

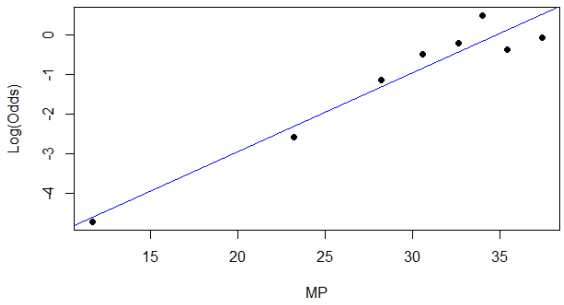

Figure 17a

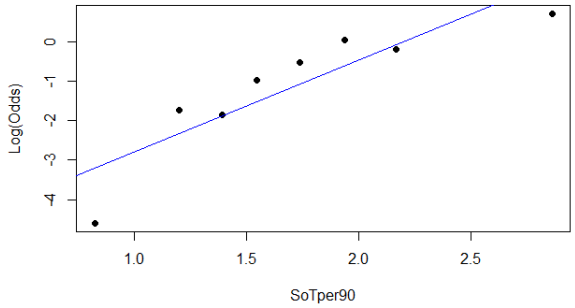

Figure 17b

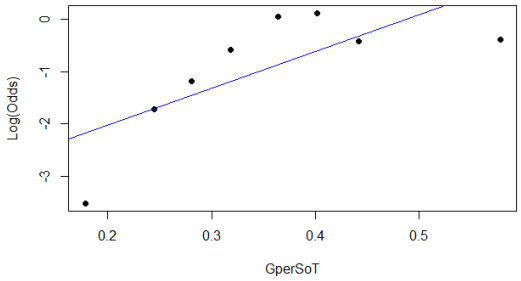

Figure 17c

I will conduct a hypothesis test for the overall effectiveness of the model using G-tests for each single predictor and then using several nested likelihood ratio tests

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_a: At\ least\ one\ \beta_i \neq 0$$

G-tests for all single predictors models yield very low each values, much lower than the significance level of 0.05. Also, every nested likelihood ratio test gave me the p-value of 0 or very close to 0, which gives me strong evidence to reject the null hypothesis. All predictors in my model are useful in predicting whether a striker scores 20 goals in a season.

However, I suspect there may be an interaction between predictors in my model. To check if it actually is the case here, I will change my model to the full interaction model and then conduct the hypothesis test.

Full interaction model in logit form:

$$\log(odds) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_1 * X_2 + \beta_5 * X_1 * X_3 + \beta_6 * X_2 * X_3$$

Hypothesis test

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0$$

$$H_a: At\ least\ one\ \beta_i \neq 0$$

We can see that the p-values for all interaction terms are very low, much lower than the significance level of 0.05. It gives me strong evidence to reject the null hypothesis and say that all interaction terms are significant in my model. I will use the model with interaction to calculate the probability of Lewandowski scoring over 20 goals in the current season, assuming will play in all remaining games of the season and all per 90 minutes statistics will stay on the same level. After using the fitted model, I get

$$\pi(Over\ 20) = 0.7476446$$

Using this model, the probability of Robert Lewandowski scoring over 20 goals in the ongoing season is around 75%. We can see a huge disproportion between the results from models 6 and 7, which may be caused by the fact explained in detail in previous parts, which is that Lewandowski is scoring more goals than he is supposed to from all his shots and the predictors used in model 7 do not account for this difference, while the single predictor in model 6 does. In fact, Lewandowski has already scored 39 goals this season, which means that the more probable outcome form both models has actually happened, and he scored more than 20 goals.

Summary

In my research, I thoroughly analyzed the performance of best European strikers of the last decade and tried to explain the phenomenal year of Robert Lewandowski using the models based on the observations of a bigger sample of players. I gave statistical evidence combined with my own conclusions why he was so successful in 2020 and tried to explain how different factors affect not only his performance, but the performance of every striker. In the last section, I calculated the probability of scoring over 20 goals in the current season assuming Lewandowski keeps performing on the same level. I hope my research helped to explain what constitutes a successful goal scorer in the best European leagues and I hope my work may be used by other sport statisticians in the future.