ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Goals Scored in Europe's Top Leagues: Final Report
Ryan Sanders

For this project, I would like to analyze the number of goals scored in the Bundesliga (Germany), Serie A (Italy), Premier League (UK), and La Liga (Spain). In the soccer world, there is a lot of debate over which league is the "best". The goal of this assignment is to determine which league has scored the most goals over the past few seasons and during which time of year the teams in the respective leagues score most frequently. I am interested in how the time of year and busyness of schedule affects the domestic leagues. Throughout the season, the teams compete in their respective leagues but some of the players are called to play in international competitions for their national teams. So with this in mind I will look at when the international breaks are, as well as when other competitions are going on and how this might affect the leagues performance in terms of goals scored.

To get my data I will go on the sites of the Leagues and look at the results for every team and calculate the total number of goals scored each month. I will then take those numbers and look up the dates for the international breaks and other competitions keep the data from goals scored in each of the months and then the dates of the other competitions in a spreadsheet or chart. I will see which league scored the most goals total, and also which month each league scored the most. Based on this data I will then be able to see how the breaks either increase or decrease the goals scored. It will also tell me which league is "best" by seeing how each league scores goals and the frequency of these goals.

Once I collect this data, I will be able to construct confidence intervals to determine the range of goals to expect for a league of my choosing given the past few seasons. Also, given that the season is almost halfway over for this year, I will be able to predict or determine how many goals will be scored in the second half of this season. With the given size of the data (months in which each season spans for four leagues), I should be able to utilize the standard normal Z table taking into account that I do not know the true population variance. I predict that my data will show me that the winter months are when each league scores the most, and that the beginnings of each season are when the least amount of goals are scored. With all this data in mind, I will be able to determine which league is "best" in terms of goals scored.

The goal of this assignment was to determine which international soccer league is the best or most exciting in terms of goals scored. I looked at the goals per game and goals per month in Europe's top four leagues (the Premier League in England, the Bundesliga in Germany, the Liga BBVA in Spain, and the Serie A in Italy) for the past two seasons, and the current season. In order to come to the conclusion of which league is the best or most exciting, I first tried to collect all the data by hand, going through each of the matches per season, counting the number of goals per game, and totaling those numbers for each month for each league. I quickly realized that this would take forever to do by hand, so I implemented a form of web scraping to find this data. Using Python, I created empty lists for all of the matches per season in each league, and then filtered through those to get the total number of goals and made new lists for each month. It was still a lengthy process, but much more efficient and accurate than if I collected the data by hand. It was difficult to find some of the data for the past seasons for some of the leagues, as the websites no longer had the results from past seasons.

From my proposal, I made some decent predictions. I split up the seasons in three parts: the beginning of the season (August, September, October), the winter months (November, December, January), and the end (February, March, April, May). For the Premier League, I was spot on. The average number of goals scored in the beginning was the least, with 79.67 goals per month, and the winter months had almost 50 goals more per month with 122.33 goals per month. This can be attributed to the international breaks, where players play with their national teams for World Cup qualifying, during the beginning of the season. The winter months have by far the busiest schedules, which leads to the most goals during those months. However, all the other leagues have far less games during the winter months, and more towards the end, and the most goals were scored towards the end of the season. The Bundesliga was the only league that scored the least goals during the winter months. From this data, the international breaks definitely affect the number of goals scored, and this could be attributed to the various playing styles of the club teams versus the national teams, or the fact that the players aren’t in their peak physical condition at the beginning of the season.

I found some interesting results. In terms of total goals scored, the Premier League scored 2376 goals since the beginning of the 2013/14 season. The Liga BBVA had 2343, the Bundesliga scored 2116 (but plays 74 less games per season), and the Serie A 2355 total goals. So in terms of total goals scored, the Premier League scores the most; however, if you calculate the average number of goals per game, the Premier League has the lowest with 2.67 goals per game. The BBVA averaged 2.7 goals per game, the Bundesliga managed 2.96 goals per game, and the Serie A just edges out the BBVA with 2.71 goals per game. This is really interesting, because in the world’s top leagues, there is very little variance between the numbers of goals per game.

The results get even more interesting when you look at the goals per month for each league. Over the past three seasons (2013/14, 2014/15, 2015/16 which is currently in progress), the Premier League averages the most goals per month with 96.65, followed by the Serie A with 93.3 goals, then the Liga BBVA with 92.55 goals per month, and finally the Bundesliga with only 85.83 goals per month. This surprised me because given the previous results, the League with the least goals per game scored the most goals overall, and tallied the most goals per month.

With this in mind, I looked deeper at the Premier League, which just so happened to have the most information readily available. I wanted to examine the total number of goals scored in the past few seasons, the average number of goals scored per month in those seasons. Using that information, I wanted to use linear regression to estimate the number of goals that will be scored in this current season. Then I constructed a 95% confidence interval to give a range with a good amount of certainty.

Collecting data from the last 15 seasons of the Premier League, I calculated the average number of goals scored and used the formulas given in class to implement linear regression. I came out with the formula:

$$Y = 970.13 + 4.325x$$

Which predicts that in the 2015/2016 season, 1039.33 goals will be scored. Using a 99% confidence interval, there will be between 965.27 and 1044.19 goals scored. This data could be inaccurate due to the sample size, as larger samples tend to behave better, and the more samples you have to work with, the more accurate the findings can be.

All this data does make sense according to my findings, and it's more or less what I expected to find. I didn't think that there would be too much difference between the leagues in terms of goals scored. I expected there to be a decent amount of variance between the numbers of goals scored per month due to the business of schedules. It would be nice to have a larger sample size for some of the calculations, but due to the accessibility of the data, it was hard enough to ascertain the data for the goals per month. For the linear regression part of my findings, I can't be as confident as I would like to be with my findings, because if I could get more data for past seasons and use that I could be more certain. Also, if I could use the average number of goals per month for all of those seasons, I could much more accurately predict the number of goals per month, and then per season. It would allow me to better see the general trend of number of goals per month, which would help me produce more accurate findings.

I think this data could be used in a lot of various ways in the real world. The leagues could use this data to promote their League as the "best" in terms of various findings such as goals per game. Broadcasting companies could utilize this to get the more views by showing certain leagues that are more "exciting" during certain months. Teams could use forms of this data to analyze their own players performances based on their opponent, as well as how their own players perform in terms of goals scored and when they produce the best results. This data could be used for a number of editorial purposes, and even gambling. People could bet on the number of goals scored per game or per month or overall using simple calculations such as linear regression.