World Cup and FIFA Rankings: An Econometric Analysis

A

paper submitted

to the

Business Area of Milligan College

as a research project

By:

Luis Esteban de la Torre González

Mentored by:

David A. Campbell, PhD

April, 2018

Table of Contents

Introduction

Soccer is without a doubt the most popular sport in the world. Its regulatory body worldwide is the Federation International Football Association (FIFA). The FIFA is in charge of organizing the FIFA World Cup once every four years. This is the biggest single-event sporting competition in the world (FIFA.com). Since FIFA became the regulatory body of soccer, countries have been ranked based on their performance. These rankings classify countries based on the performance of their national teams through the course of the last four years. These rankings are believed to provide a decently good estimate of how good a country's national team is relative to another. Throughout the years, FIFA Rankings have served as parameters for spectators and experts to reach their own conclusions and try to estimate who the winner of a certain World Cup will be. FIFA Rankings have also helped gamblers take an educated guess when betting for or against a country's national team. This paper explains the relationship of FIFA Rankings and the outcomes of the FIFA World Cup by using an econometric analysis. This paper also takes into consideration other parameters, with the purpose of finding their relationship to certain outcomes of the FIFA World Cup. Although this paper analyzes the relationship between dependent and explanatory variables, it also attempts to find a predictive model for the Russia 2018 FIFA World Cup results, which will take place this coming summer.

Empirical Approach

Before digging into the actual research, it is important to understand the basic concepts discussed in this paper. A FIFA Ranking, as briefed above, is the score given to a country based on its national team's performance over the course of the previous four years. The calculation for a country's FIFA Ranking takes into account two different average numbers of points: the ones gained from matches during the past 12 months, and the ones gained from matches older than 12 months (these depreciate yearly). The total number of points obtained by a country's national team are calculated using the following formula: P = M x I x T x C. In this formula, P stands for total points gained from a specific match; M stands for the points the team gets from the Match result. A team gets 3 points if it wins the match, 1 point if it ties, and 0 points if it loses. I stands for the Importance of the match. For a friendly match, I = 1, for a FIFA World Cup qualifier or confederation-level qualifier, I = 2.5, for a Confederation-level final competition or FIFA Confederations Cup match, I = 3, and for a FIFA World Cup final competition match, I = 4. T stands for the strength of the opposing Team. The strength of an opponent is measured by the following formula: 200 - the ranking position of the opponent. This formula has a couple exceptions to the rule. The strength of the team ranked #1 will always be 200, while teams ranked past 150th receive a strength of 50. The calculation of the strength of each team is made by using the latest posted FIFA Ranking previous to the match. Finally, C stands for the strength of the Confederation. FIFA classifies every country into different confederations. These classifications are made based on the location of each country in the world. According to FIFA, there are six different confederations in the world. AFC stands for Asian Football Federation. CAF stands for Confédération Africaine de Football and it oversees the activities of African soccer. CONCACAF stands for Confederation of North, Central America and Caribbean

Association Football. CONMEBOL stands for Confederación Sudamericana de Fútbol, and it is in charge of South American soccer. OFC stands for Oceania Football Confederation. UEFA stands for Union des Associations Européenes de Football - UEFA, and it oversees soccer in the European Union. The strength of a confederation is calculated based on the number of wins by countries of that confederation in the last three FIFA World Cups. As of now, and until Russia 2018 is over, CONMEBOL strength is 1, UEFA's is .99, and the strength of AFC, CAF, OFC, and CONCACAF is .85. When two teams from different confederations play each other, the value of C is the mean of both confederations' strengths.

Econometric Approach

Now that the empirical explanation of outcomes has been established, the econometric explanation will take place. In order to find a predictive model for the outcomes of the Russia 2018 FIFA World Cup, an econometric analysis of previous FIFA World Cups has been made. The first part of the analysis was of the South Africa 2010 FIFA World Cup. To do so, the data retrieved includes information from the Korea-Japan 2002 FIFA World Cup and the Germany 2006 FIFA World Cup, as well as other information about the countries that participated in the South Africa 2010 FIFA World Cup at the time. The dependent variables on the South Africa 2010 FIFA World Cup analysis are its own outcomes. The second part of the analysis was made by doing the same thing but with the Brazil 2014 World Cup, where the dependent variables included information from the two previous FIFA World Cups, Germany 2006 and South Africa 2010, as well as other facts about the participating countries at the time. The dependent variables for this second part of the analysis were outcomes of the Brazil 2014 FIFA World Cup. Finally, the third part of the analysis was the combination of the explanatory and dependent variables of South Africa 2010 and Brazil 2014 combined into one data set. This data set was the one used to run all of the regressions pertaining this research.

An econometric analysis is based on regressions ran using one dependent variable (y) and one or more explanatory or independent variables (x). After running a regression, each explanatory variable obtains a coefficient, which explains the amount of change in the dependent variable as the explanatory variable changes, all things held constant. An econometric analysis also includes an error term in its formula, this is omitted in practice because of the nature of it. It cannot be calculated. This leads to the assumption that the error term in a model is equal to zero. A normal formula in econometrics takes the form of a linear equation $y = m x + b$. The difference with the application of econometric principles is that the y is only an estimate, it does not provide a 100% accurate value of what the actual value really is. The difference from the actual value y to the estimated value y-hat is called standard error. Another assumption made about the outcomes of a regression is that the standard distribution is normal, this means that the mean of the values is found around the estimated y-hat. Their saturation is assumed to be around it as well. The most common format obtained for a Multiple Linear Regression (MLR) formula is $y = \beta o + \beta_1 x_1 + \beta_2 x_2 + \ldots + u$. In this formula, y represents the estimated y-hat of the model, βo represents the intercept (where the line of fit crosses the y axis), β_1 is the coefficient for x_1, which is an explanatory variable, β_2 is the coefficient for x_2, which is another explanatory variable, and so on. u is the error term.

Data Set

It is important to point out that the predictive model obtained with this research is based on previous and present data. It does not represent the actual outcomes, since these have not taken place yet. In other words, the error term in this model is mostly composed by circumstantial outcomes, such as injuries, cards obtained by players, weather, personal lives of the players, etc. In this particular research, the predictive model for the Russia 2018 FIFA World Cup was obtained by having different dependent variables and using them to explain different aspects of the outcomes. This is with the purpose of getting a model that turns out to be as predictive as possible, minimizing the place for errors as much as possible. Each one of these dependent variables were analyzed individually and then collectively. This model, among its dependent and independent variables, includes binary or dummy variables. These take the value of 1 if yes or the value of 0 if no. The dependent variables used for the estimation of the model were: Points out of group stage, Difference of goals, Games won, Round of 16 (dummy), Quarter finals (dummy), Semifinals (dummy), Final (dummy), and Won (dummy).

On the other hand, this model's explanatory variables are compounded by data from previous FIFA World Cups and data that is not explained by FIFA World Cups. Of the explanatory variables in this model, those related to previous FIFA World Cups outcomes are: Games won 2 World Cups ago, Games won previous World Cup, Games tied 2 World Cups ago, Games tied previous World Cup, Difference of goals 2 World Cups ago, Difference of goals previous World Cup, World Cups won, Host (dummy), and Previous World Cup appearances (sq.). Something to take into account when choosing these variables was to avoid perfect multicollinearity. This is why games won and games tied are included, but games lost are omitted. Another thing to point out is that instead of using goals scored and goals against, this model uses the difference of goals, which is an addition of goals scored and goals against. Finally, it should be noted that the variable of previous World Cup appearances is squared. Of the explanatory variables in the model, those that do not relate to previous World Cups yet help analyze their outcomes better are: Professional teams in the country's first league, Average age of players in the roster, FIFA Ranking, FIFA Ranking (sq.), Population (log), and GDP per capita (log). Out of these variables there are a couple other things to point out as well. The FIFA Ranking variable includes its original form as well as its squared form, while the Population and GDP per capita variables only include the logarithmic version of them.

Research

As mentioned before, different experimentations were done to find the most suitable model overall while researching. The table below explains what the outcomes were when running Points out of group stage as the dependent variable.

	Points out of group stage		
	Regression 1	Regression 2	Regression 3
R square	0.669	0.446	0.585
Adjusted R square	0.557	0.288	0.571
Intercept	0.022 (5.993)	-6.441 (7.341)	7.274 * (0.438)
games won 2 World Cups ago	0.004 (0.308)	-0.117 (0.388)	-
games won previous World Cup	-0.129 * (0.237)	-0.186 * (0.299)	-
games tied 2 World Cups ago	-0.285 (0.394)	-0.617 (0.492)	-
games tied previous World Cup	-0.671 (0.351)	-0.957 (0.441)	-
difference of goals 2 World Cups ago	-0.029 (0.143)	0.149 (0.175)	-
difference of goals previous World Cup	-0.017 (0.113)	0.119 (0.140)	-
world cups won	-0.265 (0.479)	-0.438 (0.605)	-
Host	2.316 * (1.367)	1.798 (1.728)	-
previous world cup appearances	0.146 (0.226)	0.269 (0.285)	-
Previous world cup appearances sq.	-0.005 (0.016)	-0.004 (0.021)	-

professional teams in country´s first league	0.075 (0.091)	0.290 * (0.105)	-
average age of players	0.174 (0.223)	0.127 (0.275)	-
FIFA Ranking	-0.185 * (0.040)	-	-0.172 * (0.026)
FIFA Ranking sq.	0.001 * (0.0003)	-	0.001 * (0.0003)
Pop log	-0.044 (0.287)	-0.798 * (0.321)	-
GDP log	0.194 (0.222)	0.489 * (0.271)	-

*Significant at the 90% level. (t-stat 1.645 of higher)

The table above possesses a few things to point out. The first regression ran was using all of the 16 variables, the second one was without using either FIFA Ranking or FIFA Ranking (sq.), and the third regression was just using the FIFA Ranking and the FIFA Ranking square. Although the first regression possesses 16 variables, only 4 of them are statistically significant at the 90% confidence level - Games won last World Cup, Host (dummy), FIFA Ranking, and FIFA Ranking (sq.). This means that the estimated value provided by these variables will be equal to the actual value 9 out of 10 times the model is run, everything else held constant. On the second regression, without FIFA Rankings, 4 out of the 14 variables are statistically significant at the 90% confidence level - Games won on the previous World Cup, Professional teams in country’s first league, Population (log), and GDP per capita (log). Finally, on the third regression, which only uses the FIFA Ranking and FIFA Ranking (sq.), both of the explanatory variables are significant at the 90% confidence level, as well as the intercept. One more thing to note about these regressions is their goodness of fit. The first regression has the highest R-square out of all of them, which is why it is the one used to run the predictive model. Even the R-square of the first regression (.669) is higher than the one of the other two, the Adjusted R-square of the third regression (.571) is slightly higher than the one of the first (.557).

The table below shows a comparison between regressions ran to find the model that best explains the dependent variable Points out of the group stage. This same type of analysis was made for all of the other variables as well. These analyses led to the observation that the best model for the dependent variables Difference of goals and Games won were obtained by using the exact same explanatory variables used for Points out of group stage. After running the above-mentioned regressions, the data was gathered and the outcomes were as follows:

	Points out of group stage	**Goal difference**	**Games won**
R square	0.6695	0.6306	0.6824
Adjusted R square	0.5570	0.5048	0.5742
SSE	282.35	755.45	117.37
SSR	139.39	442.55	54.63

As the table above shows, the regression that possesses the most explanatory power out of the three is the one that has Games won as its dependent variable, with an R-square of .6824 and an adjusted R-square of .5742. The next one is the one that has Points out of the group stage as its dependent variable and finally the one with Goal difference. Something else to pay attention to on these regressions are the sums of squares. The Games won regression has the lowest SST, as Goal difference has the highest. This shows that as the model attempts to explain more, the relationship between SSR and SSE will provide a lower R-square.

Finally, after finding the formulas for the “normal” dependent variables, the ones for the dummy dependent variables needed to be found. This was done with the aid of already-known dependent variables. The format of the FIFA World Cup allows 16 out of the 32 teams that participate to advance past the group stage, 8 to the quarter finals, 4 to the semifinals, 2 to the finals, and only one wins. This means that not all of the teams get to play the same amount of games. Once a team did not make it past a stage, it is out. Since this is a factor in the FIFA World Cup, the model needed to account for it. The way that the model accounted for this format was by finding what teams will make it past the group stage and including the results in the regression for the teams that will make it, dropping the least significant dependent variables from each one of them. These variables also turned out to have extremely low coefficients. By doing this, the explanatory models for the dummy variables become more significant and explanatory, boosting their R-squares and the significance of their variables. The same process was followed with all of the dummy variables. The following tables explain the process more graphically:

Round of 16		Quarter finals	
Drop: goal diff. 8 years ago	Add: points out of group stage	Drop: professional teams	Add: round of 16
0.00147 (0.03656)	0.18093* (0.02623)	-0.00002 (0.01705)	0.31134* (0.09671)

*Significant at the 99% level. (t-stat 2.576 of higher)

Semifinals		Final	
Drop: pop (log)	Add: quarter finals	Drop: GDP (log)	Add: Semifinals
0.00407 (0.04418)	0.30526* (0.11149)	0.00122 (0.02906)	0.53928* (0.09597)

*Significant at the 99% level. (t-stat 2.576 of higher)

Win		
Drop: FIFA Ranking	Drop: FIFA Ranking sq.	Add: Final
-0.00251 (0.00376)	0.00002 (0.00003)	0.45325* (0.07796)

*Significant at the 99% level. (t-stat 2.576 of higher)

As the tables above show, the explanatory variables added to each one of these models are significant at the 99% confidence level, while the ones dropped have extremely low t-stats and coefficients, which makes them virtually worthless if left included in the model.

Results

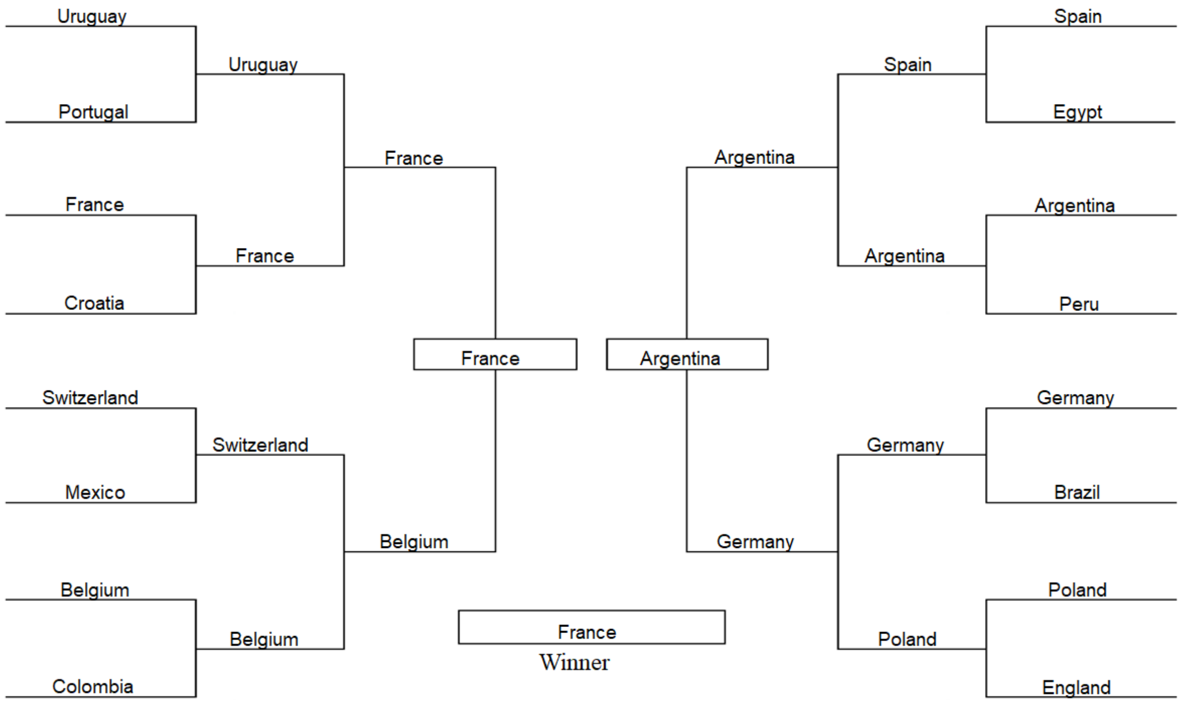
As it has been established in this paper, the ultimate purpose of this research was to find a predictive model for the Russia 2018 FIFA World Cup using an econometric analysis. After running plenty of regressions and filling in the data for the explanatory variables pertaining Russia 2018, predictions using the model were achieved. Here is a broad overview of them:

	High	Low
Points out of group stage	Spain (8)	Nigeria (0)
Goal difference	Germany (7)	Russia (-11)
Games won	Belgium & Germany (4)	Panama (-1)

The model obtained after the research predicted which teams will advance to the round of 16, as well as their place within their group. The results are as follows:

	Group A	Group B	Group C	Group D	Group E	Group F	Group G	Group H
1st	Uruguay	Spain	France	Argentina	Switzerland	Germany	Belgium	Poland
2nd	Egypt	Portugal	Peru	Croatia	Brazil	Mexico	England	Colombia

Finally, the last predictions consisted on analyzing the percentage chance that each team had of advancing to the next round. Given these results, the prediction for the final bracket and winner of the Russia 2018 FIFA World Cup are as follows:


Uruguay
Portugal
France
Croatia
Switzerland
Mexico
Belgium
Colombia
Uruguay
France
Switzerland
Belgium
France
Belgium
France
Argentina
France
Winner
Argentina
Germany
Spain
Argentina
Germany
Poland
Spain
Egypt
Argentina
Peru
Germany
Brazil
Poland
England

References:

Bleacher Report. (n.d.). Retrieved from http://bleacherreport.com/

ChartsBin. (n.d.). FIFA World Cup Teams by Number of Appearances. Retrieved from http://chartsbin.com/view/q5y

Dyte, D., & Clarke, S. (2000). A Ratings Based Poisson Model for World Cup Soccer Simulation. *The Journal of the Operational Research Society, 51*(8), 993-998. doi:10.2307/254054

Enikolopov, Ruben, Maria Petrova, and Konstantin Sonin. "Social Media and Corruption." *American Economic Journal: Applied Economics*10, no. 1 (January 2018): 150-74.

FIFA World Cup 2014 Brazil teams by average player age | Statistic. (n.d.). Retrieved from https://www.statista.com/statistics/303661/fifa-world-cup-2014-brazil-teams-by-average-player-age/

F. (n.d.). Fédération Internationale de Football Association (FIFA). Retrieved from http://www.fifa.com/

Lago, C. (2007). Are winners different from losers? Performance and chance in the FIFA World Cup Germany 2006. *International Journal of Performance Analysis in Sport*, *7*(2), 36-47.

Nigeria Population (LIVE). (n.d.). Retrieved from http://www.worldometers.info/world-population/nigeria-population/

Ridder, G., Cramer, J., & Hopstaken, P. (1994). Down to Ten: Estimating the Effect of a Red Card in Soccer. *Journal of the American Statistical Association, 89*(427), 1124-1127. doi:10.2307/2290942

Sylla, M. D. M. (2016). Prediction of the World Cup Soccer Winner: Using Two Statistical Methods.

WINSTON, W. (2009). RATING SPORTS TEAMS. In *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football* (pp. 266-282). Princeton University Press. Retrieved from http://www.jstor.org/stable/j.ctt7sj9q.45

World Bank Group - International Development, Poverty, & Sustainability. (n.d.). Retrieved from http://www.worldbank.org/

World Cup - Champions. (n.d.). Retrieved from http://www.worldfootball.net/winner/wm/