Open Access Library Journal
2019, Volume 6, e5264
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Regressive Prediction Is the Best Way to Forecast Sports Outcomes: Evidence from Brazilian Soccer

Murilo Silva, Sergio Da Silva*

Graduate Program in Economics, Federal University of Santa Catarina, Florianopolis, Brazil
Email: *professorsergiodasilva@gmail.com

How to cite this paper: Silva, M. and Da Silva, S. (2019) Regressive Prediction Is the Best Way to Forecast Sports Outcomes: Evidence from Brazilian Soccer. *Open Access Library Journal*, **6**: e5264.
https://doi.org/10.4236/oalib.1105264

Received: February 19, 2019
Accepted: March 3, 2019
Published: March 6, 2019

Abstract

We illustrate through a case study that regressive prediction is the best method to forecast sports outcomes. By taking predictions of promotion to first division soccer from a mathematician from one of the most famous sports websites in Brazil, we show that making Bayesian updates is misleading when we expect regression to the mean. The expert failed to realize that the more extreme the results are, the more regression is expected, because extremely good scores suggest very lucky days.

Subject Areas

Behavioral Economics, Mathematical Analysis

Keywords

Soccer, Sports Outcomes Forecast, Regression to the Mean, Regressive Prediction, Luck

1. Introduction

Sports competitions are an attractive field of study. A sports league forms a relatively isolated system, with few external and unrestrained influences. Besides, it is replicated over time almost in the same conditions and under the same rules. Also, the large amount of data available makes it possible to learn their statistical patterns, not to mention their popularity, which fuels a multibillion-dollar business that includes television, advertisements and a huge betting market [1].

For those reasons, it makes sense that sports media asks mathematicians and statisticians to predict tournament outcomes. In Brazil, this occurs especially in soccer and it becomes a topic of discussion among fans and experts. This work

takes one of those forecasts and investigates whether it uses regressive prediction. We show that it does not use, and this leads to incorrect outcomes.

Section 2 presents the predictions used in the study; Section 3 discusses them and Section 4 concludes this report.

2. Analysis of Predictions

Soccer is the most popular sport in Brazil and in the world [2]. Thus, match prediction has become interesting for several groups of people, including fans, experts, bettors and different types of media. In order to achieve a good forecast, it is necessary to take into account a large number of factors that affect results, such as the quality of players, injuries, cards, goal scores and luck. Even for experts in the field, like reporters, commentators and coaches, predicting a game outcome is a complex exercise.

The website https://globoesporte.globo.com/, one of the most-followed sports websites in Brazil, published a series of articles with probabilities of promotion for the soccer clubs disputing the 2018 Brazilian second division (called *Série B*) with odds calculated by mathematician T.G.. He is a specialist in soccer probabilities and his website contains predictions for the major Brazilian tournaments. The first online article was published with four games remaining for each club, the second after the 36th round and the last one before the 38th round. We show the *Série B* standings for the top 10 clubs in the last four rounds and the promotion probabilities assigned to each club in Table 1.

There are 20 clubs in *Série B*. During the course of a season (from April to December), each club plays the others twice (a double round-robin system), once at its home stadium and once at its opponents, for 38 games. Clubs receive three points for a win and one point for a draw. No points are awarded for a loss. The top four clubs are promoted to Brazil's first division (*Série A*).

After 34 games played, Fortaleza clinched a promotion, while eight clubs were

Table 1. Top 10 *Série B* standings and promotion odds.

34th round	Pts	Prob. (%)	35th round	Pts	36th round	Pts	Prob. (%)	37th round	Pts	Prob. (%)	Final round	Pts
1. Fortaleza*	64	100	1. Fortaleza*	65	1. Fortaleza*	68	100	1. Fortaleza*	71	100	1. Fortaleza*	71
2. CSA	57	91	2. CSA	58	2. CSA	59	90	2. Goiás*	60	100	2. CSA*	62
3. Avaí	56	81	3. Goiás	57	3. Goiás	57	70	3. Avaí	60	90	3. Avaí*	61
4. Goiás	54	64	4. Avaí	57	4. Avaí	57	56	4. Ponte Preta	59	65	4. Goiás*	60
5. Vila Nova	52	20	5. Londrina	54	5. Ponte Preta	56	44	5. CSA	59	45	5. Ponte Preta	60
6. Londrina	51	18	6. Ponte Preta	53	6. Londrina	55	25	6. Atlético/GO	56	1	6. Atlético/GO	59
7. Atlético/GO	51	18	7. Atlético/GO	52	7. Vila Nova	55	13	7. Vila Nova	56	0	7. Vila Nova	57
8. Ponte Preta	50	6	8. Vila Nova	52	8. Atlético/GO	53	2	8. Londrina	55	0	8. Londrina	55
9. Guarani	49	4	9. Guarani	50	9. Guarani	50	0	9. Guarani	51	0	9. Guarani	54
10. Coritiba	46	-	10. São Bento	46	10. Coritiba	49	-	10. Coritiba	49	-	10. Coritiba	52

Notes: *Club achieved promotion.

still in the race for the three remaining spots. The other three clubs completing the top four were given the best chance to qualify for first division soccer, namely CSA (91%), Avaí (81%) and Goiás (64%). The five remaining clubs were given relatively low odds: Vila Nova (20%), Londrina (18%), Atlético-GO (18%), Ponte Preta (6%) and Guarani (4%).

Only after the 36th round, another online article with new probabilities was released. One team dropped out from battle and one gained relevance, Guarani and Ponte Preta, respectively. The latter ascended to a 44% chance of promotion coming from two wins in a row. But the three teams with the most chances were still the ones among the top four, CSA (90%), Goiás (70%) and Avaí (56%).

Prior to the last round, a new online article was disclosed. There were only two spots left because Goiás clinched the promotion. For the first time, there was a change in the ranking of clubs with the most chances of qualification—CSA (45%) gave its place to Ponte Preta (65%). Avaí continued with a high probability (90%) and Atlético-GO had a 1% chance.

In the end, the teams that achieved promotion were the ones that were in the top four remaining four rounds: Fortaleza, CSA, Avaí and Goiás (first column in Table 1). In other words, the first prediction was right, but the mathematician kept making updates in a manner such that the last one ended up wrong. Although he does not reveal the model used to make forecasts, we can infer that he was doing non-regressive Bayesian updates. The mathematician was altering the probabilities of promotion to each club based on their positions in previous rounds. That is why Ponte Preta was given more chances each round—they won three games in a row prior to the last round, while CSA had their chances reduced because they drew two games and lost one before the final round. However, when we expect regression to the mean, doing non-regressive Bayesian updates is misleading.

3. Discussion

As pointed out by Daniel Kahneman [3], regression to the mean was discovered and named late in the 19th century by Sir Francis Galton in an article published in 1886 under the title "Regression towards Mediocrity in Hereditary Stature", which reports measurements of size in successive generations of seeds and in comparisons of the height of children to the height of their parents.

Wainer [4] warns that Kelley's equation, which indicates that the truth is estimated best when its observed value is regressed toward the mean of the group that it came from, is one of the most dangerous equations to ignore. Kahneman [3] argues that regression effects can be found wherever we look, but we do not recognize them for what they are. According to him, the main reason for the difficulty is that our mind is strongly biased toward causal explanations and does not deal well with "mere statistics". When our attention is called to an event, associative memory will look for its cause—more precisely, activation will automatically spread to any cause that is already stored in memory. Causal explana-

tions will be evoked when regression is detected, but they will be wrong because the truth is that regression to the mean has an explanation but does not have a cause.

To support this view, Kahneman [3] shares one of his favorite equations: success = talent + luck. To demonstrate the idea that luck often contributes to success, the author applies it to the first two days of a high-level golf tournament. He focuses on a player who did very well on the first day. An immediate inference is that the golfer is more talented than the average participant in the tournament. The formula for success suggests that another inference is equally justified: the golfer who did well on Day 1 probably enjoyed better-than-average luck on that day. If you accept that talent and luck both contribute to success, the conclusion that the successful golfer was lucky is as warranted as the conclusion that he is talented.

Suppose you are asked to predict his score on Day 2. You expect the golfer to retain the same level of talent on the second day, so your best guess will be "above average". Luck, of course, is a different matter. Since you have no way of predicting the golfers' luck on the second (or any) day, your best guess must be that it will be average, neither good nor bad. This means that in the absence of any other information, your best guess about the players' score on Day 2 should not be a repeat of their performance on Day 1. The golfer who did well on Day 1 is likely to be successful on Day 2 as well, but less so than on the first day, because the unusual luck he probably enjoyed on Day 1 is unlikely to hold.

Kahneman [3] tells that his students were always surprised to hear that the best predicted performance on Day 2 is more moderate, closer to the average than the evidence on which it is based (the score on Day 1). He states that this is why the pattern is called regression to the mean. The more extreme the original score, the more regression is expected, because an extremely good score suggests a very lucky day. The regressive prediction is reasonable, but its accuracy is not guaranteed. A few of the golfers who did well on Day 1 will do even better on the second day, if their luck improves. But most will do worse, because their luck will no longer be above average.

We can apply the theory of regression to the mean to the subject of our study. Based on the standings until the 34th round, we can say CSA was more talented than Ponte Preta and the luck of both teams was distributed evenly through these games. Now, looking into the following three rounds, Ponte Preta won them all, whereas CSA drew two and lost one game, which made it reach the final round tied in points. Suppose you are asked to predict which club, between these two (Ponte Preta and CSA), is going to achieve promotion to first division in the last round. Which one do you choose?

Implementing regressive prediction, the obvious choice would be CSA. Due to random fluctuation in the quality of performance of Ponte Preta, three wins in a row, a pattern not seen before by the club in the tournament, we can say the team was experiencing above average luck. At some point, luck runs out. As observed, the more extreme the results, the more regression we expect, because

those suggest very lucky days. So, we could expect a different outcome for Ponte Preta's last game, probably worse—something that became a fact. The team drew its last game.

We can exercise the opposite reasoning with CSA. They had three bad results in sequence, a sign of below average luck. The bad luck was not going to hold for a very long time, and the best guess was a better performance in the last round, which happened. They won and got the promotion to *Série A*. Over time, low scores on the first occasions will, on average, improve, whereas high scores will decline.

Pluchino and coauthors [5] observe that luck is almost always underestimated by successful people. This happens because randomness often plays out in subtle ways, therefore it is easy to construct narratives that portray success as having been inevitable. A tendency called "narrative fallacy" by Nassim Taleb [6]. Kahneman and Tversky [7] claim that individuals overemphasize new information. It seems that this happened to the mathematician who authored the articles in https://globoesporte.globo.com/. He gave too much weight to results in each of the final rounds and too little weight to the rest of the tournament. He failed to consider those outcomes as extreme.

Kahneman and Tversky [7] assert that any significant activity of forecasting involves a large component of judgment, intuition and educated guesswork. Opinions and intuitions play an important part even where the forecasts are obtained by a mathematical model or simulation. Intuitive judgments enter in the choice of variables that are considered in such models, the impact factors that are assigned to them, and the initial values that are assumed to hold.

Kahneman and Tversky [7] state that one of the basic principles of statistical prediction, which is also one of the least intuitive, is that the extremeness of predictions must be moderated by considerations of predictability. In intermediate situations, which are the most common, the prediction should be regressive. That is, it should fall between the class average and the value that best represents one's impression of the case at hand. The lower the predictability, the closer the prediction should be to the class average. Intuitive predictions tend to be non-regressive as people often make extreme predictions on the basis of information whose reliability and predictive validity are known to be low.

Kahneman and Tversky [7] go on to affirm that the rationale for regressive prediction is most clearly seen in the prediction of the result of a repeated performance or a replication, like in our soccer example. The laws of chance entail that a very high score on the first observation is likely to be followed by a somewhat lower score on the second, while a poor score on the first observation is likely to be followed by a higher score on the second. This phenomenon is a mathematical consequence of the presence of uncertainty. Therefore, the best prediction for a repeated performance of aclub is less extreme than the initial score. Intuitive predictions violate this principle, though. People usually make predictions as if measures of performance were equally probable to change toward the average and away from it.

Kahneman [3] argues that intuitive predictions need to be corrected because they are not regressive and therefore are biased. You still make errors when your predictions are unbiased, but the errors are smaller and do not favor either high or low outcomes. The expert could point out, correctly, that this procedure will usually produce conservative predictions that are not far from the average of the class, and is very unlikely to predict an exceptional outcome. Kahneman and Tversky's [7] answer to this objection is that a fallible predictor can retain a chance to correctly predict a few exceptional outcomes only at the cost of incorrectly identifying many other cases as exceptional. In most situations, this bias is costly, and should be eliminated.

As observed, the model supposedly developed by mathematician T.G. to predict the chances of each club to achieve promotion had a Bayesian touch. Gelman [8] addresses the subjective component of Bayesian inference. In his opinion, people tend to believe in results that support their presumptions and disbelieve results that surprise them. He says that Bayesian methods encourage this undisciplined mode of thinking and motivate even the best-intentioned researchers to get stuck in a rut of prior beliefs. Kahneman [3] also draws attention to the fact that intuitive impressions of the diagnosticity of evidence in Bayesian reasoning are often exaggerated.

4. Conclusions

We analyzed the predictions made by a mathematician about the probabilities of club promotion to first division soccer in Brazil. Instead of performing regressive predictions, he made Bayesian updates, which led him to a misguided forecast in the last round of the tournament. The specialist failed to notice that in the presence of uncertainty, the best prediction for a repeated performance of a club is less extreme than the initial score.

Extreme good results suggest lucky days; therefore, regression to the mean is expected because the unusual luck is unlikely to continue. Of course the accuracy of regressive prediction is not guaranteed, but at least it is reasonable. You still incur errors when your forecasts are unbiased, but those are reduced. Other procedures favor exceptional outcomes at the cost of mistakenly missing most cases. We do not recognize regression effects for what they are. Even a mathematician can end up neglecting a statistical phenomenon.

Acknowledgements

Financial support from CNPq and Capes is acknowledged.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Aoki, R.Y.S., Assuncao, R.M. and De Melo, P.O.S.V. (2017) Luck Is Hard to Beat:

The Difficulty of Sports Prediction. *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, 13-17 August 2017, 1367-1376. https://doi.org/10.1145/3097983.3098045

[2] Da Silva, S., Matsushita, R. and Silveira, E. (2013) Hidden Power Law Patterns in the Top European Football Leagues. *Physica A*: *Statistical Mechanics and Its Applications*, **392**, 5376-5386. https://doi.org/10.1016/j.physa.2013.07.008

[3] Kahneman, D. (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

[4] Wainer, H. (2007) The Most Dangerous Equation. *American Scientist*, **95**, 249-256. https://doi.org/10.1511/2007.65.1026

[5] Pluchino, A., Biondo, A.E. and Rapisarda, A. (2018) Talent versus Luck: The Role of Randomness in Success and Failure. *Advances in Complex Systems*, **21**, 1-31. https://doi.org/10.1142/S0219525918500145

[6] Taleb, N.N. (2007) The Black Swan: The Impact of the Highly Improbable. Random House, New York.

[7] Kahneman, D. and Tversky, A. (1982) Intuitive Prediction: Biases and Corrective Procedures. In: Kahneman, D., Slovic, P. and Tversky, A., Eds., *Judgment under Uncertainty*: *Heuristics and Biases*, Cambridge University Press, Cambridge, 414-421. https://doi.org/10.1017/CBO9780511809477.031

[8] Gelman, A. (2008) Objections to Bayesian Statistics. *Bayesian Analysis*, **3**, 445-449. https://doi.org/10.1214/08-BA318