Open Access Library Journal
2019, Volume 6, e5603
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Role of Regression to the Mean and Loss Aversion in Brazilian Soccer Club Performance

Leon Esquierro, Sergio Da Silva*

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

How to cite this paper: Esquierro, L. and Da Silva, S. (2019) Role of Regression to the Mean and Loss Aversion in Brazilian Soccer Club Performance. *Open Access Library Journal*, **6**: e5603.
https://doi.org/10.4236/oalib.1105603

Received: July 15, 2019
Accepted: July 28, 2019
Published: August 2, 2019

Abstract

Causal explanations are often favored to explain club performance in soccer tournaments, and the role that luck plays is usually neglected. Here, we consider three recent seasons of the first-division Brazilian soccer league to examine the relative importance to club success of loss aversion (a causal explanation) and regression to the mean (luck). We find that club performance depends on both, and quantify this finding.

Subject Areas

Behavioral Economics

Keywords

Soccer, Regression to the Mean, Loss Aversion, Luck

1. Introduction

Fluctuation in club performance is common to occur in the course of soccer tournaments. Both pundits and fans usually explain such a phenomenon relying exclusively on causal explanations, such as: "injuries impaired team A;" "refereeing was biased toward team B;" "team C chickened out in the final rounds;" "a change in management negatively affected the environment in club D;" and so on. A narrative is seemingly always available to account for success or failure. However, this is at odds with the story of success in sports and elsewhere [1] [2] [3].

Causal explanations often ignore a key aspect of reality: random fluctuations in performance, that is, plain luck [2]. Narratives and causal explanations are the default mode of our automatic mind (a.k.a. System 1). Moreover, System 1 has difficulties in recognizing statistical phenomena, such as regression to the mean

[4] [5] [6]. In predicting results of the second-division Brazilian soccer league, it has been argued that regressive predictions are more accurate than non-regressive ones [7].

System 1's insistence on causal explanations is due to both: 1) the analytical mind (a.k.a. System 2)'s difficulties in understanding the nuances of the notion of randomness, and 2) the cognitive ease brought about by simple causal explanations that give an air of inevitability to past events [2]. This has been called the "narrative fallacy" [8] [9].

Here, we consider data from the three more recent seasons of the first-division Brazilian soccer league, called *Serie A*. Section 2 considers the issue of regression to the mean. First, we evaluate whether club performance in the first half-season of 2016, 2017 and 2018 is predictive of performance in the second half-season. We ignore earlier seasons because two extra clubs could join *Copa Libertadores* qualifiers from 2016 onward, thus altering the structure of incentives the teams faced. Regression to the mean is detected if the clubs that are doing well in the first half-season end up performing relatively poorly in the second half-season.

Section 3 investigates whether a causal explanation is also at play—namely, loss aversion. In this case, part of regression to the mean could be genuinely addressed to a causal explanation. In sports, loss aversion has been shown to occur in golf tournaments, where professional golfers putt more accurately for par than for a birdie [2]. By analogy, here we focus on the performance of clubs fighting relegation whenever they play against those looking for promotion to *Copa Libertadores*. Loss aversion is detected if the clubs that are struggling to escape the relegation zone perform relatively better than the clubs that are aiming at promotion to *Copa Libertadores*. Loss aversion means the underdogs have more to lose than the favorites.

Section 4 discusses which one is more important—regression to the mean or loss aversion—and Section 5 concludes this report.

2. The Hypothesis of Regression to the Mean

First, we test whether clubs with points won above the median in the first half-season tend to score relatively fewer points in the second, whereas clubs that score points below the median in the first half-season tend to end up with relatively more points in the second. We consider the median rather than the mean because our data have outliers. Indeed, when sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency; and we use the median to describe the middle of a set of data that does have an outlier.

Using ordinary least squares regressions for each of the three seasons, we take as an independent variable the deviations from the median of the points won by a club in the first half-season. The dependent variable is the deviations from the median of the points won by the same club in the second half-season relative to those won in the first half-season. Here, we cannot dismiss the hypothesis of regression to the median if the angular coefficient of the estimated regression line

is negative.

We first consider a sample of the 20 clubs that take part in *Serie A* for each of the three seasons. Then, we drop outliers, that is, the club that lost more points in the second half-season as compared to their performance in the first, as well the club that won more points. Doing so, we assess whether results are not being affected by extreme data points. We also check for the robustness of results by running the regressions without the linear coefficient. Lastly, we repeat our analysis by considering the data pooled for the three seasons.

2.1. The 2016 Season

This season was the most idiosyncratic of all. The beginning (until the eighth matchday) was marked by the leadership of two teams that ended up relegated (Santa Cruz and Internacional). Besides, overall club performance was superior in the second half-season (522 versus 515), which means an expected positive intercept for the regression. Table 1 shows the final standings in alphabetical order.

Table 2 shows that the results displayed the expected signs for both estimated coefficients. However, these were not significant even at the 10 percent level. This circumstance will not happen in the subsequent two seasons, as we will see.

Table 1. Final standings in the 2016 season of the Brazilian soccer league *Serie A* in alphabetic order.

Club	Points won in the first half-season (P1)	Deviation from the median (24.5)	Points won in the second half-season (P2)	P2-P1
América MG	13	−11.5	15	2
Atlético MG	35	10.5	27	−8
AtléticoPR	30	5.5	27	−3
Botafogo	20	−4.5	36	16
Chapecoense	24	−0.5	28	4
Corinthians	34	9.5	21	−13
Coritiba	21	−3.5	25	4
Cruzeiro	19	−5.5	32	13
Figueirense	21	−3.5	16	-5
Flamengo	34	9.5	37	3
Fluminense	25	0.5	22	−3
Grêmio	32	7.5	21	−11
Internacional	22	−2.5	21	−1
Palmeiras	36	11.5	44	8
Ponte Preta	27	2.5	26	−1
Santa Cruz	18	−6.5	13	−5
Santos	33	8.5	38	5
São Paulo	26	1.5	26	0
Sport	23	−1.5	24	1
Vitória	22	−2.5	23	1

Table 2. Regression with intercept for all 20 clubs in the 2016 season.

	Coefficient	Standard error	*t*-statistic	*p*-value
Intercept	0.79	1.59	0.49	0.62
Deviation from the median	−0.35	0.24	−1.44	0.16

Moreover, an R squared of 0.104 was the lowest of the three seasons. Alas, ignoring the intercept does not change results a great deal. However, results do improve after dropping from analysis of the two outliers Santa Cruz and Internacional because the angular coefficient turns significant at 10 percent (Table 3). All in all, we cannot dismiss regression to the median in the 2016 season. Yet this phenomenon will be more pronounced in the subsequent seasons, as we will see next.

2.2. The 2017 Season

This season was characterized by an outstanding performance by champions Corinthians. However, this club scored 47 points in the first half-season, but only 25 in the second (Table 4). This strong regression to the median began in the 13th round and was anticipated by Grêmio coach Renato Portaluppi, who seems to have had a glimpse of the phenomenon. In turn, Atlético GO, which ended up at the bottom of the table, also experienced strong regression to the median.

Table 5 shows the expected sign for the angular coefficient, which was significant at 1 percent. However, the linear coefficient was not significant, though the R squared was 0.71. Thus, we cannot dismiss the hypothesis of regression to the median. Also, dropping the constant from the regression does not alter results much.

The results in Table 6 suggest those in Table 5 are unlikely to be explained by outliers. Indeed, dropping Corinthians and Atlético GO does not change results too much, apart from a reduced R squared of 0.54.

2.3. The 2018 Season

Regression to the median also seems to have occurred in this season. São Paulo's excellent performance in the first half-season (41 points won) was followed by a very disappointing second half-season (22 points) (Table 7). In contrast, champions Palmeiras came from behind and ended up scoring incredible 47 points in the second half-season. At the bottom of the table, Ceará and AtléticoPR also performed better in the second half-season.

Table 8 shows a negative angular coefficient for the variable "regression to the mean" (−0.52), which was significant at the 5 percent level. Thus, despite a low R squared of 0.25, we cannot dismiss regression to the median.

A large *p*-value for the intercept in Table 8 may have been caused by the peculiar fact that the sum of points won by all the clubs in the first half-season exactly matched the points of the second, that is, 515. This means we should expect a zero value for the intercept, a hypothesis that could not be rejected. Indeed,

Table 3. Regression with intercept for 18 clubs in the 2016 season (apart from outliers Santa Cruz and Internacional).

	Coefficient	Standard error	*t*-statistic	*p*-value
Intercept	1.59	1.72	0.91	0.37
Deviation from the median	−0.46	0.25	−1.77	0.094

Table 4. Final standings in the 2017 season of the Brazilian soccer league *Serie A* in alphabetic order.

Club	Points won in the first half-season (P1)	Deviation from the median (25)	Points won in the second half-season (P2)	P2-P1
Atlético GO	12	−13	24	12
Atlético MG	23	−2	31	8
Atlético PR	26	1	25	−1
Avaí	18	−7	25	7
Bahia	23	−2	27	4
Botafogo	25	0	28	3
Chapecoense	22	−3	32	10
Corinthians	47	22	25	−22
Coritiba	25	0	18	−7
Cruzeiro	27	2	30	3
Flamengo	29	4	27	−2
Fluminense	25	0	21	−4
Grêmio	39	14	23	−16
Palmeiras	32	7	31	−1
Ponte Preta	22	−3	16	−6
Santos	35	10	28	−7
São Paulo	19	−6	31	12
Sport	28	3	17	−11
Vasco	24	−1	32	8
Vitória	19	−6	24	5

Table 5. Regression with intercept for all 20 clubs in the 2017 season.

	Coefficient	Standard error	*t*-statistic	*p*-value
Intercept	0.75	1.14	0.66	0.51
Deviation from the median	−1.00	0.14	−6.77	2.38E−06

Table 6. Regression with intercept for 18 clubs in the 2017 season (apart from outliers Corinthians and AtléticoGO).

	Coefficient	Standard error	*t*-statistic	*p*-value
Intercept	0.91	1.26	0.71	0.48
Deviation from the median	−1.03	0.23	−4.40	0.000441

Table 7. Final standings in the 2018 season of the Brazilian soccer league *Serie A* in alphabetic order.

Club	Points won in the first half-season (P1)	Deviation from the median (22.5)	Points won in the second half-season (P2)	P2-P1
América MG	22	−.5	18	−4
Atlético MG	33	10.5	26	−7
Atlético PR	21	−1.5	36	15
Bahia	25	2.5	23	−2
Botafogo	22	−.5	29	7
Ceará	16	−6.5	28	12
Chapecoense	21	−1.5	23	2
Corinthians	26	3.5	18	−8
Cruzeiro	26	3.5	27	1
Flamengo	37	14.5	35	−2
Fluminense	23	.5	22	−1
Grêmio	36	13.5	30	−6
Internacional	38	15.5	31	−7
Palmeiras	33	10.5	47	14
Paraná	14	−8.5	9	−5
Santos	21	−1.5	29	8
São Paulo	41	18.5	22	−19
Sport	20	−2.5	22	2
Vasco da Gama	21	−1.5	22	1
Vitória	19	−3.5	18	−1

Table 8. Regression with intercept for all 20 clubs in the 2018 season.

	Coefficient	Standard error	*t*-statistic	*p*-value
Intercept	1.69	1.79	.94	0.35
Deviation from the median	−0.52	0.21	−2.40	0.0273

Table 9 shows a similar angular coefficient also significant at 5 percent in a regression run without the intercept.

These results seem to depend on outliers, however. Table 10 shows how results change after dropping São Paulo and Atlético PR from the regression. Though keeping their expected signs, coefficients significantly change in value, and *p*-values increase.

2.4. Pooling the Seasons

Regression to the median could not be discarded after pooling the three seasons, either. The angular coefficient was negative and significant at 1 percent. However,

Table 9. Regression without intercept for all 20 clubs in the 2018 season.

	Coefficient	Standard error	*t*-statistic	*p*-value
Deviation from the median	−0.44	0.19	−2.21	0.0391

Table 10. Regression with intercept for 18 clubs in the 2018 season (apart from outliers São Paulo and Atlético PR).

	Coefficient	Standard error	*t*–statistic	*p*–value
Intercept	0.93	1.60	0.58	0.56
Deviation from the median	−0.26	0.21	−1.26	0.22

the intercept became non-significant and the R squared was 0.33. Moreover, running regressions without either the intercept or outliers did not change the results a great deal (Table 11).

In sum, we cannot dismiss the hypothesis of regression to the median in the 2016, 2017 and 2018 seasons of the Brazilian soccer league *Serie A*. Luck played a role in club success or failure. However, though we do not need it, a causal explanation can still overlap the statistical fact of regression to the mean. For this reason, we turn next to evaluate how a causal explanation might also matter.

3. The Hypothesis of Loss Aversion

One credible narrative is that club performance can also depend on loss aversion, that is, clubs fighting relegation have more incentive to win matches than clubs targeting promotion to *Copa Libertadores*. To test this hypothesis, we consider the eight teams on top and the eight at the bottom in Matchday 28 of 38 in each of the three seasons, as in Table 12.

Table 13 shows that performance of the last eight clubs across the three seasons improved 9.11 percent on average (that is, 41.32 minus 32.21) with 10 matchdays remaining. This translates into an average advantage of 0.2733 per match for the eight clubs at the bottom in their direct confrontation matches with the top eight. However, the 2016 season showed an anomalous behavior, possibly for the reasons involving Santa Cruz and Internacional, as discussed in Section 2.

To evaluate the statistical significance of the mean differences in Table 13, we conducted a paired sample *t*-test. A value of 15.42 (with a Student's *t* distribution using two degrees of freedom) suggested a significance at 1 percent.

4. Discussion

Loss aversion can explain club performance as an regression to the mean. Puzzling enough, what was found as regression to the mean in Section 2 could have been explained by loss aversion as well. Likewise, what was found as loss aversion in Section 3 could have been explained by regression to the mean, too. Then, how can we disentangle the role of each in club performance? Here, some arithmetic might be useful.

Table 11. Regression with intercept for 60 data points after pooling the 2016, 2017 and 2018 seasons.

	Coefficient	Standard error	*t*–statistic	*p*–value
Intercept	1.21	0.89	1.36	0.17
Deviation from the median	−0.64	0.11	−5.44	1.08E−06

Table 12. The bottom eight clubs fighting relegation versus the top eight clubs targeting promotion in Matchday 28 in each of the three seasons.

2016		2017		2018	
Relegation	Promotion	Relegation	Promotion	Relegation	Promotion
América MG (35)	Atlético MG	AtléticoGO (36)	Botafogo	AméricaMG (38)	AtléticoMG
Cruzeiro	AtléticoPR	Avaí (38)	Corinthians	Bahia	Flamengo
Figueirense (36)	Botafogo	Chapecoense	Cruzeiro	Ceará	Fluminense
Internacional (38)	Corinthians	Coritiba (38)	Flamengo	Chapecoense	Grêmio
Santa Cruz (35)	Flamengo	Ponte Preta (38)	Grêmio	Paraná (32)	Internacional
São Paulo	Fluminense	São Paulo	Palmeiras	Sport (38)	Palmeiras
Sport	Palmeiras	Sport	Santos	Vasco	Santos
Vitória	Santos	Vitória	Vasco	Vitória (37)	São Paulo

Note: Figures in parentheses show the matchday where relegation actually occurred, and thereafter a relegated club is left out from the sample.

Table 13. Results of the direct confrontation matches between the eight clubs fighting relegation and the eight clubs targeting promotion across the three seasons.

	2016	2017	2018	Mean
Number of direct confrontation over the first 10 matchdays	31	42	31	34.67
Percent of points won by the clubs fighting relegation over the first 10 matchdays	37.8	34.12	24.72	32.21
Number of direct confrontation over the last 10 matchdays	39	32	29	33.33
Percent of points won by the clubs fighting relegation over the last 10 matchdays	33.33	45.82	44.82	41.32

First, assume there is no role for luck in the points won by the clubs, that is, only loss aversion explains performance. Note that a club fighting relegation has eight confrontation matches against the clubs aiming at promotion in the second half-season. This means there is an expected total average advantage for the club of 2.19 points (that is, 8×0.2733). However, we also computed the average performance for those clubs below the median in the three seasons as −4.03. Considering the regression in Table 11, a club with a deviation from the median of −4.03 has an expected performance in the second half-season of 3.79 points, that is, $1.21 - 0.64 \times (-4.03)$. Because $3.79 > 2.19$, the residual 1.6 means it is not correct to assume that only loss aversion explains club performance.

Now hypothesize loss aversion plays no role. In this case, we have to assume

the linear coefficient of the estimated regression line in Table 11 is nil. This puts a lower bound to regression to the mean. Note that, in this case, a club fighting relegation's expected points would be 2.58, that is $-0.64 \times (-4.03)$. By further assuming that a club fighting relegation faces the same degree of difficulty to beat each of the other 12 clubs that are not fighting relegation, this means an expected total average advantage of 3.28 (that is, 12×0.2733). The residual 0.7 means it is also not correct to assume that only regression to the mean explains club performance.

Thus, the two exercises above demonstrate that club performance should be explained by both regression to the mean and loss aversion.

5. Conclusions

Pundits and fans alike fall prey of the narrative fallacy when explaining soccer club performance by talent alone. This tendency is ingrained in our automatic mind. However, success is also a matter of luck. Here, we consider loss aversion to explain club success and failure, but also the role luck plays through the commonly neglected phenomenon of regression to the mean.

Our data refer to recent seasons of the first-division Brazilian soccer league, called *Serie A*. To test regression to the mean, we examine whether the clubs scoring above the median in the first half-season tend to score relatively fewer points in the second, whereas the clubs that score points below the median in the first half-season tend to end up with relatively more points in the second. And to test loss aversion, we investigate whether the clubs struggling to escape the relegation zone perform relatively better than the clubs aiming at promotion to *Copa Libertadores*. Here, loss aversion means the underdogs have more to lose than the favorites.

In the end, we find that club performance should be explained by both regression to the mean and loss aversion, and provide an exercise to quantify the role each play.

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] Gladwell, M. (2008) Outliers: The Story of Success. Little, Brown and Company, New York.

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

[3] 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

[4] Galton, F. (1886) Regression towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246-263. http://www.stat.ucla.edu/~nchristo/statistics100C/history_regression.pdf https://doi.org/10.2307/2841583

[5] Bulmer, M. (2003) Francis Galton: Pioneer of Heredity and Biometry. Johns Hopkins University Press, Baltimore, MD.

[6] Wainer, H. (2007) The Most Dangerous Equation. *American Scientist*, **95**, 249-256.

[7] 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

[8] Turner, M. (1996) The Literary Mind: The Origins of Thought and Language. Oxford University Press, New York.

[9] Taleb, N.N. (2010) The Black Swan: The Impact of the Highly Improbable. Random House, New York.