How cann soccer improve statistical learning?

Enivaldo C. Rocha
Federal University of Pernambuco - Brazil
enivaldorocha@sigobr.com

Dalson Britto Figueiredo Filho
Federal University of Pernambuco - Brazil
dalsonbritto@sigobr.com

Ranulfo Paranhos
Federal University of Alagoas - Brazil
ranulfoparanhos@sigobr.com

José Alexandre Silva Jr.
Federal University of Alagoas - Brazil
jasjunior@sigobr.com

Denisson Silva
Federal University of Alagoas - Brazil
denissonsilva@sigobr.com

Abstract

This paper presents an active classroom exercise focusing on the interpretation of ordinary least squares regression coefficients. Methodologically, undergraduate students analyze Brazilian soccer data, formulate and test classical hypothesis regarding home team advantage. Technically, our framework is simply adapted for others sports and has no implementation cost. In addition, the exercise is easily conducted by the instructor and highly enjoyable for the students. The intuitive approach also facilitates the understanding of linear regression practical application.

Keywords: statistics; linear regression; quantitative methods; soccer; fun.

"If I had only one day left to live, I would live it in my statistics class: it would seem so much longer"

INTRODUCTION[1]

The embracing of fun procedures to advance statistical learning is widely supported by both experimental studies (GARNER, 2006; BERK and NANDA, 1998; FRIEDMAN, HALPERN and SALB, 1999; BERK e NANDA, 2006) and theoretical literature (LUNDBERG and THURSTON, 1992; FRIEDMAN, FRIEDMAN and AMOO, 2002). In particular, humor strengthens the relationship between student and teacher (KAPLAN and PASCOE, 1977; RUNYON, 1977), reduces stress (BRYANT, COMISKY, CRANE, and ZILLMANN, 1980; BLUEMENFELD and ALPERN, 1985), makes a course more interesting (PYRCZAK, 1998; FLOWERS, 2001; WHISONANT, 2001) and enhances material recall (FRIEDMAN, FRIEDMAN and AMOO, 2002). Still regarding teaching procedures, the American Statistical Association GAISE Report (2010) recommends to: (1) emphasize statistical literacy and develop statistical thinking; (2) use real data; (3) stress conceptual understanding; (4) foster active learning in the classroom; (5) use technology for developing conceptual understanding and analyzing data and (6) use assessments to improve and evaluate student learning (ASA, 2010).

This paper outlines a structured classroom exercise that covers most of ASA (2010) recommendations. The focus relies on the intuitive interpretation of ordinary least squares regression coefficients. Methodologically, undergraduate students analyze 2013 Brazilian national soccer data, formulate and test classical hypothesis regarding home team advantage. Technically, our framework is simply adapted for others sports and has no implementation cost. In addition, the exercise is easily conducted by the instructor and highly enjoyable for the students. The intuitive approach also facilitates

[1] This classroom exercise was pioneered applied by professors Dalson Figueiredo and Enivaldo Rocha during the 2013.2

the understanding of linear regression practical application. The remainder of the paper is divided as follows. Next section outlines our structured classroom exercise[2]. The final section presents our concluding remarks.

STRUCTURED CLASSROOM EXERCISE

The exercise is structured as follows:

(1) Introduce linear regression basic principles[3]and assumptions[4]
(2) Present the hypothesis regarding home team advantage
(3) Share the dataset with students
(4) Write down the model
(5) Run the model
(6) Interpret the results

Introductory textbooks in Statistics and Econometrics teach that regression can be used to estimate the effect of different independent variables on a dependent variable. As long assumptions are met, ordinary least squares estimates will be unbiased and efficient. The basic model is:

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

The intercept (α) represents the best guess of the dependent variable (Y) in a model without any independent variables. β_1 is the average effect observed in Y when X (independent variable) increases by one unit. The ε represents the error term.

The second step is to present the hypothesis regarding home team advantage.

$$H_o: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

The null hypothesis (H_o) asserts that there is no home team advantage ($\beta_1 = 0$). In other words, we should expect no difference between the average number of points acquired playing as home team compared to playing as a visiting team. Differently, the alternative hypothesis (H_a) states that the effect of playing at home is different from zero ($\beta_1 \neq 0$). Specifically, conventional sports wisdom inform that home teams have a positive advantage ($\beta_1 > 0$).

The third step consists in share the datasets with the students. As long the classroom exercise is conducted in the laboratory, the instructor distributes the file using the course institutional email. Students download the file and open it using STATA, version 12.

[2] Data are available at http://dx.doi.org/10.7910/DVN/25240

[3] Students interested in learning more about ordinary least squares regression should check the following. For intuitive introductions see Lewis-Beck (1980), Berry and Feldman (1985) and Schroeder, Sjoquist and Stephan (1986). To a more advanced approach see Tabachnick and Fidell (2007) and Gelman and Hill (2007). To an intuitive introduction introduction in portuguese see Figueiredo Filho *et al*. (2011).

[4] (1) linearity; (2) no systematic measurement error; (3) expected mean of error term equals to zero; (4) homoscedasticity; (5) no autocorrelation; (6) no correlation between independent variables and the error term; (7) correct model specification; (8) no multicollinearity; (9) error term follows a normal distribution and (10) there is a adequate proportion between the number of cases and the number of parameters estimated.

Next the instructor writes the model down in the blackboard, emphasizing the intuitive meaning of each model component:

$$\text{Number of points} = \alpha + \beta_1 \textit{Hometeam} + \varepsilon$$

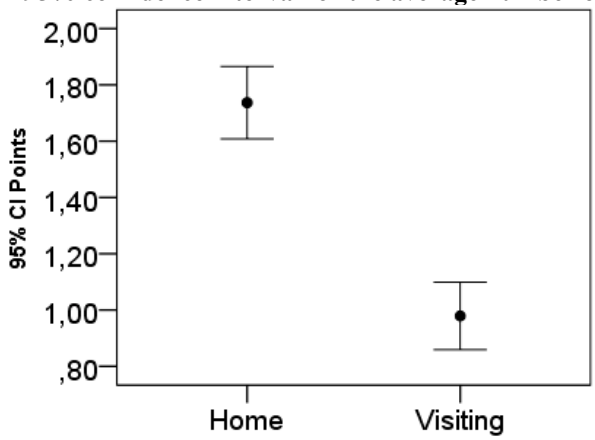
The dependent variable, the number of points acquired during the 2013 Brazilian soccer national championship, is a linear function of the intercept plus a dummy variable coded as "1" when the team plays at home and "0" otherwise[5].

The fifth step is to run the model. Students receive the Stata do file. In addition, the instructor runs the computational code using a PowerPoint projector. The teacher assistant checks if all students properly run the model, helping those facing any difficulties. The final step is to interpret the estimated coefficients:

$$\text{Number of points} = .979 + .758 \textit{Hometeam} + \varepsilon$$

This model summarizes the difference in the average number of points between home and visiting teams. The intercept, .979, is the average (or predicted) number of points for a visiting team. Algebraically, we just need to plug 0 into the equation (visiting teams = 0). To obtain the average number of points of home teams we need to plug 1 into this equation to obtain .979+.758∗1=1.737. The difference between the two groups is equal to the coefficient of *Hometeam*. This regression coefficient tells us that home teams have an average positive difference of .758 points compared to visiting teams. Figure 1 displays the 95% confidence interval of the average number of points for each group.

Figure 1 - 95% confidence interval for the average number of points



Another procedure to evaluate the effect of the independent variable is to graphically examine the means of each group. As long there is no overlap between the confidence intervals we can safely conclude that there is a statistically significant difference between the two groups. Then, we should reject the null hypothesis of home teams have no advantage compared to visiting teams ($\beta_1 = 0$).

After students understand the interpretation of the linear regression model with a binary predictor, the next step is to include a quantitative independent variable in the model.

$$\text{Number of points} = \alpha + \beta_1 \text{Hometeam} + \beta_2 \text{Ngoals} + \varepsilon$$

[5] We replicate Gelman and Hill (2007) linear regression coefficients interpretation.

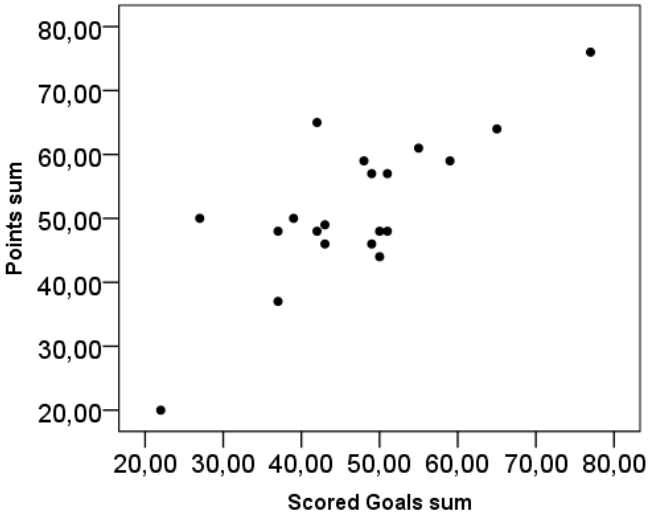
The estimated coefficients are the following:

$$\text{Number of points} = .294 + .432\ \textit{Hometeam} + .688\ \textit{Ngoals} + \varepsilon$$

The intercept. If a soccer team is visiting and scores no goals, then the model predicts an average of .294 points. The coefficient of home team. Comparing teams with the same number of goals but who differed in whether they are playing at home, the model predicts an average difference of .432 points. Put differently, home teams have an .432 points advantage, controlling for the number of scored goals.

Finally, the coefficient of Ngoals. Comparing teams in the same condition of match location, but who differed by 1 goal, the model predicts an average difference of .688 points. Or, each additional goal increases, on average, the number of points in .688, controlling for the match location. Figure 2 displays the correlation between the sum of number of goals and the number of points per team.

Figure 2 - Aggregated correlation between scored goals and number of points



Individual level data suggest a positive correlation (.640) between the number of goals and the number of points (n = 760; p-value<.000). Aggregated data indicate the same conclusions (r = .773; n = 20; p-value<.000).

CONCLUSION

We firmly believe that "sadistics" can be replaced by "funtistics" by adopting structured humor classroom exercises. This paper outlined an intuitive active classroom exercise to teach the interpretation of ordinary least squares regression coefficients. Methodologically, undergraduate students analyzed Brazilian soccer matches data, formulated and tested classical hypothesis regarding home team advantage. Technically, our framework is simply adapted for others sports and has no implementation cost. In addition, the exercise is easily conducted by the instructor and highly enjoyable for the students. The intuitive approach also facilitates the understanding of linear regression practical application. With this paper we hope to help students not only enjoy learn statistics but also apply statistical reasoning in both academic and personal life.

REFERENCES

Berk, R. A., and Nanda, J. P. (1998). "Effects of jocular instructional methods on attitudes of anxiety and achievement in statistics courses", *HUMOR: International Journal of Humor Research*, 11, 383-409.

Berk, R. A., and Nanda, J. P. (2006). "A randomized trial of humor effects on test anxiety and test performance", *HUMOR: International Journal of Humor Research*, 11, 383-409.

Blumenfeld, E., and Alpern, L. (1985). *The Smile Connection: How to Use Humor in Dealing With People*, Englewood Cliffs, NJ: Prentice Hall.

Bryant, J.; Comisky, P.W.; Crane, J. S.; and Zillmann, D. (1980). "Relationship between college teachers' use of humor in the classroom and students' evaluations of their teachers," *Journal of Educational Psychology*, 72, 511-519.

Flowers, J. (2001). "The value of humor in technology education," *Technology Teacher*, 60, 10-13.

Figueiredo Filho, D. B. *et al.* (2011). O que Fazer e o que Não Fazer com a Regressão: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO). *Revista Política Hoje,* Vol. 20, n. 1, p. 44-99.

Friedman, H. H.; Friedman, L. W.; and Amoo, T. (2002). "Using humor in the introductory statistics course", *Journal of Statistics Education*, 10(3) www.amstat.org/publications/jse/v10n3/friedman.html

Friedman, H. H.; Halpern, N.; and Salb, D. (1999). "Teaching statistics using humorous anecdotes", *Mathematics Teacher*, 92, 305-308.

Garner, R. L. (2006). "Humor in Pedagogy: How Ha-Ha Can Lead to Aha!," *College Teaching*, 54, 177-180.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press; 2006

Kaplan, R. M., and Pascoe, G. C. (1977). "Humorous lectures and humorous examples: Some effects upon comprehension and retention", *Journal of Educational Psychology*, 69, 61-65.

Lewis-Beck, Michael (1980). Applied Regression: an introduction. Series Quantitative Applications in the Social Sciences. SAGE University Paper.

Lundberg, E., and Thurston, C. M. (1992). *If They're Laughing, They're Not Killing Each Other*, Fort Collins, CO: Cottonwood Press.

Pyrczak, F. (1998). *Statistics With a Sense of Humor*, Los Angeles, CA: Pyrczak Publishing.

Runyon, R. P. (1977). *Winning with Statistics: A Painless First Look at Numbers, Ratios, Percentages, Means, and Inference*, Reading, MA: Addison-Wesley.

Schroeder, L. D.; Sjoquist, D. L.; & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide*. Beverly Hills: Sage Publications.

Tabachnick, Barbara; Fidell, Linda. (2007). Using multivariate analysis. Needham Heights, Allyn e Bacon.

Whisonant, R. D. (1998). *The Effects of Humor on Cognitive Leaning in a Computer-Based Environment*, Unpublished doctoral dissertation, Virginia Polytechnic Institute, Blacksburg, VA.