▼ Journal of Research (Science), Bahauddin Zakariya University, Multan, Pakistan.
Vol. 24-25, Nos. 1-4 January-October 2013-14, pp. 39-46 ISSN 1021-1012

INFERENCE ABOUT THE FIFA MATCHES: AN APPLICATION OF THE LOGISTIC REGRESSION AND CLASSICAL TESTS

Haris Khurram and Muhammad Aslam*
Department of Statistics, Bahauddin Zakariya University Multan-Pakistan.
email: hariskhurram2@gmail.com, aslamasadi@bzu.edu.pk.

Abstract
The basic idea behind this work is to apply some classical inferential techniques for the goals scored in the football world cup tournaments. The Fédération International de Football Association (FIFA) organizes this event after every four years. According to the FIFA ranking, first three ranks are given to the national teams of Spain, Germany and Argentina, respectively. We take 56 matches of Spain, 99 matches of Germany and 70 matches of Argentina. The logistic regression is used to find the match-win prediction using different factors. As a result Germany is found to be an unpredictable team. For testing the mean goal of the said teams, we used the Wald, LR and Score tests. To compare the power of these tests empirically and graphically, the Monte Carlo simulation scheme is used.

Keywords: LRT, Match-win prediction, Null rejection rate, Power curve, Score test, Wald test.

INTRODUCTION
Now a days, sports and related issues are popular topics for researchers. Football (soccer) is one of the most popular sports of the world. Many researchers move toward the modeling and simulation for this game. FIFA (Fédération International de Football Association) was founded in Paris in 1904. FIFA has a ranking system and it provides ranks to the football teams according to their performance. Using this ranking system in 2014, Spain, Germany and Argentina were ranked at first, second and third position, respectively (www.fifa.com).

It has been quite interesting for the researchers in sports to statically analyze the outcomes of the popular games, namely football. For instance, Willoughby (2002) used logistic regression on the Canadian Football League. The team with greatest number of win was chosen as very good, the team with mostly equal win and loss was chosen as average and the team with lowest win was chosen as poor. Radoman and Smajic [2008] used statistical techniques to establish the impact of game stoppage on the final score of the football match. Recently, Nyberg [2014]

* Corresponding author

presented a new statistical test for market efficiency in football betting market. This test is based upon multinomial logit model where maximum likelihood estimator (MLE) was used and likelihood based tests for the efficient market hypothesis.

The present study focuses on statistical analysis of the goals, scored by the top three teams (Spain, Germany and Argentina) in the FIFA world cup matches from 1938 to 2010. Summary statistics for these teams are shown in Table 1.

Table 1: Summary Statistics of the Football Teams of Spain, Germany and Argentina.

Team	Matches Played	Total Goals	Matches Won	Matches Lost	Average Goals	Points
Spain	56	88	29	27	1.571	1513
Germany	99	206	65	34	2.081	1311
Argentina	70	123	40	30	1.757	1266

(Source: wikipedia.org)

LOGISTIC REGRESSION

It is non-linear regression model which can be converted into linear model by using a simple transformation. Logistic regression is used to find the probability of the occurrence of an outcome of interest. A conventional logistic regression model is written as

$$Y_i = \frac{1}{1+e^{-\left(\alpha+\sum_{j=1}^{k}\beta_j X_{ji}\right)}}, \qquad (1)$$

where Y_i is the response variable and is defined as:

$$Y_i = \begin{cases} 1: \text{if the } ith \text{ match is won;} \\ 0: \text{if the } ith \text{ match is lost.} \end{cases}$$

Whereas X's are the predictors and are defined below with their respective notations used in place of X_{ji} in (1) with $j = 1, 2, \ldots, k$ and $i = 0,1$.

$$N_i = \begin{cases} 1: ith \text{ match played at night;} \\ 0: ith \text{ match played at day.} \end{cases}$$

$$O_i = \begin{cases} 1: \text{Own goal in the } ith \text{ match;} \\ 0: \text{Not any own goal in the } ith \text{ match.} \end{cases}$$

$$E_i = \begin{cases} 1: \text{Extra time given in } ith \text{ match;} \\ 0: \text{No extra time given in } ith \text{ match.} \end{cases}$$

$P_i =$ Numbers of penalty shoot − outs in ith match.

$R_i =$ Numbers of red cards issued in ith match.

In the present study, model (1) is used for estimation of match winning probability using the above stated factors for the above stated football teams.

SOME CLASSICAL PARAMETRIC TESTS

Let Z_i denotes the number of goals in ith match then Z_i follows the Poisson distribution with some unknown mean θ and the MLE of θ, denoted as $\hat{\theta}$, is the sample mean $\bar{Z}$. It may be of keen interest for researchers to test average goal per match for a football team. If θ_o is some hypothetical value of θ then the typical null hypothesis is $H_0 : \theta = \theta_0$. For this purpose, the classical parametric tests are the likelihood ratio test (LRT) [Neyman and Pearson 1928], Wald test [Wald 1943] and score test [Rao 1948]. However, these tests are asymptotically equal [Buse 1982].

THE LRT

The LRT is used to compare the fit of two models [Neyman and Pearson 1928]. LR test statistic which is usually denoted by "$\lambda(\boldsymbol{x})$" can be defined as:

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_o, x)}{L(\hat{\theta}, x)}, \tag{2}$$

where $\hat{\theta}$ the MLE of θ.

For the given case of the Poisson distribution,

$$\lambda(\boldsymbol{x}) = \frac{e^{-n\theta o} \times \theta_0^{\sum x_i}}{e^{-n\hat{\theta}} \times \hat{\theta}^{\sum x_i}} . \tag{3}$$

According to Wilks theorem [Wilks 1938], the asymptotic distribution of

$$R = -2ln\left[\frac{e^{-n\theta o} \times \theta_0^{\sum x_i}}{e^{-n\hat{\theta}} \times \hat{\theta}^{\sum x_i}}\right]$$

is Chi-square with 1 degree of freedom (d.f).

WALD TEST

If $\hat{\theta}$ has a limiting normal distribution and the Fisher information $I(\theta)$ is consistently estimated by $I(\hat{\theta})$ then

$$W = (\hat{\theta} - \theta_0)^2 I(\hat{\theta}) \tag{4}$$

will have a limiting Chi-square distribution with 1 d.f.

The statistic, given in (4), is called the Wald statistic [Wald 1943].

For the stated case,

$$I(\hat{\theta}) = -E\left[\frac{\partial^2}{\partial \theta^2} \Sigma lnf(X; \theta)\right] = \frac{n}{\hat{\theta}}.$$

Since the MLE of $\theta$is the sample mean $\bar{X}$, Eq. (4) becomes

$$W = (\bar{X} - \theta_o)^2 \left(\frac{n}{\bar{X}}\right).$$

SCORE TEST

This test focuses on the characteristics of the log likelihood function when the restriction of the null hypotheses is imposed. The main advantage of this test is that the estimation of unknown parameter is not required under the null hypotheses [Rao 1948].

Let $S(\theta_o)$ be the score function and $I(\theta_0)$ be the Fisher information under H_0. The test statistic for the score test is defined as

$$T = S^2(\theta_0)I^{-1}(\theta_0). \quad (5)$$

For the stated case,

$$S(\theta_o) = \frac{\partial}{\partial\theta_0}\Sigma lnf(X;\theta_0) = \frac{\sum x_i}{\theta_o} - n,$$
$$I(\theta_o) = \frac{n}{\theta_o}.$$

Thus, Eq. (5) becomes,

$$T = \frac{\left[\frac{\sum x_i}{\theta_o} - n\right]^2}{\frac{n}{\theta_o}},$$

which is asymptotically Chi-square distributed with 1 d.f.

RESULTS AND DISCUSSION

For the national team of Spain, results for logistic regression are shown in Table 2.

Table 2: Logistic regression analysis for the national team of Spain

Predictor	Coefficient	*Z*	*p-value*	*OR*
Constant	-0.4965	-1.16	0.245	-
Ni	0.6565	1.11	0.269	1.93
O_i	-1.2372	-0.81	0.416	0.29
E_i	-0.4303	-0.46	0.647	0.65
P_i	1.9235	2.28	**0.023**	**6.85**

Table 3: Logistic regression analysis for the national team of Germany

Predictor	Coefficient	*Z*	*P*	*OR*
Constant	0.4211	1.38	0.169	-
Ri	-1.0366	-1.25	0.211	0.35
N_i	0.5277	1.19	0.235	1.69
E_i	-0.1963	-0.27	0.791	0.82
P_i	1.2438	1.17	0.242	3.47

In Table 2, *p*-value of factor P_i (penalty shoot-out) is 0.023 which means this factor has significant role in the winning of the Spanish team. As the coefficient of this factor is positive this shows that penalty shoot-out has positively effect on the winning of the national team of Spain. The *OR* for penalty shoot-out is 6.85 which show that Spain wins a match 6.85 times more than a match in which there is no penalty shoot-out.

For the national team of Germany, the results for logistic regression are shown in Table 3.

After analysis, it is found that there is no significant factor for the winning of the national team of Germany. Thus, the results reveal that the national team of Germany is unpredictable team, regarding the stated factors.

Similarly, Table 4 displays the results for the national team of Argentina.

Table 4: Logistic regression analysis for the national team of Argentina

Predictor	Coefficient	*Z*	*P*	*OR*
Constant	0.8894	2.23	**0.026**	**-**
R_i	-1.7746	-2.08	**0.038**	**0.17**
N_i	-0.7795	-1.46	0.143	0.46
P_i	0.8229	0.93	0.354	2.28

In Table 4, just three factors were taken instead of five due to the available information for the Argentinean team. Here, the only significant factor is "red card" (*Ri*) and its coefficient is -1.7746. Its negative value shows that this factor has negative impact on the wining of the national team of Argentina. The OR for this factor is 0.17, showing that the national team of Argentina loses its 17% matches due to receiving of red cards.

TESTING

On the basis of the descriptive statistics, given in Table 1, we desire to test whether the average goals per match of each of the teams is 2 or not i.e., it is desired to test

$$H_0 : \theta = 2.$$

Table 5 displays the test results for the Wald, LR and Score tests for all the three teams.

Table 5: Testing of average goals per match

Team	Tests	Test statistic	p-value
Spain	Wald	6.5454	0.0105
	LRT	5.5555	0.0184
	Score	5.1428	0.0233
Germany	Wald	0.3107	0.5772
	LRT	0.3190	0.5722
	Score	0.3232	0.5697
Argentina	Wald	1.7716	0.1832
	LRT	1.6434	0.1998
	Score	1.5845	0.2081

The results in Table 5 reveal that the average of goals per match is less than 2 for the Spanish team but the other two teams score 2 goals, on average, in each match. Furthermore, all the three test statistics provide the similar conclusions for each of the teams. Therefore, for finite sample comparison, we plan to conduct some Monte Carlo simulations to evaluate the empirical null reject rate (NRR) and power of all the three stated tests. For this purpose, we take different samples of size n = 25, 50, 100 and 200 and run 5000 simulations for each sample size. For the

simulations, we generate random data from the Poisson distribution with mean $\theta = 2$. Table 6 shows the empirical NRR for all the three tests at the nominal level of significance i.e. 5%. Obviously for $\theta = 2$, the NRR represents the empirical level of significance while for the rest values of θ, empirical power of the tests are presented. The figures in Table 6 reveal that there is very less size distortion in the Wald tests for small sample (n = 25) as the empirical NRR is found to be very closer to the nominal level of significance (5%). With increase in sample size, all the rejection rates also tend to increase as expected. Furthermore, power of the Wald test is also larger as compared to the other two tests for the small sample of size 25. However, for the large samples, all of the three tests become identical in terms of the NRR. These findings are consistent with those given in Buse [1982]. Fig. 1 represents the findings of Table 6 graphically.

Table 6: The empirical NRR at 5% level of significance

Parameter	Wald Test				LRT				Score Test			
θ	n = 25	54	100	200	n = 25	54	100	200	n = 25	54	100	200
2	0.0490	0.0479	0.0522	0.0469	0.0463	0.0514	0.0490	0.0480	0.0537	0.0463	0.0522	0.0452
2.1	0.0925	0.0960	0.1240	0.1777	0.0543	0.0831	0.0983	0.1660	0.0567	0.0700	0.1240	0.1542
2.2	0.1468	0.2115	0.2972	0.5105	0.1160	0.1870	0.2504	0.4924	0.0910	0.1604	0.2972	0.4686
2.3	0.2234	0.3690	0.5718	0.8365	0.1830	0.3320	0.5152	0.8260	0.1444	0.2966	0.5718	0.8118
2.4	0.3167	0.5628	0.7906	0.9716	0.2691	0.4853	0.7511	0.9681	0.2221	0.4853	0.7906	0.9681
2.5	0.4312	0.7426	0.9263	0.9975	0.3808	0.6745	0.9162	0.9965	0.3809	0.6745	0.9263	0.9965
2.6	0.5908	0.8619	0.9813	0.9999	0.4779	0.8196	0.9734	0.9999	0.4779	0.8196	0.9813	0.9999
2.7	0.6865	0.9354	0.9970	1.0000	0.6374	0.9221	0.9952	1.0000	0.5823	0.9221	0.9970	1.0000
2.8	0.7860	0.9734	0.9994	1.0000	0.7395	0.9684	0.9992	1.0000	0.6944	0.9684	0.9994	1.0000
2.9	0.8550	0.9928	1.0000	1.0000	0.8210	0.9903	1.0000	1.0000	0.7801	0.9903	1.0000	1.0000
3	0.9078	0.9975	1.0000	1.0000	0.8847	0.9963	1.0000	1.0000	0.8847	0.9963	1.0000	1.0000
3.1	0.9561	0.9993	1.0000	1.0000	0.9255	0.9991	1.0000	1.0000	0.9255	0.9991	1.0000	1.0000
3.2	0.9776	1.0000	1.0000	1.0000	0.9695	0.9999	1.0000	1.0000	0.9593	0.9999	1.0000	1.0000
3.3	0.9862	1.0000	1.0000	1.0000	0.9802	1.0000	1.0000	1.0000	0.9732	1.0000	1.0000	1.0000
3.4	0.9940	1.0000	1.0000	1.0000	0.9919	1.0000	1.0000	1.0000	0.9879	1.0000	1.0000	1.0000
3.5	0.9970	1.0000	1.0000	1.0000	0.9960	1.0000	1.0000	1.0000	0.9960	1.0000	1.0000	1.0000
3.6	0.9992	1.0000	1.0000	1.0000	0.9986	1.0000	1.0000	1.0000	0.9979	1.0000	1.0000	1.0000
3.7	0.9995	1.0000	1.0000	1.0000	0.9994	1.0000	1.0000	1.0000	0.9992	1.0000	1.0000	1.0000
3.8	0.9998	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	1.0000
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

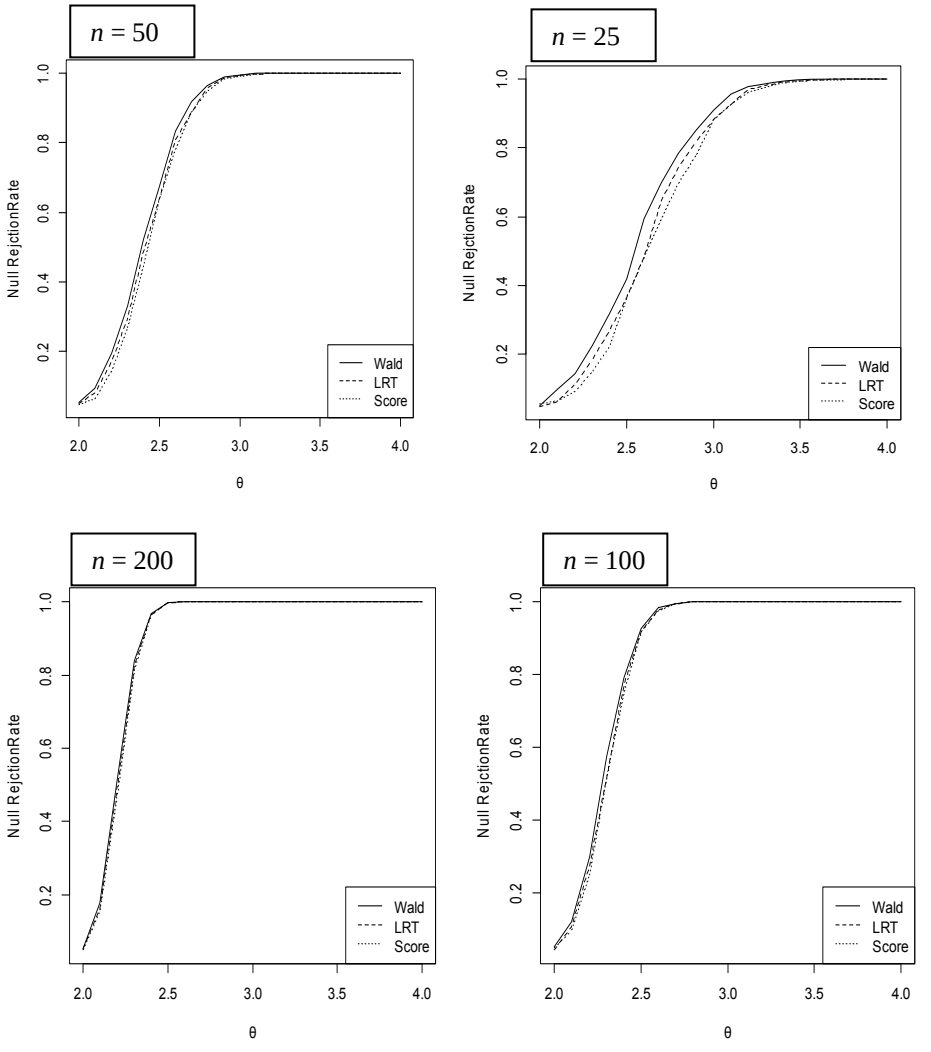

Fig. 1: Power curves for n = 25, 54, 100 and 200.

CONCLUSIONS

When the top three football teams are analyzed, it is concluded that the factor of penalty shoots remains quite favorable in the winning of the Spanish football team. On the other hand, German team is unpredictable while the Argentinean team may lose its match significantly if its players receive red cards.

It is further concluded that for small samples, the Wald test remains the best, yielding highest power and all the three tests i.e. LRT, Wald and Score tests are asymptotically equivalent.

References

Buse, A. (**1982**) “The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note”, *Amer. Stat.* 36(3), 153-157.

Neyman, J. and Pearson, ES. (**1928**) “'On the use and interpretation of certain test criteria for purposes of statistical inference, Part I”, *Biomet.* 20A, 175-240.

Nyberg, H. (**2014**), “A multinomial Logit-based statistical test of association football betting market efficiency”, Discussion Paper No. 380, University of Helsinki, Finland.

Radoman, M. and Smajic, M. (**2008**) “The important elements of game stoppage for the final score of a football match”, *Serb. J. Sports Sci.* 2(1-4), 53-58.

Rao, CR. (**1948**) “Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation”, *Math. Proceed. Cambridge Phil. Soc.* 44, 50-57.

Wald, A. (**1943**) “Tests of statistical hypotheses concerning several parameters when the number of observations is large”, *Transacs. Amer. Math. Soc.* 54, 426-482.

Wilks, SS. (**1938**). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses", *Annals Math. Stat.* 9, 60–62.

Willoughby, KA. (**2002**) “Winning games in Canadian football: a logistic regression analysis”, *The College Math. J.* 33(3), 215-220.

www.fifa.com/fifa-world-ranking/ranking-table/men/rank=236/index.html