ISSN 0976 - 1101
Vol-18, Special Issue (2)

PREDICTING THE OUTCOME OF ICC CRICKET WORLD CUP MATCHES

Shiny Raizada[1], Amritashish Bagchi[1], Harishankar Menon[2] and Nayana Nimkar[3]
[1]Assistant Professor, Symbiosis School of Sports Sciences, Symbiosis International (Deemed University), Pune
[2]Student, Symbiosis School of Sports Sciences, Symbiosis International (Deemed University), Pune
[3]Professor, Symbiosis School of Sports Sciences, Symbiosis International (Deemed University), Pune

ABSTRACT
The purpose of the study was to develop a model to predict the outcome of ICC Cricket World Cup ODI matches (Limited Overs) on the basis of first innings data. These probabilities can assist a team captain or management in considering a certain aggressive or defensive batting or bowling strategy for the next innings. The data was collected from last two world cup tournaments i.e. 2011 and 2015. Data of 98 matches were recorded, out of which 8 matches were not taken into consideration due to three reasons which were 1. Match Abandoned 2. Match Tied 3. Matches resolved by Duckworth Lewis Method. The dependent variable selected for this study was Match Outcome (Win/Loss). Team score, Total Wickets Lost, Toss, Runs Scored in Powerplay, Wickets lost in Powerplay, Team Run Rate and the Total number of Dot balls were selected as the predictor variables. For the purpose of this study only the first innings data was used and in statistical technique Binary Logistic regression was used to predict the outcome of a match (Win/Loss). It was found that the developed Logistic regression Model was significant. According to the statistical significance of the predictor variables, they were numerically weighted and can be used to predict the match outcome. Out of seven predictor variables only the variable Team score was included in the prediction model with coefficient of determination (R^2) of .272 (Cox & Snell) and .363 (Nagelkerke). 72.2 % of match results were correctly classified by the model.

Keywords: Cricket, ICC Cricket World Cup, Prediction model, Win and Loss

INTRODUCTION
Cricket is one of the many sport that require a sphere ball and a bat to play, with a different set of rules, which makes this game unique and different from others. It has evolved over the years starting from a test match followed by one day matches and from past few years T20 cricket has taken a lot of attention. But still the ICC Cricket World Cup is the most prestigious tournament of the all, which is a form of limited overs match (50 overs).

Studies have been done in cricket in terms of physiological, psychological or physical demands of batsmen, wicket keepers, spinners and pace-bowlers in different format of play (Noakes, and Durandt, 2000; Christie and King, 2008; Thelwell, Weston and Greenlees, 2007; Jo-Anne, 2012; Weissensteiner, Abernethy, Farrow, and Gross, 2012; Bagchi and Raizada, 2015). Recently few of the studies have focused on the performance analysis of individual players or a whole team by calculating the effect size (Peterson et al., 2008a; Najdan, Robins and Glazier, 2014). But to the best of my knowledge none of the studies have focused to develop a prediction model to predict the outcome of the match on the basis of first innings match data.

Developing prediction models in sports could be one of the solutions to predict the match outcome. It will help the team captain, coaches and team managers to make different tactics

ISSN 0976 - 1101
Vol-18, Special Issue (2)

during the half time. In Statistics, logistical regression is a popular method for predicting an outcome (binary or multinominal) from a dataset in which one or more independent variables are involved. These variables are also known as predictor variables and can be scale or categorical in nature. In some cases, unstable parameters occur when the total number of Covariates is large or highly correlated.

METHODOLOGY

Purpose of the study was to develop a model to predict the outcome of ICC Cricket World cup matches on the basis of first innings data. The data was collected from last two world cup tournaments i.e. 2011 and 2015 ICC Cricket World cup. Data of 98 matches were recorded, out of which 8 matches were not taken into consideration due to three reasons which were 1. Match Abandoned 2. Match Tied 3. Matches resolved by Duckworth Lewis Method. As one of the few assumptions in logistic regression is that, it requires the dependent variable to be binary in nature. Therefore, the dependent variable selected for this study was Match Outcome (Win/Loss). Team Score (TS), Total Wickets Lost (TWL), Toss, Runs Scored in Powerplay (RSP), Wickets lost in Powerplay (WLP), Team Run Rate (TRR) and the Total number of Dot balls (TNDB) were selected as the predictor variables. All the data were collected from the ESPNcricinfo website. For the purpose of this study only the first innings data was used and in statistical technique Binary Logistic Regression was used to develop the prediction model. Descriptive statistics was used to see the nature of data.

All the assumptions were taken care of before running the analysis. For this purpose Statistical Package for Social Science (SPSS) version 24.0 was used. The level of significance was set at 0.05.

FINDING AND RESULT

Logistic regression does not have much key assumptions similar to linear regression and general linear models that are based on ordinary least squares algorithms, such as linearity, normality, homoscedasticity, and measurement level. Therefore, only the descriptive statistics (i.e. mean, standard error of mean, standard deviation, skewness, kurtosis etc) was used to see the nature of data and the correlation matrix was used to check the assumption of high multicollinearity among the variables, which is one of the few assumptions that need to be fulfilled.

Table 1- Descriptive Statistics of all Scaled Variables

	TS	WL	RSP	WLP	TRR	TNDB
Mean	260.2111	8.1000	44.9889	1.3889	5.4058	145.4000
Std. Error of Mean	8.35919	.21379	1.54479	.11027	.13891	2.52591
Std. Deviation	79.30222	2.02817	14.65521	1.04607	1.31781	23.96289
Skewness	-.376	-.751	.627	.784	.016	-.055
Std. Error of Skewness	.254	.254	.254	.254	.254	.254
Kurtosis	-.341	.075	.014	1.063	-.596	-.131
Std. Error of Kurtosis	.503	.503	.503	.503	.503	.503

ISSN 0976 - 1101
Vol-18, Special Issue (2)

Table 2- Correlations Matrix

		Toss	TS	TWL	RSP	WLP	TRR	TNDB
Toss	Point Biserial correlation	1	.234*	-.152	.016	.004	.246*	-.028
TS	Pearson Correlation	.234*	1	-.716**	.403**	-.490**	.969**	-.407**
TWL	Pearson Correlation	-.152	-.716**	1	-.158	.342**	-.717**	.472**
RSP	Pearson Correlation	.016	.403**	-.158	1	-.439**	.454**	-.372**
WLP	Pearson Correlation	.004	-.490**	.342**	-.439**	1	-.461**	.319**
TRR	Pearson Correlation	.246*	.969**	-.717**	.454**	-.461**	1	-.568**
TNDB	Pearson Correlation	-.028	-.407**	.472**	-.372**	.319**	-.568**	1
*. Correlation is significant at the 0.05 level (2-tailed).								
**. Correlation is significant at the 0.01 level (2-tailed).								

One of the assumption in logistic regression is that there should not be high multicollinearity among the independent variables. The above table of correlation matrix shows the correlation coefficient between sets of variables and was used to check the multicollinearity assumption.

Although there is a significant correlation between the variables but none of the variable was found to be highly correlated. And this was checked by calculating Variance Inflation Factor (VIF) using SPSS. For all the variables the VIF value was 1, which means there was no multicollinearity between the independent variables. Hence, we can continue with the logistic regression analysis.

Table 3 - Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	28.621	1	.000
	Block	28.621	1	.000
	Model	28.621	1	.000

As compared to -2 Log Likelihood value (i.e. 124.589) of the null model, the omnibus test of model coefficients shows a significant decrease in the -2 Log Likelihood value (i.e. 95.967), it means the developed model is significantly better fit than the null model.

Table 4 - Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	95.967[a]	.272	.363
a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.			

Unlike linear regression in logistic regression there is no actual R^2 (Coefficient of Determination) value, which summarizes the proportion of variance in the dependent variable, explained by the independent variable selected by the model. Higher the proportion better will be the model. It can be seen from the above table that in the second model the value of Nagelkerke

ISSN 0976 - 1101
Vol-18, Special Issue (2)

R^2 is .363 and the value of Cox & Snell R-square is found to be .272. Both Nagelkerke and Cox & Snell R-square values are the approximation of actual R^2 value. The Nagelkerke R^2 value was considered for the developed model because in Cox & Snell R-square even for a "perfect" model with categorical outcomes, it has a theoretical maximum value of less than 1. Nagelkerke R^2 is the adjusted version of the Cox & Snell R-square that adjusts the scale of the statistic to cover the full range from 0 to 1 ("IBM Knowledge Center", 2018). The value of Nagelkerke R^2 is .363 which means 36.3 % of the variability in the dependent variable is explained by the selected independent variables.

Table 5 - Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.629	8	.889

The Hosmer-Lemeshow test (HL test) is a goodness of fit test for developed logistic regression model. It tests the null hypothesis that the fitted model is correct, which means the p – value should be insignificant to reject the null hypothesis. In the above table, the p – value is .889 which is greater than .05. Hence the model fit is good, in other words the observed event rates match the expected event rates in population subgroups.

Table 6 - Classification Table[a]

Observed			Predicted		
			MATCH RESULTS		Percentage Correct
			LOSS	WIN	
Step 1	MATCH RESULTS	LOSS	27	16	62.8
		WIN	9	38	80.9
	Overall Percentage				72.2
a. The cut value is .500					

The above table shows the summary of correct and wrong classification of the subjects in match Outcome (i.e. Loss or Win) on the basis of the developed regression model. It unveils the number of wins predicted by the logistic regression model compared to the number actually observed and similarly the number of losses predicted by the logistic regression model compared to the number actually observed. Overall 72.2 % of matches were correctly classified on the basis of selected independent variables.

Table 7 - Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	TS	.018	.004	19.093	1	.000	1.018
	Constant	-4.580	1.108	17.079	1	.000	.010
a. Variable(s) entered on step 1: Team Score.							

The above table provides the regression coefficient (B), the Wald statistic (used to test the significance of individual coefficients in the model) and the all-important Odds Ratio (Exp (B)). "B" coefficients are also known as unstandardized coefficients and are used to develop the regression equation (Bewick, Cheek & Ball, 2005). Only the variable team score is selected by the model. Cricket is an unpredictable game where fortunes can change in a matter of time. The result depends on many factors which works together and makes this game unpredictable. These factors are fall of wickets in crucial situation/time, drop catch of a good batsmen, set batsmen wicket, wrong decisions by umpire, whether condition change suddenly, pitch condition, unexpected Run out and many more. One of the other reason that many of the variables were not

ISSN 0976 - 1101
Vol-18, Special Issue (2)

selected by the model is - in ICC World Cup the level of teams are not same, for example if Canada won the toss against Australia, no matter whatever is the decision the chances of winning the match against Australia is too less.

The unpredictable nature of this game can be understood by the following two events of cricket timeline -

1983 WORLD CUP

In 1983, India won the ODI world cup under Kapil Devs Captaincy. It wasn't the best team in the tournament. It simply played the best cricket that fortnight (Ghosh, 2017). West Indies was in its best form; they won the toss and selected to bowl first. In that era, they were considered to be a team with world's best bowling attack. They bowled well and India got all – out at 183 runs, which was not a good target against West Indies. India bowled extremely well, Amarnath and Madan Lal (3–31) each took three wickets, and one memorable moment was the sight of Kapil Dev running a great distance (about 18–20 yards) to take a catch to dismiss Richards, the West Indies top scorer and regarded as one of the greatest batsmen of all time ("1983 Cricket World Cup Final", 2018). And at the end India won the cup.

2017 ICC CHAMPIONS TROPHY

At the time of ICC Champions Trophy, India was the world No 1 in all formats. And in the ICC Champions Trophy opening match of Group B India vs. Pakistan, Pakistan lost the first match by 124 runs, outclassed by India in all departments. But the results reversed in the final match of the tournament. Even though in ICC Champions Trophy finals India won the toss and opted to bowl first, it was Pakistan who turned out to be victorious at the end to clinch the trophy as they defeated India by a huge margin of 180 runs (Alter, 2018).

REGRESSION EQUATION

Using regression coefficients (B) of the model shown in the table 7, the regression equation was developed which is as follows:

Logit = -4.580 + .018 (Team Score)

Odds = $e^{logit} = e^{-4.580+ .018\ (\text{Team Score})}$

P(Y) = $\frac{odds}{1+odds}$

The above regression equation can be used to predict the match outcome (i.e. Win/Loss) of the future ICC World Cup Cricket Matches on the basis of one predictor/ independent variables (i.e. Team Score) of the first innings data. It will only explain 36.3 % of variability in the dependent variable, the remaining percentage of the variability (63.7 %) may explain by some other variables.

CONCLUSION

The developed Logistic regression Model was found to be significant. According to the statistical significance of the predictor variables, they were numerically weighted and were used to predict the match outcome. Only one variable i.e. Team Score out of seven variables is selected by the model with coefficient of determination (R^2) of .272 (Cox & Snell) and .363 (Nagelkerke). 72.2 % of match results were correctly classified by the model.

ISSN 0976 - 1101
Vol-18, Special Issue (2)

REFERENCES

- 1983 Cricket World Cup Final. (2018). En.wikipedia.org. Retrieved 11 February 2018, from https://en.wikipedia.org/wiki/1983_Cricket_World_Cup_Final
- Alter, J. (2017). How Pakistan won the ICC Champions Trophy - Times of India. The Times of India. Retrieved 11 February 2018, from https://timesofindia.indiatimes.com/sports/cricket/champions-trophy-2017/top-stories/how-pakistan-won-the-icc-champions-trophy/articleshow/59214643.cms
- Bagchi, A., & Raizada, S. (2015).Anthropometric and Physical Variables as Predictors of Off-Spin Performance in Cricket: A Multiple Regression Study. International Journal of Sports Sciences & Fitness, 5 (2), 314 – 322.
- Bewick V, Cheek L, & Ball J. (2005). Statistics review 14: logistic regression. Crit Care, 9 (1), 112–8
- Christie, C. J. & King, G. A. (2008). Heart rate and perceived strain during batting in a warm and cool environment. International Journal of Fitness, 4, 33- 38.
- Cricket Records | Records | ICC Cricket World Cup, 2010/11 | | Match results | ESPNcricinfo. (2018). Cricinfo. Retrieved 1 January 2018, from http://stats.espncricinfo.com/icc_cricket_worldcup2011/engine/records/team/match_results.html?id=4857;type=tournament