31
Applications of Logistic Regression to Shots at Goal in Association Football

Jake Ensum[1], Richard Pollard[1] and Samuel Taylor[2]
[1]California Polytechnic State University, San Luis Obispo, California
[2]Research Institute for Sport and Exercise Sciences, Liverpool John Moores University

1. INTRODUCTION

Pollard and Reep (1997) introduced the use of logistic regression to analyse shots at goal in soccer. This approach allowed researchers to investigate factors thought to affect the chance of scoring from a shot. Factors identified as significant were utilised to quantify a shooting chance by estimating a shot's scoring probability. These shot probabilities were then used as an outcome measure to quantify the effectiveness of playing strategies and to construct contour diagrams for scoring probability. The following study expands upon Pollard and Reep's approach. More factors are analysed to investigate their importance in scoring from shots and shot probabilities are used as a comparative tool to evaluate shooting performance of 2002 World Cup semi-final teams (the quality of chance a team creates/concedes and the individual efficiency in scoring/saving these chances). Some limitations of logistic regression analysis are discussed and some directions for future work are suggested.

2. METHODS

Altogether, 1099 shots and 117 goals from 48 matches in the 2002 World Cup were recorded by five observers. For each shot, 9 factors were entered into Excel. Factors comprised those at the moment of the shot (definitions 1–5 (Table 1)) and those preceding the moment of the shot (6–9). A shot was defined as "an attempt to propel the ball into the opposition's goal".

Inter-reliability studies were conducted on 2.5 matches. These revealed that the shooting technique of kicked shots and the event preceding the shot (other than a cross) were not consistently recorded by observers and were rejected from the analysis. Other factors excluded were the number of passes involved in the movement, the match time and the match position (whether a team was winning, losing or drawing) as they were unsuitable for logistic regression. As in Pollard and Reep's study, shots blocked within 1 m of their origin were not included in the analysis. Shots taken directly from free kicks and penalties were excluded due to their infrequency and differing characteristics from other types of shot (both are closed rather than open skills). In total 729 kicked shots with

77 goals and 163 headed shots with 27 goals remained. Due to the technical differences between kicked and headed shots, both were analysed separately. The GLIM program was used to perform the logistic regression analyses. Semi-final teams' performances were used for evaluation purposes as they provided the most data.

Table 1. Definitions of variables.

	Variable	Definition
1.	Distance	The distance between the shot position and the nearest goal-post.
2.	Angle	The angle between the shot position and the nearest goal-post.
3.	Space	The shot-taker had>1 m of space from an opposition player.
4.	GK position	The goalkeeper was positioned between the shot-taker and goal.
5.	# players	The number of outfield players between the shot-taker and goal.
6.	SP origin	The movement leading to the shot originated from a dead ball.
7.	RP area	The pitch position from which possession is regained, divided into attacking, midfield and defending thirds.
8.	Cross	A pass directed towards the opposition's penalty box from a wide area (this is from the by-line to 2 m inside the penalty box).
9.	# touches	The number of touches completed by the player to take the shot, including the final touch to shoot.

3. RESULTS AND DISCUSSION

3.1. Investigation and quantification of factors

An initial logistic regression analysis showed that no second order interactions were significant. Subsequent analyses therefore omitted interaction terms.

The factors of 'distance', 'angle', 'cross', 'space' and '# players' all had a significant effect (where $P<0.05$) on the success of a kicked shot (Table 2). For headed shots, only 'distance' and 'space' were significant factors (Table 3).

Table 2. Significant factors for kicked shots.

Variable	Coefficient	P value	Odds Ratio
Distance	−0.115	0.001	0.89
Angle	−0.021	0.003	0.98
Cross	0.695	0.038	2.00
Space	0.734	0.019	2.08
# players	−0.311	0.004	0.73

Table 3. Significant factors for headed shots.

Variable	Coefficient	Pvalue	Odds Ratio
Distance	−0.522	0.000	0.59
Space	1.140	0.019	3.13

3.1.1. Significant factors

For most significant factors, their importance to shot scoring is self-evident and will not be discussed further. However, the 'cross' factor warrants further examination. Whilst the odds of scoring increase when a shot originates from a cross, this does not mean that crosses are the most effective means of penetrating a defence; many fail to meet their target. Principally, a cross changes the angle and area of play more than any other type of preceding event, thus placing greater attentional and decision making demands on defending players. For the defending players, attention has to take account of, and change to, this new area of play. Attention may also need to be directed towards the multiple points at which the ball might be met across the goal. Consequently, defenders and goalkeepers may have less time to organise and pick up cues that help them to anticipate play and there may be more space for shot-takers. For goalkeepers, the situation may also require them to move across the goal so that they have momentum in one direction. Therefore, they may not be well balanced to save in all potential directions. A further development in examining crosses would be to compare cross types. Partridge and Franks (1989, a and b) provided an interesting analysis of this question.

3.1.2. Non-significant factors

It was unexpected that 'GK position' was not significant, perhaps due to the small number of instances where the goalkeeper was not in the goal. Most factors preceding the shot were not significant. This result emphasises the importance of measuring factors as close as possible to the moment of the shot. In addition, the '# touches' factor may have been multi-modal and thus unsuitable for logistic regression as the data would violate the assumption of a sigmoidal relationship with the dependent variable (a goal).

3.1.3. Quantification of factors

Whilst most of the significant results might be expected, another value of logistic regression is to quantify factors, so informing coaches how important these factors are. The odds ratios in Tables 2 and 3 indicate this. For instance, for kicked shots, for each yard (0.9 m) the shot is taken away from goal, the odds of scoring reduce by 11 %, or, if a player has space, the odds of scoring double. A coach may choose to use these figures to motivate players and aid their understanding.

When comparing kicked and headed shots, 'distance' and 'space' become more important for headed shots. This may be due to headers generally having less power and their technique being impeded more easily through contact than for kicked shots.

3.2. Comparison with Pollard and Reep's (1997) study

Pollard and Reep's results of the 1986 World Cup are provided in Table 4. The '# touches' was recorded as a dichotomous variable of one touch or greater than one touch. The 'angle' measurement was originally reported in radians. This has been converted into degrees for ease of comparison. The '# touches' was not analysed (NA) for headed shots as all but one shot involved one touch.

Table 4. Pollard and Reep's results (1997).

Sample	Constant	Distance	Angle	Space	SP Origin	# Touches
410 kicked shots	1.245	–0.219	−0.026	0.947	−1.069	NS
74 headed shots	1.520	−0.237	−0.052	NS	−1.784	NA

Direct comparison is not wholly appropriate as different combinations of factors were analysed. Nevertheless, for kicked shots, both studies have broad agreement in their findings of significant factors, bar the factor of whether the movement originated from a dead ball ('SP origin'). It may be that as set plays are usually associated with an increased number of players advancing or retreating into attacking or defensive positions, the 2002 sample may more accurately account for the negative effect 'SP origin' has on scoring from shots by recording the number of players in front of goal.

For headed shots, 'distance' was significant for both samples. The other discrepancies may best be explained by the small sample size.

When comparing coefficient values for kicked shots, 'angle' and 'space' were similar. However, for the 2002 sample, 'distance' was half that of the 1986 sample (−0.113 compared to −0.219 (Tables 2 and 4)). This observation is hard to account for. It may be due to improvements in accuracy and power in 2002 and/or climatic locational differences. Less consistent recording of shots from distance in the 2002 sample is also possible, due to the number of observers involved. It may also indicate that more data are required for the values to stabilise.

For headed shots, coefficient values varied between samples but are not discussed further due to the small sample size. In general, analysis of all factors would have benefitted from greater reliability testing (no intra-reliability tests were conducted and only 2.5 matches were tested for inter-reliability whilst the data were recorded by 5 observers). Also, given the inconsistent findings between the two samples, it would be inappropriate to make generalisations for universal soccer performance at this stage.

3.3. Calculation of shot scoring probabilities

Despite the inconsistency of the results, calculating shot scoring probabilities may still provide useful information for this particular sample and their method of calculation and their applications are of interest.

The equations for calculating the 'y' value of kicked (1) and headed (2) shots are outlined below where 'y' is the dependent variable, 'D' represents the distance factor, 'A' angle, 'C' cross, 'S' space and 'P' the number of players.

that are not accurately accounted for. For instance, it might be expected that there are certain thresholds at which 'distance' deviates from this relationship, i.e. when players change technique to increase accuracy or power when closer or further away from goal, or, at a certain distance where a goalkeeper's reaction time is not quick enough to react to the shot.

4.2. Inclusion of all relevant factors

If all relevant factors are not analysed, a significant predictor's variance may be wrongly attributed as it may be shared with the variance of factors that are not included. Equally, some factors cannot be appropriately analysed, as above, which suggests that there is always the possibility for error when analysing shots at goal with logistic regression. In addition, some factors may not be included in the analysis as further work is required to identify all the factors that are significant to scoring from shots and some of these factors may be difficult to record, e.g. the goalkeeper's anticipation speed.

4.3. Net effect

The coefficients for each factor give a net effect. In some situations factors can have both positive or negative effects on scoring from a shot. For example, the number of players in front of goal would reduce the amount of the goal area available although, in some instances, they might also unsight the goalkeeper.

5. FUTURE WORK

Given the inconsistency in findings between the two samples studied and the utility of the information which logistic regression could provide, it would be desirable to record an appropriate sample in terms of validity and reliability for elite soccer. This would enable the reliable quantification of factors and comparison of performance between competitions, teams and players. With this sample, it might not be necessary for more practical match analysts (who wish to assume the study's reliability and validity or are unable to collect a large enough sample size) to collate and analyse data for logistic regression. The process could be 'short-cut' by using the study's coefficient values with the chosen sample so that performance could be compared as desired.

6. CONCLUSION

Despite the issues of reliability, sample size and the problems inherent in logistic regression, the present report has demonstrated some useful applications of logistic regression to shots at goal in soccer. Factors important to scoring goals can be identified and quantified and the probability of a shot scoring estimated. The calculation of shot scoring probabilities allows evaluation of the effectiveness of playing strategies, construction of scoring probability contours and assessment of performance associated with shooting. In this latter regard, the value of measuring shot-scoring probabilities over