World Cup Possession

Ethan Turner

2022-12-20

INTORDUCTION AND GOALS:

This study explores the effect of possession and types of possession on team performance using the most recent World Cup (in association football). We will attempt to answer the following questions:

1. If a team possesses more of the ball on average, do they progress further in the tournament?
2. If a team possesses more of the ball on average in attacking areas, do they progress further in the tournament?
3. If a team has a greater success rate with dribbling, do they progress further in the tournament?

METHOD:

This study is conducted using the software R, and its methods of linear regression analysis. The dataset used is from fbref.com.

The progression of a team in the tournament is measured in this dataset by the amount of 90-minute matches played. By this system, 3 corresponds to teams eliminated at the first round, then 4 at the next, and so forth until 7 for both finalists and both participants in the Third Place Playoff.

One complication with this dataset is that matches that advance to extra time after scores that are level after 90 minutes slightly inflate the amount of time spent on the pitch by each team. Thus, any conclusion made using the amount of total time played as indicative of the progress of a team through the tournament must be viewed with some caution.

RESULTS:

library(readr)

First, I examine the average percentage of possession for each team as a predicting variable for the number of 90-minute periods played (used here as an indicator of success in the competition).

Before this, here is a sample of the data that will be shown for all of the 32 World Cup teams.

WCP = read.csv("https://raw.githubusercontent.com/ethantur/World-Cup-Possession/main/WorldCupPossession.csv")  
head(WCP)

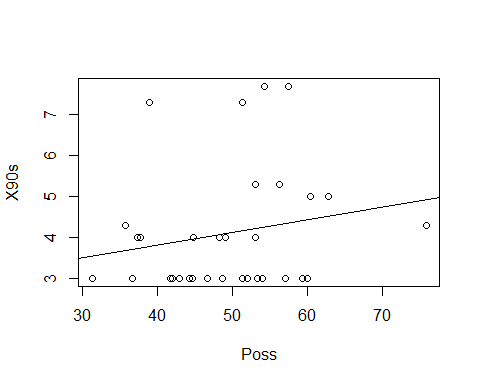
## Squad X..Pl Poss X90s Touches.Touches Touches.Def.Pen Touches.Def.3rd  
## 1 ar Argentina 24 57.4 7.7 5388 352 1424  
## 2 au Australia 20 37.8 4.0 2155 306 882  
## 3 be Belgium 20 57.0 3.0 2172 251 828  
## 4 br Brazil 26 56.2 5.3 3772 254 938  
## 5 cm Cameroon 22 41.7 3.0 1571 226 604  
## 6 ca Canada 19 52.0 3.0 1889 155 517  
## Touches.Mid.3rd Touches.Att.3rd Touches.Att.Pen Touches.Live Dribbles.Succ  
## 1 2716 1293 157 5383 45  
## 2 910 378 46 2155 13  
## 3 995 369 52 2172 19  
## 4 1839 1020 162 3771 32  
## 5 666 324 50 1571 24  
## 6 946 441 56 1889 31  
## Dribbles.Att Dribbles.Succ. Dribbles.Mis Dribbles.Dis Receiving.Rec  
## 1 112 40.2 99 81 3868  
## 2 46 28.3 61 36 1236  
## 3 45 42.2 47 24 1579  
## 4 110 29.1 88 58 2718  
## 5 55 43.6 39 24 943  
## 6 82 37.8 52 30 1289  
## Receiving.Prog  
## 1 210  
## 2 65  
## 3 95  
## 4 186  
## 5 58  
## 6 83

As we can see, each team is classified by over a dozen measurements: number of players used, average percentage of possession, number of 90-minute periods completed, total ball touches, total touches in each penalty area and third of the pitch, live-ball touches, number of successful dribbles, number of attempted dribbles, dribble success rate, miscontrols, dispossessions, passes received, and progressive passes received.

simplepossmod = lm(X90s~Poss, data = WCP)  
summary(simplepossmod)

##   
## Call:  
## lm(formula = X90s ~ Poss, data = WCP)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.4263 -1.0306 -0.5769 0.5057 3.5228   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.57167 1.43593 1.791 0.0834 .  
## Poss 0.03091 0.02854 1.083 0.2874   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.503 on 30 degrees of freedom  
## Multiple R-squared: 0.03763, Adjusted R-squared: 0.005549   
## F-statistic: 1.173 on 1 and 30 DF, p-value: 0.2874

plot(X90s~Poss, data = WCP)  
abline(simplepossmod)



In this situation, we conduct a hypothesis test of the efficacy of ball possession as a predictor of progress through the tournament. The null hypothesis is that the coefficient of average possession percentage in this predictive model is zero. The alternative hypothesis is that the coefficient of average possession percentage in this predictive model is nonzero.

With a p-value of .2874 > .05, we conclude by failing to reject the null hypothesis. We have no significant evidence that the created model predicting the amount of 90-minute matches played by World Cup teams has a nonzero coefficient of Average Possession Percentage.

By plotting the explanatory variable, the progression in the tournament, by the predictor of possession, we see that there is not a clear linear relationship on display. Overall, there is no obvious indication that a greater percentage of ball possession over their games at the World Cup leads to progress further in the tournament.

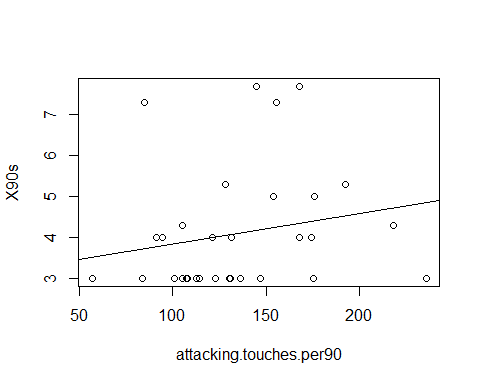
There are still more subliminal pieces of this dataset, however, which might prove useful. Perhaps the amount of possession, or touches of the ball, in purely attacking areas might prove to be a better predictor.

To conduct this line of investigation, I create a new variable to measure the amount of touches in the attacking third per 90 minutes for each team. This alleviates the problem of using total touches over the course of the tournament as an indicator of success, since a team playing seven games total would clearly be expected to log more total touches than a team playing three. We are instead trying to judge whether there is a connection the average amount of touches a team had in attacking areas over the tournament and how deeply the team traveled into the final stages.

attacking.touches.per90 = (WCP$Touches.Att.3rd / WCP$X90s)  
  
attackmod = lm(X90s~attacking.touches.per90, data = WCP)  
summary(attackmod)

##   
## Call:  
## lm(formula = X90s ~ attacking.touches.per90, data = WCP)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.8512 -0.9666 -0.4709 0.4608 3.5707   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.101857 0.951110 3.261 0.00276 \*\*  
## attacking.touches.per90 0.007423 0.006792 1.093 0.28313   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.502 on 30 degrees of freedom  
## Multiple R-squared: 0.03829, Adjusted R-squared: 0.006234   
## F-statistic: 1.194 on 1 and 30 DF, p-value: 0.2831

plot(X90s~attacking.touches.per90, data = WCP)  
abline(attackmod)



Here, we conduct a hypothesis test of the efficacy of touches in the attacking third as a predictor of progress through the tournament.

The null hypothesis is that the coefficient of touches in the attacking third in the above model is zero, while the alternative hypothesis is that this same coefficient is nonzero.

From the model’s summary output, a p-value of 0.28313 indicates that we should not reject the null hypothesis, and we lack significant evidence that touches in the attacking third has a nonzero coefficient in this model.

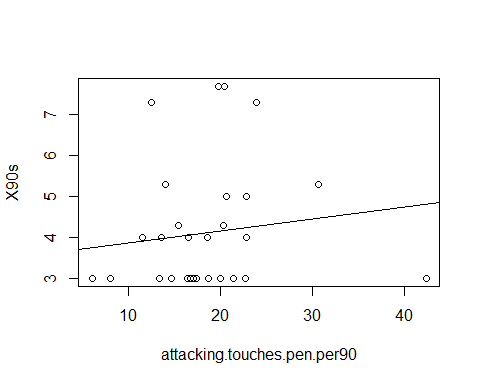
Plotting the data, and seeing a lack of obvious linear relationship, supports this conclusion.

After failing to establish the concrete predictive efficacy of touches in the attacking third, I will attempt to narrow the investigation further by including touches only in the attacking penalty area.

attacking.touches.pen.per90 = (WCP$Touches.Att.Pen / WCP$X90s)  
  
attackpenmod = lm(X90s~attacking.touches.pen.per90, data = WCP)  
summary(attackpenmod)

##   
## Call:  
## lm(formula = X90s ~ attacking.touches.pen.per90, data = WCP)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.8146 -1.0602 -0.4902 0.4012 3.5516   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.56625 0.79952 4.460 0.000106 \*\*\*  
## attacking.touches.pen.per90 0.02949 0.04161 0.709 0.483941   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.519 on 30 degrees of freedom  
## Multiple R-squared: 0.01647, Adjusted R-squared: -0.01632   
## F-statistic: 0.5024 on 1 and 30 DF, p-value: 0.4839

plot(X90s~attacking.touches.pen.per90, data = WCP)  
abline(attackpenmod)



The null hypothesis is that the coefficient of touches in the attacking penalty area in the above model is zero, while the alternative hypothesis is that this same coefficient is nonzero.

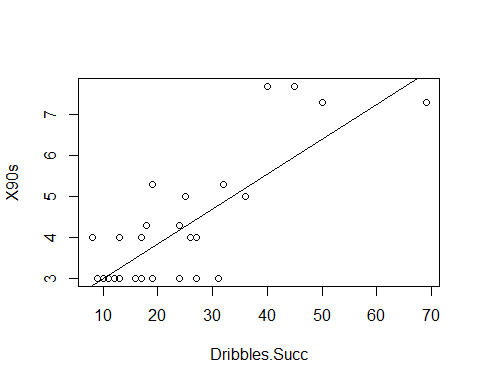
From the model’s summary output, a p-value of 0.483941 indicates that we should not reject the null hypothesis, and we lack significant evidence that touches in the attacking penalty area has a nonzero coefficient in this model.

Now, I will depart slightly from the purely possession-based statistics and advance to dribble success rate, a measurement linked exclusively with territorial advancement on the pitch. As previously, I am attempting to find a predictor for the amount of 90-minute matches played by World Cup teams.

dribblemod = lm(X90s~Dribbles.Succ, data = WCP)  
summary(dribblemod)

##   
## Call:  
## lm(formula = X90s ~ Dribbles.Succ, data = WCP)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.78059 -0.61984 -0.02285 0.65902 2.14772   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.12255 0.33730 6.293 6.17e-07 \*\*\*  
## Dribbles.Succ 0.08574 0.01263 6.788 1.58e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9619 on 30 degrees of freedom  
## Multiple R-squared: 0.6057, Adjusted R-squared: 0.5925   
## F-statistic: 46.08 on 1 and 30 DF, p-value: 1.577e-07

plot(X90s~Dribbles.Succ, data = WCP)  
abline(dribblemod)



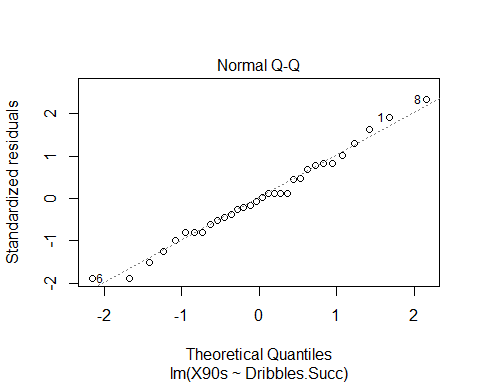
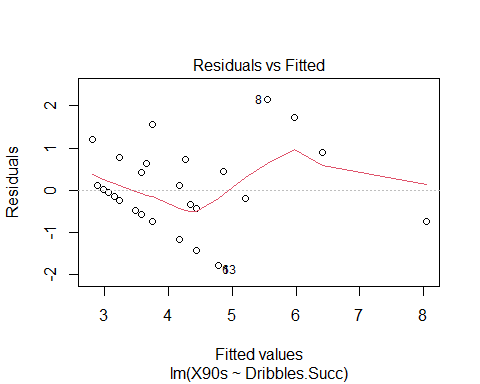
The null hypothesis is that the coefficient of dribble success rate in the above model is zero, while the alternative hypothesis is that this same coefficient is nonzero.

Using a hypothesis test from the summary output, we obtain a p-value of 1.58e-07 < .05. We reject the null hypothesis with significant evidence that the coefficient of dribble success rate as a predictor for matches completed is nonzero. From the graph of these two variables, we also notice a positive linear relationship.

From the model, we can expect for a team’s progression through the tournament (by number of 90-minute games played) to increase by .08574 for each additional percentage in dribbling success rate. With the given intercept of 2.12255, we can project a team with an average dribbling success rate of around 10% to be eliminated at the group stage after 3 matches, while a team of around 57% would be projected to be a semi-finalist or finalist.

We will now examine the conditions of linear models to see if our created predictive model for progress in the World Cup meets the general linear model requirements.

plot(dribblemod, 1:2)



In evaluating for the conditions of linear models, we run into a few areas of concern. In a single tournament, it might be difficult to assert the independence of the data that we have extracted to make our conclusions over the significance of possession and attacking possession at the World Cup. Factors such as the location of the tournament or the time of year could influence the general trends of play.

Away from this, the distribution of residuals does appear to be Normal and centered at zero. The spread is not quite uniform, but the relatively small sample size of 32 teams alleviates these concerns somewhat, although should be examined in further studies.

CONCLUSION:

This study is by no means an endpoint in the investigation into the utility of prioritizing ball possession in high-level association football. Although we have established a connection within the given data from the 2022 World Cup between dribbling success rate and advancement within the competition (and failed to do so for possession in general and possession in attacking areas), studies in more competitions would yield far more concrete results. Limiting our findings to the scope of this study, teams succeeding in carrying the ball forward appears to carry more importance than simply keeping the ball, even in attacking areas. Whether or not the failure of teams who dominated possession but succeeded with a lesser rate of dribbles is evidenced in a larger trend of the sport or an anomaly in this edition of the World Cup should be further examined.

WORKS CITED:

“2022 World Cup Possession Stats”. fbref.com. Accessed December 18 2022.  
<https://fbref.com/en/comps/1/possession/World-Cup-Stats>.