RAMAKRISHNA MISSION RESIDENTIAL COLLEGE

(AUTONOMOUS), NARENDRAPUR,

KOLKATA-700103

A STATISTICAL PROJECT WORK

NAME: *SAHIL MALLICK*

ROLL NO: *STUG/024/18*

YEAR: *2020-2021*

TOPIC

A comparative study between the forward lines of 3 of the biggest football clubs in the World.

INTRODUCTION

Football is one of the most loved and most watched sports around the world. Nearly 1 billion people watched the world cup final 2019. There are nearly 250 million registered football players all over the world and moreover 5 million registered referees and nearly 500 million people are working for this game. Football has the highest global television audience in sport.

In many parts of world football evokes great passion and plays an important role in the life of individual fans, local communities, and even nations. *R. Kapuscinsky* says that the people who are polite, modest or even humble fall easily into rage with playing or watching soccer games. *The Côte d'Ivoire* national team helped secure a truce to the nation's civil war in 2006 and it helped further reduce tensions between the government and rebel forces in 2007 by playing a match in the rebel capital of Bouake`, an occasion that brought both armies together peacefully for the first time. By contrast, football is widely considered to have been the final proximate cause for the Football War in June 1969 between El Salvador and Honduras.

In the history of football there are some greatest footballers of all time in the world whose name will be in the hearts of the football fans forever i.e., *Pele*, *Diego Maradona, Ronaldo Nazario, Kaka, Puyol, Messi, C. Ronaldo, Ronaldinho, Iniesta, Xavi* etc.. Their names will be remembered by the football fans until the death of the football.

A football match is being controlled on the field by referees, linesmen and VAR (introduced in 2019). Off the field the football matches are controlled by various boards like **IFAB, FIFA, UEFA**, and boards run by various nations.

At a professional level, most matches produce only a few goals. For example, the 2005-06 season of the English Premier League produced an average of 2.48 goals per match. Broadly a team includes four main categories: strikers, whose main task is to score goals; defenders, who specialise in preventing the opponents from scoring; midfielders, who dispossesses the opposition and keep possession of the ball in order to pass it to the forwards on their team; and goalkeeper, whose job is to prevent the opponent from

scoring goals. The layout of a team is known as *formation* and the team's formation and tactics is usually the prerogative of the team's *manager*.

In the 21st century, for the football world the club football is the most important. Football fans are crazy for the club they support. Whenever a club wins, their fans are the happiest person at that moment. For the win the most important part is to score goals more than the opponent. Various times we see that a team with more ball possession scores more goals. By contrast we also can see the exactly opposite scenario, i.e. one team dominates the ball while other team scores the goal and take the winning points. Like Mourinho, a great manager in the new world says "*You take the ball, I will take three points*". Scoring goals is mainly the job of the forward line (strikers) of a team.

So as a student of statistics**, I shall compare the forward lines of 3 of the biggest football clubs in the World based on conversion rate of a team.**

I used the program "Microsoft Excel" to do my required computation.

STRATEGY

For testing purpose I have collected data on conversion rate (%) and ball possession (%) for 10 matches each for 3 best teams of Europe. These data are collected on the basis of entire 2020-2021 season.

SAMPLING METHOD

The new football era is dominated by European football clubs. Among the football clubs of Europe I chose three best teams as per performance of 2020-2021 season as I am interested in the comparison of these best clubs' performances. The clubs I choose are ***CHELSEA F.C., F.C. BARCELONA*** & ***MANCHESTER CITY F.C.*** Then I have randomly chosen 10 matches played by each team in the season randomly using SRSWOR among 40 matches. Then I have collected data of conversion rate (%) & ball possession (%) for all the matches chosen.

These data are available on http://www.footballdatabase.com//data,

http://www.en.wikipedia.org//la-liga-2020-2021,
http://www.premierleague.com ,

https://1xbet.whoscored.com/Statistics

CONCEPT

Linear Regression:

Linear Regression analysis is a statistical method used to predict the value of a dependent variable based on the values of dependent variables.

The value being predicted is termed dependent variable because its outcome or values depends on the behaviour of other variables. This independent variable generally ascertained from the population or sample.

Here we consider 'Y' as the dependent variable defining the percentage of conversion rate of a team in a match, and the independent variable X defining the percentage of ball possession throughout the match by the team.

According to our given data our model is:

$Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij}$, i=1(1)3, j=1(1)10,

Where,

Y_{ij} = value of the conversion rate (%) of the jth chosen match of the ith chosen team.

α_i = y intercept.

β_i = regression co-efficient of the ith chosen team.

X_{ij} = value of the ball possession (%) possessed in the jth chosen match played by ith chosen team.

ε_{ij} = Random error.

Now we are going to test the following hypothesis

H01: ß1=0 **ag** H11:ß1 not equal to zero.

H02: ß2=0 **ag** H12: ß2 not equal to zero.

H03: ß3 = 0 **ag** H13: ß3 not equal to zero.

Then we are going to find a linear regression equation of Y with the variables whose coefficients reject its corresponding null hypothesis.

Now our gathered information for the matches of the three clubs Chelsea F.C., F.C. Barcelona &Manchester City F.C. are as follows:

CHELSEA F.C.:

TABLE NO. 1

Match No.	Conversion rate (%) (Y1)	Ball possession(%)(X1)
1	50	68
2	20	66
3	20	44
4	40	52
5	33.3	65
6	11.1	43
7	16.6	49
8	16.6	44
9	20	47
10	28.6	55

F.C. BARCELONA:

TABLE NO. 2

Match no.	Conversion rate (%) (Y2)	Ball possession (%) (X2)
1	33.3	68
2	25	66
3	60	73
4	42.8	55
5	25	59
6	40	60
7	83	75
8	28.5	51
9	25	48
10	20	49

MANCHESTER CITY F.C.:

TABLE NO. 3

Match no.	Conversion rate (%) (Y3)	Ball possession (%) (X3)
1	50	62
2	20	51
3	66	83
4	20	44
5	14.5	51
6	33.3	37
7	40	44
8	16.6	47
9	40	70
10	25	48

ANALYSIS OF OUR MODEL

Here our model is,

$Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij}$, i=1(1)3, j=1(1)10,

We assume εij ~ N(0, σ^2) for all i & j where σ^2 is unknown.

The least square estimates are

$\beta i^* = S_{xy(i)} / S_{xx(i)}$

$\alpha i^* = \overline{y_i} - \bar{x} * \beta i^*$

$S_{xy(i)} = \sum_j (Xij - \overline{Xio})(Yij - \overline{Yio})$ and $S_{xx(i)} = \sum_j (Xij - \overline{Xio})^2$, i=1,2,3

SSE= $\sum_i (Syy(i) - \beta i^{*^2} * S_{xx(i)})$ and MSE=SSE/d.f.

We define

$ti = \beta i^* / \sqrt{MSE/sxx(i)}$

ti ~ t_8 (under Hoi)

We reject Hoi at α level of significance if | ti|>$t_{\frac{\alpha}{2};8}$,i=1,2,3

RESULTS

THE LINEAR REGRESSION SUMMARY:

- TABLE 1 SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.675833739
R Square	0.456751242
Adjusted R Square	0.388845147
Standard Error	9.52959942
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	610.8298792	610.8298792	6.72621867	0.031939087
Residual	8	726.5061208	90.8132651		
Total	9	1337.336			

	Coefficients	*Standard Error*	*t Stat*
Intercept	-19.40203247	17.6192193	-1.101185708
ball possession	0.844691041	0.325695979	2.593495454

1) Here we can see that the value of the t-stat (t1) is greater than the value $t_{\frac{\alpha}{2};8}$=2.301 . So we reject the null hypothesis H01 at 5% level of significance & we can say that conversion rate of forward line is dependent on the ball possession of that team CHELSEA F.C..

- The regression line is $\mathbf{Y_{1j}}$ **= -19.402+0.845** $\mathbf{X_{1j}}$

- TABLE 2 SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.734031203
R Square	0.538801806
Adjusted R Square	0.481152032
Standard Error	14.16144414
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1874.331999	1874.331999	9.346121707	0.01565034
Residual	8	1604.372001	200.5465002		
Total	9	3478.704			

	Coefficients	*Standard Error*	*t Stat*
Intercept	-50.68116613	29.4355538	-1.721767033
ball possession	1.472535863	0.481670629	3.057142736

1) Here we can see that the value of the t-stat (t2) is greater than the value $t_{\frac{\alpha}{2};8}$=2.301. So we reject the null hypothesis H01 at 5% level of significance and we can say that conversion rate of forward line is dependent on the ball possession of that team BARCELONA F.C..

- The regression line is $\boldsymbol{Y_{2j} = -50.681 + 1.4725X_{2j}}$

➢ TABLE 3 SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.72612975
R Square	0.527264414
Adjusted R Square	0.468172466
Standard Error	12.12322351
Observations	10

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1311.403614	1311.403614	8.922779322	0.01741073
Residual	8	1175.780386	146.9725482		
Total	9	2487.184			

	Coefficients	*Standard Error*	*t Stat*
Intercept	-13.91826951	16.01848013	-0.868888271
ball possession	0.865144684	0.289626748	2.987102161

1) Here we can see that the value of the t-stat (t3) is greater than the value $t_{\frac{\alpha}{2};8}$=2.301. So we reject the null hypothesis H01 at 5% level of significance and we can say that conversion rate of forward line is dependent on the ball possession of that team MANCHESTER CITY F.C. .

- The regression line is $\boldsymbol{Y_{3j} = -13.918 + 0.865X_{3j}}$

We come to know that conversion rate of a forward line of a team is dependent on the ball possession of that team. Now our target is to compare the forward lines of the three teams. For this we have to test whether the regression lines are identical or not. We pool all the three samples to test this.

Our model is same as the above:

$Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij}$, i=1(1)3, j=1(1)10,

$\varepsilon_{ij} \sim N(0, \sigma^2)$ where σ^2 is unknown.

We are going to test the following hypothesis

Ho: α1=α2=α3 & β1=β2=β3 ag H1 : Ho is not true

Now , SSE = $\sum_{i=1}^{3}(S_{yy(i)} - \beta i *^2 * S_{xx(i)})$

Syy(i) =$\sum_j (Y_{ij} - \overline{Y_{io}})^2$

Sxx(i) =$\sum_j (X_{ij} - \overline{X_{io}})^2$

d.f. = 30-6 =24

Under H0 our model is

Y_{ij} = α0+β0 *X_{ij}+εij , εij $\sim N(0, \sigma^2)$ where σ^2 is unknown .

Where α0 , β0 are the common values of α and β respectively.

SSEH0 =Syy -$\widehat{\beta_0}^2$*Sxx

Where,

Syy=$\sum_i \sum_j (Y_{ij} - \overline{Y_{oo}})^2$

Sxx=$\sum_i \sum_j (X_{ij} - \overline{X_{oo}})^2$

d.f. = 30-2=28

SSHo = SSEHo-SSE, d.f. = 28-24=4

F= $(\frac{SSHo/4}{SSE/24})$; F~$F_{\alpha;4,24}$

Ho is rejected iff $F_{obs} > F_{\alpha;4,24}$

Here from the Excel sheet we get that

SSEHo= 3992.878 ; SSE=3506.658; SSHo =486.2196 ; F_{obs}=0.8319

Here $F_{obs} < F_{0.05;4,24}$=2.78

So our null hypothesis is accepted, i.e. , the regression lines for the three teams are identical.

Calculation procedure :

I performed the analysis and calculations using the software MS Excel. Given the dataset, I proceeded in the following way:

1) Click on Excel options present on the toolbar. Click on add ins.

2) Click on Data Analysis toolpack under the section application add ins.

3) Click on “Data” on the quick access toolbar.

4) Select “Regression” under the section “analysis tools”.

5) Select the column containing y values corresponding to “input y values” and select the column containing X value under the section “input x values”.

6) Click OK.

7) We are doing the same thing for Table 1, 2, 3 simultaneously.

MS Excel provides us with the required summary data which shall be used for our analysis.

CONCLUSION

From this statistical study on the data of 3 of the biggest football clubs of the World I have got many useful & important conclusions stated as follows:

- *The conversion rate of forward line of a team is dependent on the ball possession neglecting other midfield effects.*

  So for any coach of a team who wants that his team scores more, must look into the ball possession possessed by his team.

- There is also an interesting fact for these three teams is stated below:

  If these three teams possesses ball during the match equally then all the conversion rate of the forward lines of these three teams will more or less equal. So I can say based on my study that none of these three teams is best. All are equally capable of scoring goal based on their ball possession.

BIBLIOGRAPHY

1. Krugman, Paul R; Obstfeld, M.; Melitz, Marc J.(2012) . *International Economics: Theory and Policy* (9th global ed.). Harlow: Pearson.
2. Draper, N.R.; Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). John Wiley.
3. Jolliffe, Ian T. (1982). *"A Note on the Use of Principal Components in Regression"*. Journal of the Royal Statistical Society, Series C. 31 (3): 300-302.
4. Rohatgi V.K., Saleh A.M.D. (2001): *An Introduction to probability and Statistics.*
5. Critical value of t and F distribution source: http://www.stattrek.com/online-calculator

ACKNOWLEDGEMENT

I would like to express my sincere thanks and gratitude to my professors of the Department of statistics, Ramakrishna Mission Residential College Narendrapur, Dr. Dilip Kumar Sahoo, Dr Parthasarothi Chakrabarti, Sri Tulsidas Mukherjee, Sri Shubhadeep Banerjee, Sri Palash Pal for encouraging me to explore different topics for the project work and helping me in each step to complete it.

I would also like to thank my parents, classmates and seniors who effortlessly helped me whenever I needed their help to complete the project.

Sahil Mallick