The Frontiers of Society, Science and Technology
ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

How Soccer Players' Box Score Statistics Effect on their Rating and Market Value

Haoyang Yan

Shanghaitech University Shanghai 201210 China

ABSTRACT. *This paper discusses the application of advance statistics in soccer. In order to valuate the performance and value of soccer players, we introduce more advance statistics and try to detect the effect on players' rating and market value from their box score statics.*

In section 2 of this paper, we introduce an advance statistic - Box Rating (BR) - to valuate the performance of a soccer player in a match. The derivation of BR comes from the rating given by Whoscored regressed by 26 series of box score statistics (including goals, assists, passes and so on) of 2024 players in European top 5 leagues last 3 competition seasons (2017-2020). We then analyse the ridge trace of the regression coefficient to test the robustness of the regression.

In section 3 of this paper, to avoid the confounders like age, injury and commercial value which have unobserved effect to players' market value, we use 35 series of player ability ratings from FIFA20 as the In- strument Variable and do two stage least squre (2SLS), getting the true effect players' box score statistics exerting on their market value.

In section 4 of this paper, we compute the correlation among ratings and market value, and do some hypotheses-test of our results. We also point out some limitation of players' statistics.

KEYWORDS: *box score statistics, rating, market value*

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

1 Introduction

Advance statistics are widely used in professional sports leagues to valuate players' performance. Based on detailed data recording and high-score matches, professional leagues of baseball, hockey and basketball have introduced plenty kinds of advanced statistics to valuate players' performance comprehensivly. Adjusted Plus-Minus (APM) in National Hockey League (NHL) [1] [2] and Regularized Adjusted Plus-Minus (RAPM) in National Basketball Association

(NBA) [3] [4] [5]are two of the representative examples of advanced statistics. The common characteristic of APM and RAPM is that they are both derived on the basis of plus-minus, which means how many scores the team win/lose when the player is on the field, independent with what the player actually do on the field, but can generally accurately valuate how strong the player effect the match.

However, when applied to soccer, statistics from plus-minus like APM and RAPM face difficult problems to solve [6] [7]. Firstly, the score of soccer match is too small, which means the sample size is not large enough to insure a low- level stochasticity. Secondly, the first string of a soccer team is relatively fixed and there are only 3 substitution quota for a team in one match. These two factors make plus-minus statistics full of stochasticity and hard to reflect soccer players' real contribution to matches.

The advance statistics used in soccer are mostly one-sided. There has been developed advance statistic to predict expected goal which valuate offence chances and quality of shoots [8] [9], advance statistic to valuate players from their actions [10], advance statistic to valuate performance of goalkeeper [11], advance statistcs to compare the team performance between competitions home and away [12].

There is still one way to valuate players' general performance on-field. To valuate the contribution of basketball players without large-scale sample and abundant plus-minus advance statistics, Daniel Myers develop the Box Plus/Mi- nus (BPM), which is a advance statistic regressing to the Regularized Adjusted Plus Minus (RAPM). [13] Through the regression function, we can easily gain BPM of players with simple box score statistics, avoiding the limitation of im- mature statistics in previous time. This method can also be used in soccer. Although there is no meaningful plus-minus statistic like RAPM for soccer, we can use the rating from media to replace.

Valuating players' personal market value is a more complex problem. Not only on-field performance, but also other factors such as injury, age, personal background effect their market value [14] [15]. Thus, detecting the true effect players' box score statistics exerting on their market value is a challenging work for us.

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

2. BR - valuate players' performance by box score statistics

2.1 preprocessing and cleaning

In this part, we use web crawler to obtain the box score and rating data from whoscored and do some preprocessing and cleaning to these data. www.whoscored.com is a unique website and one of the fastest growing in the sports industry, spe- cializing in the in-depth analysis of detailed soccer data.

Firstly, we filter out data of soccer players in the European top 5 league (England Premier League, Spanish La Liga, Italian Serie A, German Bundesliga,

French Ligue 1) last 3 competition seasons (2017-2020). Secondly, we filter out that a player played more than 5 matches in a competition season to avoid the error of small size sample. Thirdly, we match the players' name to the players' name in the dataset of FIFA2020, cleaning up players who cannot match (for players with the same name, we use his height and weight to confirm). After that, we get 4718 rows of data of 2024 different players in last 3 years.

In the vertical direction of the data, we reserve 1 column of players' average rating in this season and 11 columns of players' essential information, including name, season name, tournament name, team name, age, height, weight, posi- tion, appearance, substitute on, minutes played. After some preprocessing and cleaning, we reserve 26 columns of box score, which are all players' average box score statistics of every match in a season. For those statistics which have mu- tiple items, such as total, won, lose, accurate, inaccurate, we only reserve the

total and rate to valuate players' yield and efficiency. What's more, we intro- duce the shot adjusted distance to estimate players' average shot distance, which

equals to 1 * shotsixyardboxrate + 2 * shotpenaltyarearate + 3 * shotoutboxrate. Detailed information of these boxscore is listed on Table 1.

2.2 linear regression

In order to detect the relationship between rating and 26 series of box score statistics, we use multiple linear regression to regress the rating by box socre statistics. The predicted value of rating, rating hat, is introduced as Boxscore Rating (BR).

The regression formula and result is as follow:

$$BR = \hat{rating} = \beta_0 + \sum_{i}^{26} \beta_i * boxscore \quad (1)$$

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

$$i=1$$

Table 1: 26 series of box score statistics and their (new) regression coefficients β_i , (γ_i) to rating

i	$boxscore_i$	β_i	γ_i
1	goal	1.0223	1.0568
2	assist	0.6939	0.7434
3	key pass	0.1356	0.1399
4	turnover	(0.0584)	-
5	dispossessed	(0.0422)	-
6	clearance	0.0387	0.0797
7	save	0.2059	0.2312
8	tackle won	0.1021	0.1117
9	interception	0.0895	0.1306
10	shots total	0.0547	0.0440
11	shot adjusted distance	(0.0121)	-
12	shot on target rate	0.0637	0.0499
13	penalty taken	(0.5345)	(0.5595)
14	pass long	(0.0053)	-
15	pass long accurate rate	0.0776	0.1932
16	pass short	0.0044	-
17	pass short accurate rate	0.1597	-
18	dribble	0.0935	0.0536
19	dribble rate	0.0908	0.1212
20	duel aerial	0.0311	-
21	duel aerial rate	0.2022	-
22	goal own	(0.6545)	(0.5834)
23	foul given	0.0365	0.0346
24	foul committed	(0.0246)	-
25	yellow card	(0.1549)	(0.1738)
26	red card		

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

	(1.1921)	
	(1.2305)	0
	constant	
	5.6716	
	5.8121	
R^2	0.8915	0.8299
$\bar{R}^2$	0.8909	0.8290

It is obvious that goals and assists are two of the most influential factors to players' rating. Besides, turnover dispossessed, penalty taken (this is reasonable because the goals' value become less when a player take penalty), goal own, yellow and red card do negative effect on rating.

The goodness of fit of the linear regression 1 R2 = 0.8915 and the adjusted

R-squred R2 = 0.8909 show good fitness, but we still need further test to certify

this multiple linear regression model is suited to explain the relationship between rating and box score.

2.3 ridge regression

To estimate the autocorrelation among these box score statistics, we need to analyse the trace of their ridge regression. Before doing ridge regression, we linearly normalize all the box score into the range[0,1]. Then we use ridge regression

$$\beta(k) = (boxscoreT * boxscore + k * I)^{-1} * boxscoreT * rating \quad (2)$$

where the ridge coefficient k [0, 10] , I is the identity matrix, and β(k) is the normalized regression coefficient under ridge coefficient k.

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

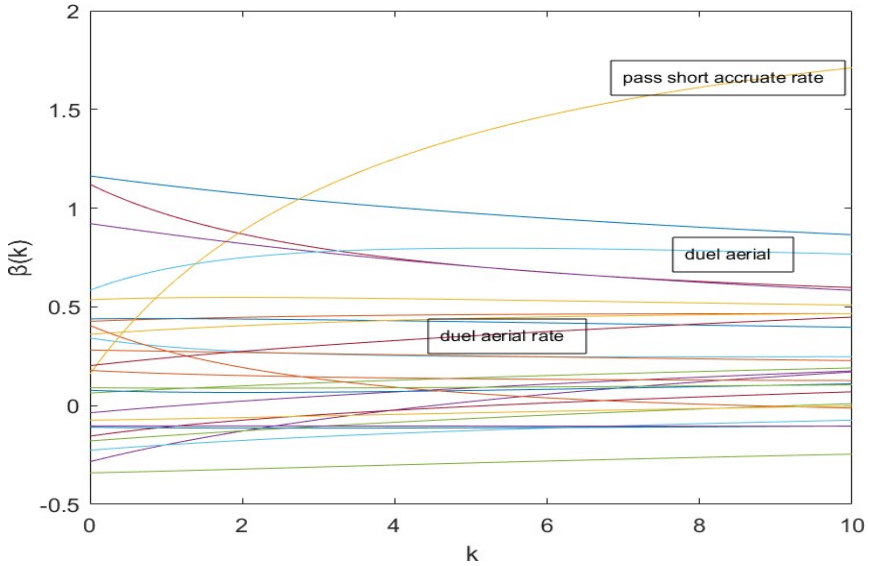
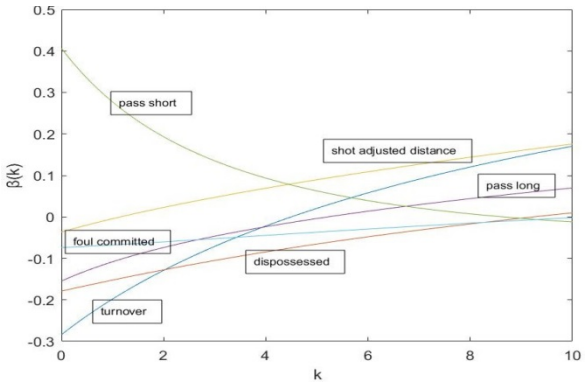


Figure 1: ridge trace of coeficients (after normalizing) (without β_0)

In the ridge trace figure, there are 3 series of regression coefficients - pass short accurate rate, duel aerial, duel aerial rate - are obviously abnormal, which increse when the ridge coefficient k increase.



ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

What's more, the absolute values of 6 series of regression coefficients - turnover, dispossessed, shot adjusted distance, pass long, pass short, foul com- mitted - have been less than 0.05 in range[0,10].

Figure 2: ridge trace of 6 series of regression coefficients less than 0.05 (after normalizing)

This abnormality implies that these series of boxscore may not satisfy the linear relation with the rating or not stable to be significant. So we delete these 9 series of box score statistics, using other 17 series to do linear regression. The new regression coefficients γi are showed in table 1, with $R2 = 0.8299$ and

$\bar{R}^2$

$R = 0.8290$.

3. how players' box score statistics effect their market value

3.1 choosing of instrument variable

In this section, we want to detect the true effect players' box score statistics exerting on their market value. However, there are many confounders correlated with both box score and market value, which may exaggerate or shrink this effect. Here are some examples, younger players are usually more valuable with similar boxscore and ability because they are more potential and may work for the team much longer; injury players do not appear much and are not in good body conditions, which means their box score may not be impressive, while their market value will not decrease that much; players whose position are forward may be well-known to spectators and have higher commercial value.

In order to remove the interference by these confounders, we use the ability ratings of players in FIFA20 as the instrument variable, which is only correlated to players' performance on field, but not to any other factors . From the instru- ment variable and two stage least square, we can seperate the endogenous and exogenous part of box score statistics, which reveal the true effect players' box score statistics exerting on their market value.

FIFA20 is one of the most popular soccer simulation video game published by Electronic Arts as part of the FIFA series. The ability rating in FIFA20 gives 35 different kinds of ratings of each player's ability, including pace, shooting, passing and so on.

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

3.2 first stage regression

In the first stage regression, we use 35 series of ability ratings to regress 26 series of box score statistics. Because the ratings of ability are mostly aggregated in range[60,100], we use the log-linear regression to fit the curve better, and regularize boxscore into range[1,e] and ability into range[0,1]. The coeficient δij and the goodness of fitting of the log-linear regression is showed in table 2.

$$log(\widehat{boxscore_j}) = \delta_{0j} + \sum_{i=1}^{35} \delta_{ij} * ability_i \tag{3}$$

$$boxscore_j = e^{\delta 0j + \sum_{i=1}^{35} \delta ij * ability i + uj}$$

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

Table 2: Regression coefficients δ_{ij} between ability ratings and box score statis- tics

δij	goal assist key-	tu rn -	dispo- tackle- shot-	clear- inter-	save shots-	s h ot -	penal ty-	p as s-	p as s-	p as s-	dribble dribble- duel-	d u el -	g o al -	fo ul -	fo ul -	yellow-	red-
	Pass	o v e r	ssess ed ance	Won	ception Total adjusted- on-	distan ce tar ge t- ra te	Tak en long	long - short	sh or t-	rate	Aerial aerial- Own	rate	Given Commit -Card	ted	Card		
								acc urat e- rat e	acc urat e- rat e								
-																	

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

pace	0.17	0.15	0.06	0.17	0.09	0.20	(0.42)	(0.01)	0.07	0.13	0.35	0.36	0.36	0.18	0.02	0.21	0.03	(0.06)	(0.26)	0.03	0.55	(0.20)	(0.40)	0.02
acceleration	(0.03)	(0.08)	(0.02)	0.02	(0.02)	0.14	0.15	(0.05)	0.03	(0.01)	(0.14)	(0.05)	(0.07)	0.04	(0.11)	(0.16)	0.01	0.06	(0.10)	(0.11)	(0.11)	0.00	0.02	0.07	0.19	(0.01)
sprint_speed	(0.05)	0.05	0.00	(0.05)	(0.01)	0.02	0.23	(0.07)	(0.12)	(0.04)	(0.21)	(0.13)	0.02	0.22	0.02	(0.02)	0.00	0.01	0.10	0.04	0.25	(0.03)	(0.14)	(0.09)	0.04	0.03
shooting	0.30	0.22	0.06	(0.12)	(0.13)	0.04	(1.04)	(0.05)	(0.02)	0.01	0.16	0.50	0.07	(1.73)	(0.35)	0.06	(0.22)	(0.29)	0.36	0.26	(0.82)	(0.11)	(0.30)	(0.21)	(0.28)	(0.37)
positioning	0.03	0.03	0.07	0.12	0.15	(0.34)	0.08	0.06	(0.11)	(0.03)	(0.07)	0.01	0.17	(0.12)	(0.04)	(0.15)	(0.08)	0.00	(0.23)	0.02	0.07	(0.04)	0.07	0.19	0.06	(0.03)
ftnishing	0.02	(0.07)	(0.06)	0.20	0.14	(0.14)	0.46	(0.06)	(0.06)	(0.01)	(0.30)	(0.02)	(0.03)	0.63	0.11	0.21)	0.05	0.18	(0.19)	(0.03)	0.26	0.04	0.17	0.09	0.02	0.09
shot_power	0.13)	(0.13)	(0.08)	(0.04)	(0.06)	(0.07)	0.19	(0.02)	(0.05)	(0.07)	0.01	(0.02)	(0.08)	0.35	0.01	0.19)	0.04	0.00	(0.17)	(0.06)	0.14	0.08	(0.06)	(0.02)	0.01	0.06
long_shots	(0.08)	(0.06)	0.02	0.04	0.05	(0.07)	0.22	(0.03)	0.03	0.04	0.02	(0.24)	(0.01)	0.36	0.13	0.05)	0.04	0.06	0.02)	(0.13)	0.18)	(0.03)	0.07	0.12	0.19	0.12
volleys	0.06	0.01	0.03	0.11	0.11	(0.09)	0.18	(0.04)	(0.03)	0.02	(0.18)	(0.05)	(0.04)	0.07	0.09	0.02	0.02	(0.03)	(0.14)	(0.03)	0.17	0.03	0.04	0.02	0.00	0.02
penalties	0.03	(0.02)	0.04	0.03	0.03	0.22	0.23	0.02	0.01	(0.02)	0.03	0.04	0.25	0.11	0.00)	0.02	(0.00)	(0.03)	0.03)	0.13	0.11	0.17	0.08	0.00	0.02	
passing	0.10)	0.00	0.11)	(0.28)	(0.10)	0.24)	0.29	0.05	0.03	(0.08)	0.09	0.16)	0.11	0.05	0.36	0.70	0.27	(0.14)	0.67)	0.13)	0.24)	0.18	0.02)	0.33)	0.35)	0.04)
vision	0.04	0.07	0.00	0.00	0.00	0.16	(0.01)	0.00	(0.13)	0.00	(0.05)	(0.16)	(0.02)	(0.04)	0.02	(0.00)	(0.00)	0.05	0.20	0.00	0.00	0.09	0.10	(0.00)	(0.00)	(0.00)

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

	)		2	7	9	)	)		0)		)	)	)	)	)	6)	4)		0	2	1	)	)	3)	9)	0)						
crossing	(0.08)	0.17)	0.2	0.0	(0.03)	0.1	(0.06)	0.10	0.08	(0.08)	(0.03)	0.08	0.0	0.0	(0.37)	(0.1	0.1	0.08	0.07)	0.03	0.1	0.01	0.01	(0.0	0.06	(0.0					
freekick	0.09	0.13	0.17	0.1	0.1	0.16	(0.02)	0.03	0.09	0.13	0.05	0.1	0.5	0.1	0.16	0.06	0.1	0.1	0.10	0.2	0.0	0.0	0.08	0.01	0.15	0.01					
short_pass	0.02	(0.12)	(0.05)	0.2	0.1	(0.13)	(0.23)	0.06	0.03	(0.1)	0.2	0.1	0.3	0.2	(0.04)	(0.35)	(0.01)	(0.03)	(0.2)	(0.3)	0.9	0.1	)	)	)	0.13	0.08	0.02			
long_pass	(0.03)	0.01)	(0.02)	(0.05)	0.1	0.10	(0.05)	(0.04)	0.06	0.06	(0.03)	0.2	0.2	0.2	0.2	(0.03)	0.02	0.09	0.20	0.05	0.5	0.1	0.0	0.0	0.02	0.05	0.01				
curve	0.02	0.04	0.04	(0.02)	(0.05)	0.7	0.1	(0.09)	(0.08)	(0.01)	0.1	0.0	0.1	0.6	(0.04)	(0.04)	(0.0)	0.0	0.1	0.02)	0.2	0.3	(0.04)	(0.10)	(0.0)	(0.08)	(0.08)	0.3			
ovr_dribbeling	(0.10)	(0.15)	(0.20)	0.0	(0.03)	(0.08)	(0.08)	(0.1)	(0.0)	(0.07)	(0.24)	(0.54)	(0.03)	0.8	0.0	(0.4)	0.00	0.02	(0.06)	(0.16)	(0.37)	(0.20)	0.0	0.36	0.77	0.06					
agility	(0.07)	(0.1)	(0.14)	(0.01)	0.0	0.0	(0.04)	0.05	(0.0)	(0.0)	0.0	(0.07)	(0.02)	(0.14)	0.0	0.7	(0.0)	0.00	(0.0)	0.0	(0.01)	(0.09)	0.0	0.1	0.04	0.04	(0.01)				
balance	(0.03)	0.01)	0.0	(0.01)	0.0	(0.11)	(0.07)	0.05	(0.0)	(0.0)	0.1	0.1	0.01)	(0.13)	0.06	(0.0)	(0.0)	0.00	0.0	(0.08)	(0.20)	(0.00)	0.0	(0.0)	(0.1)	(0.3)	(0.1)				
reactions	0.49	0.45)	0.2	(0.07)	(0.05)	0.4	0.6	(0.1)	0.16	0.33	(0.08)	0.2	0.6	0.0	0.2	0.6	0.40	0.01	0.02	0.08)	0.4	(0.04)	0.0	0.6	0.0	(0.2)	(0.1)	(0.0)			
ball_control	0.01	0.06)	0.6	0.0	0.0	0.0	(0.07)	(0.04)	0.01	0.18	0.4	0.0	0.5	0.02)	0.16)	0.2	0.3	14.0)	0.09	0.6	0.3	0.9	0.2	0.1	0.1	0.07)	0.01)	(0.01)			
dribbling	(0.03)	0.04)	0.1	0.2	0.3	0.15	0.3	0.23	0.03	(0.01)	0.0	0.1	0.3	0.05	0.54	0.1	0.12	0.0	0.37	0.1	0.0	0.2	0.7	0.02	0.1	0.9	(0.0)	(0.3)	(0.9)	(0.0)	
comp	0.	0.	0.	(0.	(0.	0.	(0.	(0	(0	0.	(0.	0.	0.	(0.	0.	0.	0.	0.	0.	(0.	0.	(0.	0.	(0	(0	0.					

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

osure	1	09	0	07	04	0	02	.1	.1	02	20	1	0	03	1	25	11	05	3	08	0	02	0	.2	.1	0
	2		8	)	)	9	)	3)	1)		)	2	7	)	6				1	)	7	)	1	2)	4)	0
defen	(0.	(0	0.	(0.	(0.	0.	0.	0.	0.	(0	(0.	(0.	0.	1.	(0.	(0	0.	(0	1.	(0.	0.	0.	(0.	(0	0.	0.
se	25	.1	0	36	25	2	4	34	75	.3	47	16	0	0	03	.0	09	.0	6	37	6	2	19	.0	25	3
	)	5)	8	)	)	8	2			4)	)	)	5	9	)	7)		4)	0	)	8	3	)	2)		3
interc	(0.	0.	(0.	(0.	(0.	(0.	(0.	0.	0.	(0	0.	(0.	(0.	(0.	0.	0.	0.	(0	(0.	0.	(0.	(0.	(0.	0.	0.	(0
eption	00	03	03	03	07	04	04	10	12	.0	0	06	03	13	1	19	06	.1	29	0	03	03	02	03	03	.0
s	)		)	)	)	)	)			0)	2	)	)	)	8			0)	)	3	)	)	)			6)
headi	0.	(0	(0.	0.	(0.	0.	(0.	(0	(0	0.	(0.	0.	0.	(0.	(0.	0.	(0	(0	(0.	0.	0.	0.	(0.	0.	(0	(0
ng	1	.0	14	0	06	2	09	.0	.0	08	06	1	0	10	08	03	.0	.1	20	3	0	0	03	01	.0	.0
	6	5)	)	5	)	3	)	8)	1)		)	3	0	)	)		4)	2)	)	2	3	3	)		1)	1)
aware	(0.	(0	(0.	(0.	(0.	0.	(0.	(0	(0	0.	0.	0.	(0.	(0.	(0.	0.	0.	(0	(0.	0.	(0.	(0.	0.	(0	(0	(0
nesss	05	.0	06	03	04	0	15	.0	.1	02	1	0	02	29	01	05	01	.0	42	0	18	05	0	.0	.1	.0
	)	3)	)	)	)	2	)	6)	3)		7	3	)	)	)			9)	)	5	)	)	3	6)	2)	9)
stand	0.	0.	0.	0.	0.	(0.	(0.	0.	(0	0.	0.	0.	(0.	(0.	(0.	0.	0.	(0	0.	(0.	0.	(0.	(0.	0.	0.	(0	(0
_tack	0	05	0	2	2	04	15	02	.1	04	2	0	11	35	03	19	.0	10	49	2	18	07	0	03	.1	.1
le	0		1	0	3	)	)		7)		4	5	)	)	)		1)		)	0	)	)	6		0)	1)
slide	(0.	(0	(0.	(0.	(0.	0.	(0.	0.	0.	(0	0.	(0.	(0.	(0.	(0.	(0	(0	(0	(0.	(0.	(0.	(0.	0.	0.	0.	(0
_tack	04	.0	09	11	12	0	04	15	10	.0	0	06	00	03	13	.0	.0	.0	14	02	02	00	0	05	08	.0
le	)	4)	)	)	)	9	)			8)	8	)	)	)	)	0)	0)	2)	)	)	)	)	3			3)
physi	0.	0.	0.	0.	0.	0.	(0.	0.	0.	0.	0.	0.	0.	(0.	(0.	(0	(0	0.	0.	0.	0.	0.	0.	0.	0.	0.
cal	1	07	0	2	1	5	39	04	14	24	5	3	0	34	07	.1	.2	05	3	5	0	0	1	35	09	0
	6		1	3	4	1	)				6	0	6	)	)	2)	6)		8	4	2	0	0			5
jumpi	(0.	(0	(0.	(0.	(0.	0.	0.	(0	0.	(0	(0.	(0.	0.	(0.	(0	(0	(0	(0	(0.	0.	0.	0.	(0.	(0	(0	(0
ng	05	.1	10	08	12	0	0	.0	03	.0	06	06	00	0	05	.0	.0	.1	07	0	1	0	10	.0	.0	.0
	)	0)	)	)	)	8	4	3)		4)	)	)	)	5	)	4)	1)	2)	)	9	2	1	)	0)	2)	1)
stami	0.	0.	0.	0.	0.	(0.	0.	0.	0.	0.	(0.	(0.	0.	0.	(0.	0.	0.	0.	(0.	(0.	0.	(0.	0.	0.	0.	(0
na	0	12	1	1	1	21	1	33	15	08	11	13	0	1	00	18	04	04	10	03	0	02	1	21	13	.0
	4		9	5	1	)	2				)	)	3	3	)				)	)	4	)	2			6)
streng	(0.	(0	(0.	(0.	0.	(0.	0.	(0	(0	(0	(0.	(0.	(0.	0.	0.	(0	0.	0.	(0.	(0.	0.	(0.	(0.	(0	(0	(0
th	10	.1	11	07	0	17	0	.1	.0	.1	15	15	03	0	0	.1	06	01	06	15	0	01	04	.0	.1	.0
	)	2)	)	)	2	)	7	2)	8)	0)	)	)	)	6	6	0)			)	)	7	)	)	7)	0)	2)
aggre	(0.	(0	(0.	0.	0.	(0.	0.	0.	(0	(0	(0.	(0.	0.	0.	(0.	(0	(0	(0	(0.	(0.	(0.	(0.	0.	0.	0.	0.
ssion	02	.0	07	0	0	16	0	02	.1	.0	12	09	0	0	04	.1	.0	.0	10	06	05	03	1	21	31	0
	)	7)	)	3	2	)	7		4)	5)	)	)	0	2	)	2)	1)	2)	)	)	)	)	2			6

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

const ant	(0. 18)	(0 .17)	(0. 10)	0. 0 4	(0. 03)	0. 2 0	0. 3 5	0. 11	0. 10	(0 .11)	0. 3 7	0. 0 5	(0. 09)	0. 3 6	0. 3 9	(0 .03)	0. 88	(0 .07)	0. 3 3	0. 0 9	0. 6 9	0. 0 2	0. 0 1	0. 14	0. 19	0. 0 4
$R2$	0. 4 7	0. 30	0. 5 3	0. 5 8	0. 5 3	0. 7 3	0. 9 1	0. 56	0. 64	0. 58	0. 6 6	0. 2 9	0. 1 9	0. 7 4	0. 1 5	0. 58	0. 59	0. 50	0. 2 5	0. 5 2	0. 5 3	0. 0 6	0. 2 4	0. 37	0. 20	0. 0 3
$R2$	0. 4 6	0. 29	0. 5 3	0. 5 8	0. 5 3	0. 7 3	0. 9 1	0. 56	0. 63	0. 58	0. 6 6	0. 2 9	0. 1 9	0. 7 4	0. 1 5	0. 58	0. 58	0. 50	0. 2 4	0. 5 1	0. 5 2	0. 0 5	0. 2 4	0. 36	0. 20	0. 0 2

Some of the coeficients are easy to explain, like the high positive correlation between the box score of pass-short (-accruate-rate) and the ability of passing. Some are not that comprehensible, like the high positive correlation between the box score of pass-long and the ability of defense. That is beacuse players with higher defensing ability are usually posited in backfield and have more chance to give long pass.

Some of these regression do not have significant goodness of fit, especially the regression to the box score statistics of penalty-taken, goal-own, yellow-card and red-card. Probably because the sample is too small or these box score statistics are mostly happened accidental.

3.3second stage regression

In this part, we use web crawler to obtain players' market value from transfer market. www.transfermarkt.co.uk is a website which has footballing information, such as scores, results, statistics, transfer news, and fixtures. Its valuation to player's market value is widely accepted among social media. The data is showed in Great Britain Pound £.

Same as the process we did in section 2.1, we match names of players who has the market value data to names already in our dataset. 3896 rows are matced and others are cleaned.

We use the coefficients δij in section 3.2 to predict the boxsˆcore. Then we do

the second stage regression, which regress the predicting boxsˆcore to the market

value in equation 4, where ζj is the regression coeficient and $(\sum_{j=1}^{26} \zeta_j * uj + \varepsilon)$

is the disturbing term. For comparision, we also do a linear regression without instrument variable, where ηj is the regression coeficient and v is the disturbing term in equation 5.

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

$$value = \zeta_0 + \sum_{j=1}^{26} \zeta_j * boxscore_j + \varepsilon = \zeta_0 + \sum_{j=1}^{26} \zeta_j * \widehat{boxscore}_j + (\sum_{j=1}^{2} \zeta_j * u_j + \varepsilon) \quad (4)$$

$$value = \eta_0 + \sum_{j=1}^{26} \eta_j * boxscore_j + v \quad (5)$$

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

Table 3: (two stage) regression coeficient (ζ_j) , η_j from box score statistics to market value

j	$boxscore_j$	ζ_j	η_j
1	goal	9.57E+07	3.69E+07
2	assist	2.63E+08	2.48E+07
3	key pass	-3.23E+07	-2.33E+06
4	turnover	5.14E+06	-1.15E+06
5	dispossessed	1.75E+07	-7.29E+05
6	clearance	1.47E+07	-1.82E+05
7	save	4.44E+07	4.09E+05
8	tackle won	5.74E+06	1.17E+05
9	interception	-2.47E+07	-2.03E+06
10	shots total	4.42E+06	3.12E+06
11	shot adjusted distance	7.98E+06	-1.84E+06
12	shot on target rate	4.71E+07	6.98E+05
13	penalty taken	-2.03E+08	-2.80E+07
14	pass long	-3.57E+06	-1.45E+05
15	pass long accurate rate	-3.45E+07	3.57E+06
1	pass short	1.11E+	3.73E+0

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

6		06	5
17	pass short accurate rate	-3.63E+07	1.81E+07
18	dribble	-3.15E+06	3.65E+06
19	dribble rate	7.83E+07	3.53E+06
20	duel aerial	-1.07E+07	-4.06E+03
21	duel aerial rate	1.92E+07	7.38E+06
22	goal own	6.68E+08	-1.18E+07
23	foul given	-1.94E+07	-1.15E+06
24	foul committed	3.48E+07	-3.10E+06
25	yellow card	-4.97E+07	-3.84E+06
26	red card	-6.53E+08	-1.37E+07
0	constant	-5.48E+07	-1.87E+07
	R^2	0.3230	0.3556
	$\bar{R}^2$	0.3183	0.3511

In table 3, some of the coeficients ζj (more specifically, penalty taken, goal own and red card) are abnormally large due to their small sample. What's more, coeficients of two stage regression ζj are mostly larger than coeficients ŋj without instrument variable. This reveal that we may have underestimated the effect players' box score statistics exerting on their market value because of those confounders. Giving an example, a player's market value will increase about £100,000,000 (that's the market value of Lionel Messi, one of the most famous soccer player in the world) if he score one more goal in every competition.

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

4.some property and hypothesis

4.1correlation among ratings and value

For each of the 3896 rows, we extract 4 series of data - the average rating of the player this year given by whoscored, the Boxscore Rating (BR) derived from the box score statistics, the overall ability rating given by FIFA20 (that's the weighted average ability rating of players) and the market value given by transfermarket. We compute the correlation coeficients among these 4 series of data.

correlation coeficient	rating	Boxscore Rating	overall ability	market value
rating	1.000	0.944	0.518	0.489
Boxscore Rating	0.944	1.000	0.474	0.462
overall ability	0.518	0.474	1.000	0.573
market value	0.489	0.462	0.573	1.000

Those correlation coeficients reveal that compared with rating and boxscore in the field, the ability of players is more highly correlated to their market value, probably because rating and boxscore can not reflect all the effect a player exert to the competition and the team.

3.2 are younger players more valuable?

In section 3.1 we refer that younger players are usually more valuable with similar boxscore and ability because they are more potential and may work for the team much longer. In this part we want to test whether this hypothesis is convincing.

We regress the overall ability rating and the age of players to their market value with multiple linear regression, then we structure the T-statistic of the regression coeficient α2 from age to value.

$$\hat{value} = \alpha_0 + \alpha_1 * overallability + \alpha_2 * age \quad (6)$$

$$\frac{\hat{\alpha}_2}{} = \frac{\hat{\alpha}_2}{}$$

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

$$T = \frac{\hat{\sigma}}{\sqrt{c_{22}}} = \frac{S_e}{\sqrt{c_{22}}\sqrt{n-3}} \sim t(n-3) \sim N(0,1) \quad (7)$$

where

c_{22} is the (3,3) element of matrix $(X^T X)^{-1}$

X is the matrix of independent variable [1, overallability, age]

$Se = \sum_{i=1}^{n} (value_i - \hat{value}_i)^2$

Because of the large sample, we can approximately regard the t-distribution as gauss distribution with mean 0 and standard deviation 1.

	constant	overallability	age		R^2
α	-1.28E+08	2.64E+06	-2.35E+06		0.5632
T-statistic			-40.33		

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

The result shows that a player's market value decrease about 2 million pounds when he get one year older whlie his ability stay stable. Due to the large absolute value of the T-statistic, we have more than 99.99% confidence ($p<0.0001$) to believe that there exists a nagetive correlation between players' market value and age.

3.3 limitation on valuating goalkeeper

In some aspect, the rating and box score statistics exist significant limi- tation on valuating players. This becomes apparently when valuating players from different positions, especially goalkeeper. To find the limitation, we divide players into two groups, 303 goalkeepers and 4415 non-goalkeepers, using Z-test to detect the difference

of rating and Boxscore Rating between these two group of players.

$$Z_{rating} = \frac{rating_{gk} - rating_{ngk}}{\sqrt{\frac{S_{gk}}{n_{gk}} + \frac{S_{ngk}}{n_{ngk}}}} = 4.798$$

$$Z_{BR} = \frac{BR_{gk} - BR_{ngk}}{\sqrt{\frac{S_{gk}}{n_{gk}} + \frac{S_{ngk}}{n_{ngk}}}} = -7.086 \quad (8)$$

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

From the Z-statistic we have more than 99.99% confidence ($p<0.0001$) to be- lieve that goalkeepers' rating and Boxscore Rating is less than non-goalkeepers'. For Boxscore Rating the difference is larger. As we all know, goalkeepers al- ways activate in penalty area and seldom have box score statistics. This means valuating players via box score statistics is not fair for goalkeepers.

References

[1] Brian Macdonald. Adjusted plus-minus for nhl players using ridge regres- sion with goals, shots, fenwick, and corsi. Journal of Quantitative Analysis in Sports, 8(3), 2012.

[2] Brian Macdonald. A regression-based adjusted plus-minus statistic for nhl players. Journal of Quantitative Analysis in Sports, 7(3), 2011.

[3] Shankar Ghimire, Justin A Ehrlich, and Shane D Sanders. Measuring in- dividual worker output in a complementary team setting: Does regularized adjusted plus minus isolate individual nba player contributions? PloS one, 15(8):e0237920, 2020.

[4] Jeremy Mertz, L Donald Hoover, Jean Marie Burke, David Bellar, M Lani Jones, Briana Leitzelar, and W Lawrence Judge. Ranking the greatest nba players: A sport metrics analysis. International Journal of Performance Analysis in Sport, 16(3):737–759, 2016.

[5] Jeremias Engelmann. A new player evaluation technique for players of the national basketball association (nba). In Proceedings of the MIT Sloan Sports Analytics Conference, 2011.

[6] Tarak Kharrat, Ian G McHale, and Javier López Peña. Plus–minus player ratings for soccer. European Journal of Operational Research, 283(2):726– 736, 2020.

[7] Francesca Matano, Lee F Richardson, Taylor Pospisil, Collin Eubanks, and Jining Qin. Augmenting adjusted plus-minus in soccer with fifa ratings. arXiv preprint arXiv:1810.08032, 2018.

[8] Masoomeh Zameni, Mahmood Fathy, and Amin Sadri. A low cost algorithm for expected goal events detection in broadcast soccer video. International Journal of Digital Content Technology and its Applications, 4(8):118–125, 2010.

[9] Andrew Rowlinson et al. Football shot quality: Visualizing the quality of soccer/football shots. 2020.

[10] Guiliang Liu, Yudong Luo, Oliver Schulte, and Tarak Kharrat. Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Mining and Knowledge Discovery, pages 1–29, 2020.

[11] Paul Power, Aditya Cherukumudi, Sujoy Ganguly, Felix Wei, Long Sha, Jennifer Hobbs, Hector Ruiz, and Patrick Lucey. Trading places–simulating goalkeeper performance using spatial & body-pose data, 2019.

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

[12] V Barnett and S Hilditch. The effect of an artificial pitch surface on home team performance in football (soccer). Journal of the Royal Statistical Society: Series A (Statistics in Society), 156(1):39–50, 1993.

appendix: code

```
rating = xlsread ('WhoScoredFinal7. xlsx ','D3:D4720');
boxs = xlsread ('WhoScoredFinal7. xlsx ','O3:AO4720');
beta = regress ( rating ,boxs);
BR = boxs * beta;
5
avgr = mean( rating );
SSR = sum((BR - avgr ).^2);
SST = sum(( rating - avgr ).^2);
R2 = SSR/SST;
10 ajR2 = 1 - (1 -R2)*(4718 -1)/(4718 -27 -1);
11
12 ma=[];
13 for i = 1:27
ma = [ma,max(boxs(: , i ))];
end
mam = repmat(ma, 4718 ,1);
norboxs = boxs./mam;
18
19 E = eye (27);
20 betak =[];
21 for k = 0:0.01:10
j = (norboxs '*norboxs + k*E)^(-1)*( norboxs ')* rating ;
betak = [betak , j ];
end
25 for i = 1:26
26 plot (0:0.01:10 , betak( i ,:))
hold on
end
saveas (1 , '01. jpg ')
hold o f f
31
32 thr = zeros (1 ,27);
33 for k = 1:27
```

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

```
34 34     for j = 1:1001
35 35     i f abs(betak(k, j )) < 0.05
36 36     thr (k) = 1;
37 end
38 end
39 end
40 40     for i = 1:26
41 41     i f thr ( i )==1
42 42     plot (0:0.01:10 , betak( i ,:))
43 hold on
44 end
45 end
46 saveas (1 , '02. jpg ')
47 hold o f f
48 48
49 49     ajboxs = boxs;
50 50     ajboxs (:,[4 ,5 ,11 ,14 ,16 ,17 ,20 ,21 ,24]) = 0;
51 gamma = regress ( rating , ajboxs );
52 nBR = ajboxs * gamma;
53 nSSR = sum((nBR - avgr ).^2);
54 SST = sum(( rating - avgr ).^2);
55 nR2 = nSSR/SST;
56 56 najR2 = 1 - (1 -nR2)*(4718 -1)/(4718 -27 -1);
57 57
58 abi = xlsread ('WhoScoredFinal7. xlsx ','AR3:CA4720');
59 abi = abi /100;
60 reguboxs = log (norboxs*(exp(1) -1)+1);
61 delta = [];
62 62     for i = 1:26
63 delta = [ delta , regress (reguboxs(:, i ), abi )];
64 end
65 reb = abi * delta ;
66 66 Rm = [];
67 67 ajRm = [];
68 68     for i = 1:26
69 avgnb = mean(reguboxs(:, i ));
70 SSR = sum(( reb(:, i ) - avgnb).^2);
71 SST = sum((reguboxs(:, i ) - avgnb).^2);
72 Rm = [Rm,SSR./SST];
73 73     ajRm = [ajRm,1 - (1 -SSR./SST)*(4718 -1)/(4718 -27 -1)];
74 74 end
75 75
```

ISSN 2616-7433 Vol. 2, Issue 15: 82-104, DOI: 10.25236/FSST.2020.021511

```
ratingnew = xlsread ('WhoScoredFinal8. xlsx ','D3:D3898');
BRnew = xlsread ('WhoScoredFinal8. xlsx ','C3:C3898');
overall = xlsread ('WhoScoredFinal8. xlsx ','AQ3:AQ3898');
age = xlsread ('WhoScoredFinal8. xlsx ','H3:H3898');
boxsnew = xlsread ('WhoScoredFinal8. xlsx ','O3:AO3898');
value  = xlsread ('WhoScoredFinal8. xlsx ','CC3:CC3898');
abinew  = xlsread ('WhoScoredFinal8. xlsx ','AR3:CA3898');
abinew = abinew/100;
mamm = repmat(ma,3896 ,1);
85 mamm(: ,27) = [];
preboxsnew = ((exp(abinew * delta )-1)/(exp(1)-1)).* mamm;
preboxsnew = [preboxsnew,ones (3896 ,1)];
zeta = regress ( value ,preboxsnew);
eta = regress ( value ,boxsnew);
90
prevalue1 = preboxsnew * zeta ;
prevalue2 = boxsnew * eta ;
avgv = mean( value );
SSR1 = sum(( prevalue1 - avgv).^2);
SSR2 = sum(( prevalue2 - avgv).^2);
SST = sum(( value - avgv).^2);
R21   = SSR1/SST;
R22  = SSR2/SST;
99    ajR21 = 1 - (1 -R21)*(3896 -1)/(3896 -27 -1);
      100       ajR22 = 1 - (1 -R22)*(3896 -1)/(3896 -27 -1);
      101
      cor = [ratingnew ,BRnew, overall , value ];
      corrcoe = zeros (4);
      for  i   = 1:4
      for  j   = 1:4
      corrcoe ( i , j ) = corr ( cor (:, i ), cor (:, j ));
      end
      end

      X = [ones (3896 ,1), overall ,age];
      111 XX = (X'*X)^( -1);
      112       c22 = XX(3 ,3);
      alpha = regress ( value ,X);
      a2 = alpha (3);
      prevalue = X * alpha ;
      Se = sum(( prevalue - avgv).^2);
      R20 = Se/SST;
```