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 Instrument Variable and do two stage least square (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
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 comprehensively. 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 statistics 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/Minus (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 immature 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.
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, specializing 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, position, 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 multiple 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 introduce the shot adjusted distance to estimate players’ average shot distance, which equals to 4 shotsixydanboxrate + 2 shotpenaltyareaerate + 3 shootoutboxrate. 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:

$$BR = rat_{hat} = \beta_0 + \beta_i * boxscore \hspace{1cm} (1)$$
Table 1: 26 series of box score statistics and their (new) regression coefficients

\[ \beta_i, (\gamma_i) \] to rating

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

26 red card
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 $R^2 = 0.8915$ and the adjusted $R$-squared $\tilde{R}^2 = 0.8299$ 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

$$\beta(k) = (\text{boxscore}^T \cdot \text{boxscore} + k \cdot I)^{-1} \cdot \text{boxscore}^T \cdot \text{rating}$$

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

\[ \begin{array}{c|cc}
1.1921 \\
1.2305 \\
\hline 
(1.2305) & 0 \\
\hline 
\text{constant} & 5.6716 \\
\text{5.8121} \\
\hline 
\end{array} \]

\[ \begin{array}{c|cc}
R^2 & 0.8915 & 0.8299 \\
\hline 
R & 0.8909 & 0.8290 \\
\hline 
\end{array} \]
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 increase when the ridge coefficient $k$ increase.
What’s more, the absolute values of 6 series of regression coefficients - turnover, dispossessed, shot adjusted distance, pass long, pass short, foul committed - 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 $\gamma_i$ are showed in table 1, with $R^2 = 0.8299$ and $\bar{R} = 0.8290$.

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 instrument variable and two stage least square, we can separate 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.
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 coefficient $\delta_{ij}$ and the goodness of fitting of the log-linear regression is showed in table 2.

\[
\log(\text{boxscore}_j) = \delta_0 + \sum_{i=1}^{35} \delta_{ij} \times \text{ability}_i
\]

\[
\text{boxscore}_j = e^{\delta_0 + \sum_{i=1}^{35} \delta_{ij} \times \text{ability}_i} + u_j
\]
Table 2: Regression coefficients $\delta_{ij}$ between ability ratings and box score statistics

<table>
<thead>
<tr>
<th>$\delta_{ij}$</th>
<th>goal</th>
<th>tu</th>
<th>dispo-</th>
<th>clear-</th>
<th>save</th>
<th>s</th>
<th>penal</th>
<th>p</th>
<th>p</th>
<th>dribble</th>
<th>d</th>
<th>g</th>
<th>fo</th>
<th>fo</th>
<th>yellow-</th>
<th>red-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>assist</td>
<td>rn</td>
<td>tackle-</td>
<td>inter-</td>
<td>shots-</td>
<td>h</td>
<td>ty-</td>
<td>as</td>
<td>as</td>
<td>dribble-</td>
<td>u</td>
<td>o</td>
<td>ul</td>
<td>ul</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>key-</td>
<td>shot-</td>
<td>ot</td>
<td>pass-</td>
<td>s</td>
<td>s</td>
<td>duel-</td>
<td>el</td>
<td>al</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>assessed</td>
<td>Won</td>
<td>conception</td>
<td>ed</td>
<td>Total</td>
<td>en</td>
<td>or</td>
<td>aerial-</td>
<td>Commit</td>
<td>Given</td>
<td>Card</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>adjusted-</td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>distance</td>
<td>acc</td>
<td>acc</td>
<td>rate</td>
<td>rate</td>
<td>ted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ce</td>
<td>urat</td>
<td>urat</td>
<td>e</td>
<td>e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tar</td>
<td>e</td>
<td>e</td>
<td>ge</td>
<td>rat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>t-</td>
<td>e</td>
<td>e</td>
<td>ra</td>
<td>rat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>te</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Frontiers of Society, Science and Technology

Published by Francis Academic Press, UK

-92-
Some of the coefficients are easy to explain, like the high positive correlation between the box score of pass-short (accurate-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 because players with higher defending ability are usually positioned 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.

### 3.3 second 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 matched and others are cleaned.

We use the coefficients $\delta_{ij}$ in section 3.2 to predict the box score. Then we do the second stage regression, which regress the predicting boxs’ core to the market value in equation 4, where $\zeta_j$ is the regression coefficient and $\left(\sum_{j=1}^{26} \zeta_j * u_j + \varepsilon\right)$ is the disturbing term. For comparison, we also do a linear regression without instrument variable, where $\eta_j$ is the regression coefficient and $v$ is the disturbing term in equation 5.
\begin{align}
\sum_{j=1}^{26} \text{value} &= \zeta_0 + \zeta_j \cdot \text{boxscore}_j + \varepsilon \\
&= \zeta_0 + \sum_{j=1}^{26} \zeta_j \cdot \hat{\text{boxscore}}_j + \left( \sum_{j=1}^{26} \zeta_j \cdot u_j + \varepsilon \right) \\
\sum_{j=1}^{26} \text{value} &= \eta_0 + \eta_j \cdot \text{boxscore}_j + v
\end{align}
Table 3: (two stage) regression coefficient ($\zeta_j$), $\eta_j$ from box score statistics to market value

<table>
<thead>
<tr>
<th>$j$</th>
<th>boxscore</th>
<th>$\zeta_j$</th>
<th>$\eta_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>goal</td>
<td>9.57E+0</td>
<td>3.69E+0</td>
</tr>
<tr>
<td>2</td>
<td>assist</td>
<td>2.63E+0</td>
<td>2.48E+0</td>
</tr>
<tr>
<td>3</td>
<td>key pass</td>
<td>3.23E+0</td>
<td>2.33E+0</td>
</tr>
<tr>
<td>4</td>
<td>turnover</td>
<td>5.14E+0</td>
<td>1.15E+0</td>
</tr>
<tr>
<td>5</td>
<td>dispossessed</td>
<td>1.75E+0</td>
<td>7.29E+0</td>
</tr>
<tr>
<td>6</td>
<td>clearance</td>
<td>1.47E+0</td>
<td>1.82E+0</td>
</tr>
<tr>
<td>7</td>
<td>save</td>
<td>4.44E+0</td>
<td>4.09E+0</td>
</tr>
<tr>
<td>8</td>
<td>tackle won</td>
<td>5.74E+0</td>
<td>1.17E+0</td>
</tr>
<tr>
<td>9</td>
<td>interception</td>
<td>2.47E+0</td>
<td>2.03E+0</td>
</tr>
<tr>
<td>10</td>
<td>shots total</td>
<td>4.42E+0</td>
<td>3.12E+0</td>
</tr>
<tr>
<td>11</td>
<td>shot adjusted distance</td>
<td>7.98E+0</td>
<td>1.84E+0</td>
</tr>
<tr>
<td>12</td>
<td>shot on target rate</td>
<td>4.71E+0</td>
<td>6.98E+0</td>
</tr>
<tr>
<td>13</td>
<td>penalty taken</td>
<td>2.03E+0</td>
<td>2.80E+0</td>
</tr>
<tr>
<td>14</td>
<td>pass long</td>
<td>3.57E+0</td>
<td>1.45E+0</td>
</tr>
<tr>
<td>15</td>
<td>pass long accurate rate</td>
<td>3.45E+0</td>
<td>3.57E+0</td>
</tr>
<tr>
<td>16</td>
<td>pass short</td>
<td>1.11E+0</td>
<td>3.73E+0</td>
</tr>
</tbody>
</table>
In table 3, some of the coefficients $\zeta_j$ (more specifically, penalty taken, goal own and red card) are abnormally large due to their small sample. What's more, coefficients of two stage regression $\zeta_j$ are mostly larger than coefficients $\eta_j$ 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.
4. Some property and hypothesis

4.1 Correlation 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 boxscore 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 coefficients among these 4 series of data.

<table>
<thead>
<tr>
<th>Correlation Coefficient</th>
<th>Rating</th>
<th>Boxscore Rating</th>
<th>Overall Ability</th>
<th>Market Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>1.00</td>
<td>0.944</td>
<td>0.518</td>
<td>0.489</td>
</tr>
<tr>
<td>Boxscore Rating</td>
<td>0.944</td>
<td>1.000</td>
<td>0.474</td>
<td>0.462</td>
</tr>
<tr>
<td>Overall Ability</td>
<td>0.518</td>
<td>0.474</td>
<td>1.000</td>
<td>0.573</td>
</tr>
<tr>
<td>Market Value</td>
<td>0.489</td>
<td>0.462</td>
<td>0.573</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Those correlation coefficients 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 coefficient $\alpha_2$ from age to value.

$$\hat{\text{value}} = \alpha_0 + \alpha_1 \cdot \text{Overallability} + \alpha_2 \cdot \text{age}(6)$$

$$\hat{\alpha}_2 = \frac{\hat{\alpha}_2}{\sqrt{\text{SE}(\hat{\alpha}_2)}}$$
\[ T = \sqrt{\frac{c}{\hat{\sigma}}} = \sqrt{\frac{S}{e}} \]  
\[ \sqrt{\frac{c}{\hat{\sigma}}} \sim t(n - 3) \]  
\[ 2 \]  
\[ c_{22} \]  
\[ n \]  
\[ 3 \]  
where \[ c_{22} \] is the (3,3) element of matrix \((X^T X)^{-1}\)  
\[ X \] is the matrix of independent variable \([1, \text{overallability}, \text{age}]\)  
\[ \text{Se} = \sum_{i=1}^{n} (\text{value}_i - \text{val'ue}_i)^2 \]  
Because of the large sample, we can approximately regard the t-distribution as gauss distribution with mean 0 and standard deviation 1.  

<table>
<thead>
<tr>
<th></th>
<th>constant</th>
<th>overallability</th>
<th>age</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha)</td>
<td>1.28E+08</td>
<td>2.64E+06</td>
<td>-2.35E+06</td>
<td>0.563</td>
</tr>
<tr>
<td>T-statistic</td>
<td></td>
<td></td>
<td>-40.33</td>
<td></td>
</tr>
</tbody>
</table>
The result shows that a player’s market value decrease about 2 million pounds when he get one year older while 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 negative correlation between players’ market value and age.

3.3 limitation on valuating goalkeeper

In some aspect, the rating and box score statistics exist significant limitation 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. 

\[
Z^{\text{rating}}_{\text{gk}} = \frac{\text{rating}_{\text{gk}} - \text{rating}_{\text{ngk}}}{\text{Sn}_{\text{gk}} + \text{Sn}_{\text{ngk}}} = 4.798 
\]

\[
Z^{\text{BR}}_{\text{gk}} = \frac{\text{BR}_{\text{gk}} - \text{BR}_{\text{ngk}}}{\sqrt{\text{Sn}_{\text{gk}}}} = -7.086
\]

---

Published by Francis Academic Press, UK

-100-
From the Z-statistic we have more than 99.99% confidence (p<0.0001) to believe that goalkeepers’ rating and Boxscore Rating is less than non-goalkeepers’. For Boxscore Rating the difference is larger. As we all know, goalkeepers always activate in penalty area and seldom have box score statistics. This means valuating players via box score statistics is not fair for goalkeepers.

References


**appendix: code**

```matlab
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;
ajR2 = 1 - (1 -R2)*(4718 -1)/(4718 -27 -1);
ma=[];
for i = 1:27
   ma = [ma,max(boxs(:,i))];
end
mam = repmat(ma,4718,1);
norboxs = boxs./mam;
E = eye(27);
betak =[];
for k = 0:0.01:10
   j = (norboxs'*norboxs + k*E)^(-1)*(norboxs')*rating;
betak = [betak,j];
end
for i = 1:26
   plot(0:0.01:10,betak(i,:))
hold on
end
saveas(1,'01.jpg')
hold off
thr = zeros(1,27);
for k = 1:27
   ...
end
```

Published by Francis Academic Press, UK

-102-
34 for j = 1:1001
35 if abs(betak(k, j)) < 0.05
36 thr(k) = 1;
37 end
38 end
39 end
40 for i = 1:26
41 if thr(i) == 1
42 plot(0:0.01:10, betak(i,:))
43 hold on
44 end
45 end
46 saveas(1, '02.jpg')
47 hold off
48
49 ajboxes = boxes;
50 ajboxes(:,:,4,5,11,14,16,17,20,21,24) = 0;
51 gamma = regress(rating, ajboxes);
52 nBR = ajboxes * gamma;
53 nSSR = sum((nBR - avgr).^2);
54 SST = sum((rating - avgr).^2);
55 nR2 = nSSR/SST;
56 najR2 = 1 - (1-nR2)*(4718 -1)/(4718 -27 -1);
57
58 abi = xlsread('WhoScoredFinal7.xlsx','AR3:CA4720');
59 abi = abi/100;
60 reguboxs = log(norboxs*(exp(1)-1)+1);
61 delta = [];
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 ajRm = [ajRm, 1 - (1 - SSR/SST)*(4718 -1)/(4718 -27 -1)];
74 74 end
75 75
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);
mamm(: ,27) = [];
preboxsnew = ((exp(abinew * delta )-1)/(exp(1)-1)).* mamm;
preboxsnew = [preboxsnew,ones (3896,1)];
zeta = regress ( value,preboxsnew);
eta = regress ( value,boxsnew);
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;
ajR21 = 1 - (1 -R21)*(3896 -1)/(3896 -27 -1);
ajR22 = 1 - (1 -R22)*(3896 -1)/(3896 -27 -1);
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];
XX = (X'*X)^( -1);
c22 = XX(3,3);
alpha = regress ( value,X);
a2 = alpha (3);
prevalue = X * alpha ;
Se = sum(( prevalue - avgv).^2);
R20 = Se/SST;