Assignment #2: Regression to the Mean

Introduction

A player's overall percentage of shots on target is a measure of his efficiency in front of goal. When given an opportunity to shoot, does the player hit the target, forcing the goalkeeper to act and keeping the attacking opportunity alive, or does the player miss the goal completely, wasting the opportunity? Even if the ball does not go into the goal, a shot on target forces the goalkeeper into a save, where the ball could be parried, keeping the attack alive.

To better analyze shooting efficiency, this report looks to analyze player's shooting ability in soccer by using regression to the mean to estimate the true talent level of Percent Shots on Target(SoT%) for players in the United Soccer League Championship(USLC), the second tier of American professional soccer.

To make this analysis, I used the American Soccer Analysis' xGoals Table, which contained the minutes, shots, Goals, Assists, Shots on Target, Key Passes, and other relevant personal soccer metrics for each USLC player during the 2022 season. For my purposes, I only looked into outfield players who had recorded at least one shot. Using shots and shots on target, I found each players SoT%.

Regression to the Mean

As defenders and attackers will naturally have different shooting abilities, I stratified field players by two position groups: players who mostly defend(CBs, WBs, DMs), and players who mostly attack(AMs, CMs, Ws, STs), running regression to the mean on each group. In finding the true talent level, I use the following formula for regression to the mean:

$$True\,Talent = \frac{(n/\sigma^2)\bar{y} + (1/\sigma_0^2)\mu_0}{(n/\sigma^2) + 1/\sigma_0^2} \tag{1}$$

For each player, I can directly estimate $n, \sigma^2, \bar{y}$, and μ_0 from the given data, but I must use a different method to estimate σ_0^2.

Estimating Population Variance

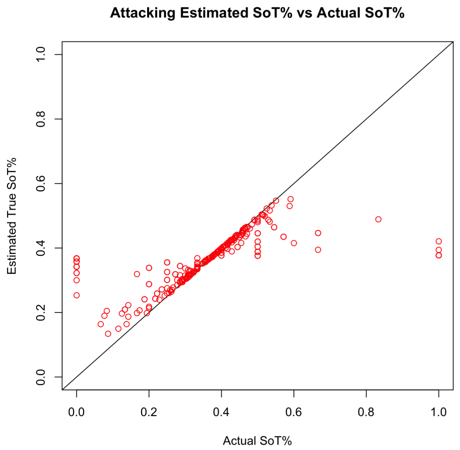
To estimate σ_0^2, I use the following formula:

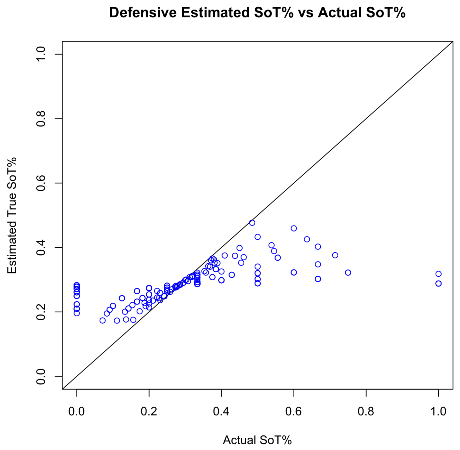
$$(\sigma_0^2)^{(t+1)} = \frac{\sum_{j=1}^{p} \hat{\sigma}_j^2/((\hat{\sigma}_0^2)^{(t)} + \hat{\sigma}^2/n_j)^2}{\sum_{j=1}^{p} 1/((\hat{\sigma}_0^2)^{(t)} + \hat{\sigma}^2/n_j)^2} \tag{2}$$

Where I initiate $\hat{\sigma}_0^2$ to be 0, repeating the formula until $\hat{\sigma}_0^2$ converges to, in my case, **0.0044** for defenders and **0.0030** for attackers. This variance is relatively small, meaning that both attackers and defenders in the USLC have a similar true talent of shots on goal percentage relative to their groups.

Analysis

I then plotted the estimated true and measured SoT% with the line $y = x$. Here, I can find which players have the highest true talent SoT%, with Luis Solignac, having an estimated true SoT% of **0.6083** as the highest attacker, and Dayonn Harris, with an estimated true SoT% of **0.6021**.





Unsurprisingly, the mean true talent for defenders(**0.2917**) was somewhat lower than the mean true talent for attackers(**0.3691**), as attacking players specialize more into shooting and scoring goals than defensive players.

For both groups, the data mostly stays around $y = x$, meaning that for most players their measured SoT% is very close to their estimated true talent. However, especially for defenders, this relationship sometimes deviates heavily from $y = x$. This discrepancy is likely from the small number of shots defensive players take, so there is more noise in their measured SoT%.

Limitations

However, one should note that SoT% is far from an all-encompassing metric surrounding a player's shooting efficiency, as the metric does not account for the difficulty of each shot. For example, a very difficult shot from 25 yards out is completely different than a simple tap-in from 3 yards out, but both are weighed the same for SoT%. Additionally, midfielders and defenders could inherently have a lower SoT% because they are more likely to shoot more difficult shots from farther out.

One way I could account for shot difficulty is to incorporate Expected Goals(xG) by finding a player's average $G - xG$ per 90 minutes, where G is 1 if the goal is scored and 0 otherwise. However, the loss of a binary result would require game-by-game data to find σ^2, which would be difficult to find for a league as small as the USLC.

Conclusion

To better gauge attacking efficiency in front of goal, this report looks to estimate the true Percent Shots on Target in the USLC for the 2022 season using regression to the mean. Through my estimation of true talent, I found that there was little spread in the true talent of SoT%, although attackers had a higher SoT% than defenders.