Problem Set 2

Donovan Doyle

2/16/2019

Question 1

It is helpful to see both because in hockey you have to play both offense and defense, as only 5 skaters are on the ice at one time. It would be impossible to evaluate a player's true impact on the ice if we only looked at his offensive production, because he could be terrible at defense. Net shots taken is a more robust statistic, as it compares the number of opportunities created while a player is on the ice versus the number of opportunities he allows. This is different from baseball in that a hitter can only affect the game with offense in one given sequence, and a fielder can only affect the game with defense in one given sequence, whereas in hockey possession is much more fluid, and any offensive move made will in turn affect your defense.

Question 2

We care about relative shot percentage versus others on your team because teams will vary in their raw number of shots, so shot percentage is a more true indicator of your output compared to your shot number. For example, if you took the 5th most shots on your team but were on the team that took the most shots in the league, people may think you handle a big workload on offense when compared to the leading shooter of the team that takes the least shots. It's similar to a fixed effect because the error would be constant across individuals if we looked at only shot numbers, while looking at relative shot percentage versus others on your team incorporates that error into the coefficient.

Question 3

The correlation between PDO and lagged_PDO is only 0.14, which is very weak, but doesn't necessarily mean they're not serially correlated. I also ran a regression that graphs lagged_PDO onto PDO, controlling for: games, age, salary, goals_sixty, assists_sixty, goals, assists, toi, corsi, cfrel_percent, all other lags included, and shots. Lagged_PDO was significant at the 99% level, meaning we are confident they are serially correlated. In the regression, it says that with a PDO of 103.38, the following year we'd expect a PDO of roughly 100.5, a slight decrease. This corresponds with the 59th percentile of the PDO data, a significant decrease from the 95th percentile the year before.

Question 4

The 95th percentile of goals is 21. If a player scored 21 goals one year, we would expect him to score around 15 the next year, a drop to the 80th percentile. The correlation between goals and lagged goals is 0.447. The 95th percentile of cfrel_percent is 6.26. If a player had a cfrel_percent of 6.26, we would expect him to have a cfrel_percent of around 3 the next year, which corresponds with a drop to the 78th percentile, roughly. The correlation between cfrel_percent and lagged_cfrel_percent is 0.63. The most replicable from year to year is goals, as this has the lowest drop in percentile. This would make sense, as goals doesn't control for volume, and a player who has a good year the year before will likely be more aggressive and take more shots the next year, in turn scoring a lot again, even if he's not that effective of a player in other metrics. This would likely be due to pressure from fans on the coach to set the player up for shots, as they see him as a high-goal scorer and more deserving of the puck.

Question 5

I first ran a regression with all lagged variables, and found that this could explain roughly 23.5% of the variation in goals, not a great predictor, but it is also statistically significant. The other possible specification would be to include age, salary and team as controls. This is much more robust, so it would be my preferred specification. Another possible prediction method would be The player with the most predicted goals would be Steven Stamkos in 2008.

Question 6

Undervalued players would be those with a low PDO, as this seems to be where regression to the mean is most present, given the drop from the player in the 95th percentile. Other undervalued players would be those with a high cfrel_percent but a low amount of goals, as they are actually extremely efficient players and their presence on the ice is correlated with opportunities for the team to score/limiting the other teams' opportunities, they just aren't the ones who score the goals themselves.

Part B

Question 1

This isn't a good metric, because, hypothetically, if in the long-run a players round average is 70, and he's shot really well the first 3 rounds with an average of 66, odds are he'll revert back to his typical mean and shoot over 70 in the last round. This unfairly punishes players for experiencing luck in the first 3 rounds that they won't necessarily experience in the fourth.

Question 2

The only way to truly test a player's skill would be to give them all the exact same group, same broadcast coverage, same weather, same standing in the tournament, same familiarity with the course, and the same pretty much anything else that could impact a player's performance outside of skill, then see how they shoot for multiple rounds. This is pretty obviously impossible/not feasible, as players likely never even end up in the same situation twice themselves.

Question 3

a) We can compare their play in the random groups to their play in the sorted groups, because if they're in a top sorted group, they're going to get more broadcast time. This is likely the best way for us to quantify the significance of pressure, given the variables we have. *b)* Having Tiger Woods in your group means you're going to have much more pressure on you, since he is the poster boy of golf and will always have the highest broadcast numbers. When casual fans watch golf, they generally always know how Tiger is doing, and if you're in Tiger's group, they're also watching you. *c)* Other players with predictive effects like Tiger might be Phil Mickelson or Rory McIlroy (these are the only two golfers who come to mind to me, and I don't follow golf at all, so that's likely a decent heuristic for who is most popular). We should include fixed effects for tournaments, because some tournaments are much more heavily broadcasted than others, so that must be factored into our analysis. It seems to be random as to who performs better when in Mickelson's group, so I don't think pressure in the first 2 rounds is that significant. Pressure likely comes more from being placed in a high-performing group for the last 2 rounds.

Question 4

Performance doesn't seem to change by player, but we include so many different specifications that our error is large. I think this likely would extrapolate to other sports that also include many specifications such as football, where there are so many different possible situations we couldn't isolate skill. Baseball, on the other hand, has few possible situations and we can actually isolate variable with a significant enough n to find meaningful results.