Sports Economics

Problem Set 2

**Due 11:59pm Monday, Feb 13**

This problem set will use data from the National Hockey League (NHL) to investigate the replicability and predictability of individual performances in hockey, with possible Moneyball-esque predictions.

Many of these questions will again be open-ended. It is more important to consider the strengths and weaknesses of different approaches we have discussed in the class.

Download nhlps2.csv from the course website. **The data contain statistics for all forwards (“goal scorers”) for any season in which they played at least 500 minutes from the 2007-2008 season until the 2015-2016 season. Note that this will be an unbalanced panel: not all players will show up in the data in every year.**

No background knowledge of hockey is necessary. Three points of interest for this assignment:

1) Players take turns on the ice, in the game, with 5 players on at a time. Unlike baseball where most players play the whole game. Hockey is tiring as players skate as fast as they can so players usually go on for shifts of about 30-40 seconds of game time. Players usually play shifts with the same teammates “lines”, throughout the game/season. Most teams rotate 3-4 lines.

2) There is no automatic switch from offense/defense as in baseball (After 3 outs in a half inning) or even like football where the other team gets the ball if you score. Thus, a strong team can play a large part of the game on offense, in the other team’s third of the ice (close to the goal they are attacking). They should also be expected to get more shot attempts as one can only shoot after getting the puck, “ball”, into the other team’s end. And even after they score, they can get the puck right back following a faceoff.

3) Goalies are very good. Equipment and technique have evolved such that on average less than 10% of shot attempts will go in. Thus, even if one player creates a lot of shots for their team, the team might not score so much. A lot of goals are the results of deflections/interference, that is perhaps out of the shooter’s/goalie’s control. In basketball, for comparison, around 50% of shots go in.

The variables of interest are:

Games = games played

Salary = salary in millions

Goals= goals scored

Assists= goals for which you provided the last or penultimate pass

Shots=number of shots you attempted

Goals\_sixty, Assists\_sixty= Adjusting your statistics to numbers per 60 minute game (length of

regular season game)[[1]](#footnote-1)

TOI = total time on ice = total time played

CorsiFor = the number of shots your team took when you were on the ice

CorsiAgainst = the number of shots the other team took when you were on the ice

Corsi = the net difference in shots taken by your team and the other team (higher is better.

Negative means the other team got more shots than your team when you were on the ice)

CFRel\_Percent = how your net corsi difference compares to your team’s overall corsi difference,

depending on your ice time and opponents. Higher is better.

PDO = think of this as a measure of your shooting percentage (goals/shots) relative to the rest of

the league where the mean is 100

Lags for some of these variables = their values in the previous period

In Part 1, we look at player performance:

Exercises:

1) Explore the data. Over this time period, in an average hockey game, each team scored roughly

2.75 goals on 30 shot attempts. Given this, why might we want to know about total shots by each team while you were on the ice versus just goals and assists? Compare hockey to basketball, where the average score is, say, 100-98 with each team scoring ~40 baskets per game on 90 shots? Recall the discussion of “expected home runs” from class?

2) What do you think the variable Corsi is trying to measure? Why might it be useful compared to just goals scored at both the team and the player level?

3) Why would we possibly care about your relative corsi percentage versus others on your team versus just your relative shot number compared to the league? How is this similar to the idea of a team fixed effect?

4) In basketball, shooting percentage is thought to be very replicable over time (we will revisit this

later in the course). For example, Stephen Curry seems to always have a 3-point shooting

percentage among the top 10 players in the NBA.[[2]](#footnote-2) Is it the case that hockey relative shooting

percentage (PDO) is highly replicable, i.e. that it is serially correlated? Pooling observations over

the full sample period, the 95th percentile in PDO was 103.38. Assuming someone had a PDO at

this level in year t, what would you forecast their PDO to be in year t+1 using just their previous year’s

performance? What percentile, over the full sample period, does this correspond to?

5) Repeat the forecasting exercise in question (3), this time using goals, instead of PDO, from one

year to predict goals in the next year. Repeat again with Relative Corsi. Which statistic is most

replicable from year to year? Does this match with your expectations?

6) Try to predict goals in season t, using only information from season t-1, employing at least two

different specifications. What is your preferred specification? For which player-season did you forecast the most goals for each specification?

Part 2- The Pythagorean Theorem for Hockey

In standings\_2018\_2019.csv, we have the standings for each hockey team from the 2018 first half of the season, 2018 full season, and 2019 full season.

GP-game played

W-wins

L-losses

OL-overtime losses (teams get a point for these)

PTS-points

Pts%- % of total points

GF-goals for (your goals)

GA-goals against (opponents’ goals)

“\*” denotes a team that made the playoffs

You should create a variable that is points per game because the number of games in the first half of the season was not identical for all teams.

7) a) Estimate Pythagorean wins using 2 different coefficients over the 2019 season. Which one fits the actual wins over the 2019 season best?

b) Now apply these two coefficients to the first half of the 2018 season. Which one minimizes the Root Mean Squared Error of predicted points per game in the second half. Is Pythagorean wins an improvement over points per game in the first half?

Bonus Question: Look up the rules on total points for games that go into overtime, in particular about shootouts. Why do you think this might make hockey a less than ideal venue for Pythagorean wins? In particular do you think relatively more weight should be given to offense or defense?

Part 3-Putting it together

8) Given what you found in part 1 and what we discussed in the Moneyball readings, what kinds of players do you think might have been undervalued in the hockey market over this period? Create a back of the envelope estimate of how much value the most underpaid player could add to a team that scores exactly as many goals as its opponents using your coefficients from 7. (There is no perfect answer, creativity is good. And hopefully you can have some fun with it. If you assume that people get paid based on their output in the current season or past season that is fine if the coding is too much.)

1. We do not want to punish players who did not get to play much, but think about how this might reward them. [↑](#footnote-ref-1)
2. If you are interested in looking more at 3-point skill at this point, see for example: https://thepowerrank.com/2020/07/28/predictability-vs-skill-in-sports-analytics-3-point-shooting/

   “compare this random assumption with actual players like Steph Curry. Steph has made 43.2% of his 3 point shots over the past six seasons. If each of Steph’s shots had a 35.6% chance to go in at random, Steph’s shooting percentage would be 9.6 standard deviations from the NBA average (based on 3,681 attempts). This is extremely unlikely. This confirms what we all know: Steph Curry is a great shooter, probably the best to have ever played the game. In contrast, Russell Westbrook has made 30.4% of his 2,150 attempts from 3. His percentage is five standard deviations below the NBA average.” [↑](#footnote-ref-2)