Hockey metrics

Michael Lopez, Skidmore College

Overview

In this lab, we'll gain experience implementing traditional and more novel approaches to measuring NHL player performance. We'll also some familiar approaches for looking at the repeatability of a statistic. First, our preamble to get the data.

```
library(RCurl)
library(mosaic)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/NHL.csv")
nhl.data <- read.csv(text = url)
dim(nhl.data)
summary(nhl.data)
head(nhl.data)
```

The data set `nhl.data` contains player-season level information for several seasons of the NHL. There are roughly 7900 player seasons, and we have 18 variables for each player within each season. Credit to the website and hockey blog war-on-ice for the data. It's an awesome site for grabbing data and learning statistics.

Some variables basic - including `Position`, `Team`, `Season`, and `Games` (number of games played). Let's start with those.

1. Use the `summary()` output to identify which seasons that we are drawing from. Additionally, how do you think `Salary` is coded?

Other variables will take some care to define. Let's go through a few other columns in our data set.

First, all productivity measures use 5 v. 5 play only. This is for a couple of reasons - power play or penalty kill performance adds noise to player performance, and a players' skill in such situations is largely dependent on his coach putting him on the ice.

Among the columns, `Goals_Sixty` and `Assists_Sixty` are rate statistics, calculated using (number of Goals/minutes on ice)*60.

Next, you'll see some terms that we referred to in class on Tuesday. These include `PDO`, `CF_Percent`, `Corsi`, `CFRel_Percent`, `CorsiFor`, and `CorsiAgainst`.

2. Look at the first line of our data set. Confirm that `Corsi` is simply a function of `CorsiFor` (total shot attempts for the player's team) minus `CorsiAgainst`.

Data cleaning

One thing you may have noticed is that the first player in the data set, Spencer Abbott, only played 5.46 minutes for Toronto during the 2013-2014 season. It doesn't make much sense to include Abbott, as it is difficult to learn properties of different metrics with such noisy data.

In addition to a sample size restriction, we also drop rows with missing observations (*note*: we can do this because we still maintain more almost all of our player-seasons). Finally, we look only at forwards, as it takes a bit more care to handle metrics for defenders. That explains the following filtering of defensemen.

```
nhl.data <- filter(nhl.data, TOI > 500, Position!="D", Position!="DL", Position!="DR")
nrow(nhl.data)
nhl.data <- na.omit(nhl.data)
nrow(nhl.data)
```

We are left with 2801 player-seasons, in which each player (using forwards only) accumulated at least 500 minutes of 5v5 play.

Corsi and shot statistics

Let's look at some of our metrics together to see how they relate to one another.

```
library(corrplot)
nhl.data <- mutate(nhl.data, Shots_Sixty = Shots/TOI*60)
cor.matrix <- cor(select(nhl.data,
   Goals_Sixty, Assists_Sixty, Shots_Sixty, CF_Percent, CFRel_Percent, PDO, Shots))
corrplot(cor.matrix, method = "number")
```

3. Which variables are most closely tied to `CFRel_Percent`? Recall - this is the fraction of shots above or below 0%, comparing a player's team and his opponent, relative to when that player is off the ice. Why do you think this number is more strongly linked to a player's shots than a player's goal rate?

4. Note the correlations with a players PDO. Why do you think goal and assist rate are linked to PDO, but not shot-based metrics such as `CF_Percent` and `CF_RelPercent`?

5. Make a scatter plot of `CF_RelPercent` as a function of `CF_Percent`, and describe why this link follows your expectation.

One thing that you may have observed in your last plot is that there were some players whose relative performance (to their team) fell above or below the line of best fit that we could have drawn in. Let's find out who those players are.

```
p<- ggplot(nhl.data, aes(CF_Percent, CFRel_Percent, label = Name))
p + geom_text() + scale_y_continuous("Relative Percentage of total shots taken") +
  scale_x_continuous("Percentage of shots taken")+
  theme_bw()
```

Note that you see some players' names twice - remember, we are using seasonal data, and so there are several players that may appear multiple times. One of the names that appears in the bottom right of the graph is Dallas Drake.

```
filter(nhl.data, Name =="Dallas.Drake")
```

Can you find Drake's point on the graph?

6. In the 2007-2008 season, Drake's team, the Detroit Red Wings, recorded 2.2% more shots than their opponents when Drake was on the ice. However, Drake's relative production (`CFRel_Percent`) was negative. Look up the 2007-2008 Detroit Red Wings on the internet, and conjecture as to why this was the case.

Repeatability (or lack thereof)

There are lots of things we may be interested in, and one interesting thing would be to look at which of these metrics are repeatable.

Some code to start:

```r
nhl.data1 <- nhl.data %>%
  arrange(Name, Season) %>%
  group_by(Name) %>%
  mutate(f.Goals = lead(Goals),
         f.Assists = lead(Goals),
         f.PDO = lead(PDO),
         f.CFRel_Percent = lead(CFRel_Percent),
         f.CF_Percent = lead(CF_Percent),
         Sh_Percent = Goals/Shots,
         f.Sh_Percent = lead(Sh_Percent),
         f.PDO = lead(PDO))
head(nhl.data1)
```

The data set `nhl.data1` contains the same information before, as well as the players' production in the season following (abbreviated with an `f.`)

Two variables that our readings told us were *not* repeatable were a players' shooting percentage (`Sh_Percent`) and his PDO (`PDO`).

Let's take a look.

```r
xyplot(f.Sh_Percent ~ Sh_Percent, data = nhl.data1)
cor(f.Sh_Percent ~ Sh_Percent, data = nhl.data1, use="pairwise.complete.obs")^2
xyplot(f.PDO ~ PDO, data = nhl.data1)
cor(f.PDO ~ PDO, data = nhl.data1, use="pairwise.complete.obs")^2
```

Only about 4% of the variability in a player's shooting percentage in year $t+1$ can be explained by his shooting percentage in year t. Similarly, there is barely any link between a players PDO from year t to $t+1$. This matches what we found in our readings.

7. Note that earlier in the lab, we did some filtering to make sure we only looked at offensive players (forwards). If we had included defensemen, why do you think the year to year correlation found above for `Sh_Percent` would be substantially higher than 4%?

Ultimately, we want to understand which hockey metrics are most reliable for predicting future success. Goals in hockey are hard to come by - there are only about 5.5 in a game - and so perhaps goal scoring alone is not enough to predict future goal scoring.

Let's look at the year to year correlations among several of our variables.

```r
cor.matrix <- cor(select(ungroup(nhl.data1),
    Goals, Assists, Sh_Percent, CF_Percent, PDO, CFRel_Percent,
    f.Goals, f.Assists, f.Sh_Percent, f.CF_Percent, f.PDO, f.CFRel_Percent),
    use="pairwise.complete.obs")
corrplot(cor.matrix[7:12, 1:6], method = "number")
```

In the above correlation matrix, note that code is organized so that instead of displaying a 12 x 12 matrix, we see a 6 x 6 one. This should make it easier for viewing the link between numbers in year t (columns) and year $t+1$ (rows).

8. Which of the 6 metrics most strongly correlates with itself in the following season? This shouldn't surprise you given our readings.
9. Is there a strong link between `f.Sh_Percent` or `f.PDO` and any of the players' stats in the previous year?
10. Interestingly, `Goals` in year *t* appear to have the same relationship strength to both `f.Goals` and `f.Assists`. Why do you think that is?

Finally, in the first row of the matrix, note that future goals are positively linked to several player metrics from the season prior, including goals, assists, `CF_Percent`, and `CFRel_Percent`. This comes up in Brian's paper, where he finds that the combination of current goals and possession statistics are better predictors of future goals that current goals alone.

Open-ended.

With a partner, work on the following question. Note that there's no exact correct answer, but we'll review our work at the end of class.

Using a multiple linear regression model, identify the link between `ln.Salary` (response variable, taken on the log scale and in millions of dollars) and one-ice metrics such as Goals, Assists, Shot Percentage, Corsi Percentage, PDO, Corsi Relative Percentage, and time on ice. Which are more relevant to a player's pay in a given season?

```r
nhl.data <- mutate(nhl.data, ln.Salary = log(Salary))
```

Recall that how to fit a multiple linear regression model (as well as how to check assumptions) was covered in Lab 3