Hockey metrics: expected goals

Michael Lopez, Skidmore College

Overview

In this lab, we'll return to the NHL play by play data set, containing shots from the 20172018 and 20182019 seasons.

```
library(RCurl); library(tidyverse)
gURL <- "https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/pbp_data_hockey.rds"
pbp_data <- readRDS(gzcon(url(gURL)))
names(pbp_data)

pbp_data <- pbp_data %>%
  mutate(coords_x_adj = ifelse(event_team == home_team,
                               -1*abs(coords_x), abs(coords_x)),
         coords_y_adj = ifelse(event_team == home_team & coords_x < 0,
                               coords_y, -1*coords_y),
         coords_y_adj = ifelse(event_team == away_team & coords_x > 0,
                               coords_y, -1*coords_y))
```

Mapping shots

Let's take a sample game, between Detroit and Edmonton in the 20182019 season. In this game, we can use the `count` command to identify the frequency with which each team shot the puck. Turns out, Edmonton had 20 more shots than Detroit.

```
sample_game <- pbp_data %>% filter(game_id == 2018020197)
sample_game %>% count(event_team)
```

Returning to the idea of mapping

```
ggplot(sample_game, aes(x = coords_x_adj, y = coords_y_adj,
                        colour = event_team, pch = event_type)) +
  geom_point()
```

One approach for improving the map above is to allow for different shapes or colorsbased on the type of shot (`event_type`)

```
ggplot(sample_game, aes(x = coords_x_adj, y = coords_y_adj,
                        colour = event_team)) +
  geom_point()

ggplot(sample_game, aes(x = coords_x_adj, y = coords_y_adj,
                        colour = event_type)) +
  geom_point()
```

1. What do the location maps above provide that the general shot map does not? Be specific .

Logistic regression modeling

We can try a few logistic regression models

```
library(broom)
pbp_data <- pbp_data %>%
```

```
  mutate(is_home = event_team == home_team)

fit_1 <- glm(event_type == "GOAL" ~ event_distance +
               event_angle + event_detail ,
             family = "binomial", data = pbp_data)
tidy(fit_1)

fit_2 <- glm(event_type == "GOAL" ~ is_home, data = pbp_data,
          family = "binomial")
tidy(fit_2)

fit_3 <- glm(event_type == "GOAL" ~ event_distance +
               event_angle + event_detail + is_home, data = pbp_data,
          family = "binomial")
tidy(fit_3)
```

2. What is the difference between `fit_1` and `fit_2`
3. Interpret the coefficient on `is_home` in `fit_2`. What does this entail about shooting the puck at home?
4. Interpret the coefficient on `is_home` in `fit_3`. What does this entail about shooting the puck at home?
5. Provide one possible explanation for your findings above.

Hosmer Lemeshow

We'll use the Hosmer Lemeshow test – review your notes from last class – as one approach to assess the fits above. Specifically, we create the goal probability `shot_prob_fit_3` as the predicted goal likelihood.

```
pbp_data$shot_prob_fit_3 <- predict(fit_3, pbp_data, type = "response")

tab_check <- pbp_data %>%
  filter(!is.na(shot_prob_fit_3)) %>%
  mutate(shot_prob_cat = cut(shot_prob_fit_3, 10)) %>%
  group_by(shot_prob_cat) %>%
  summarise(ave_exp_goals = sum(shot_prob),
            ave_act_goals = sum(event_type == "GOAL"),
            n_shots = n())

tab_check <- tab_check %>%
  mutate(diff_sq = (ave_exp_goals - ave_act_goals)^2/
           ((ave_exp_goals)*(1-ave_exp_goals/n_shots)))

tab_check
```

6. Descrive the distribution between observed and expected goals in the ten bins above. Where does the model tend to fit well? Where are there more goals than expected? Where are there less goals than expected?

Hosmer Lemeshow

```
hm_test <- tab_check %>%
  summarise(test_stat = sum(diff_sq))
hm_test

1-pchisq(hm_test$test_stat, df = 8, lower.tail = TRUE)
```

7. State the null and alternative hypotheses for the HL test. Would we reject or fail to reject the null hypothesis? What does this say about `fit_3`?