Lecture 10: Statistics in soccer

Skidmore College

Goals

- Multiple logistic regression
- Score effects
- Expected goals
- Advanced shot mapping

Set-up:

NHL shot data

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

```
##  [1] "season"         "game_id"        "game_date"      "session"
##  [5] "event_index"    "game_period"    "game_seconds"   "event_type"
##  [9] "home_team"      "away_team"      "home_skaters"   "away_skaters"
## [13] "home_score"     "away_score"     "event_detail"   "event_team"
## [17] "event_player_1" "event_player_2" "coords_x"       "coords_y"
## [21] "home_goalie"    "away_goalie"    "event_circle"   "event_distance"
## [25] "event_angle"    "shot_prob"
```

Adjusted coordinates

```
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))
```

Player metrics

```
season_2018 <- pbp_data %>%
  filter(season == 20172018) %>%
  group_by(event_player_1, season) %>%
  summarise(n_goals_18 = sum(event_type == "GOAL"),
            n_xGs_18 = sum(shot_prob),
            n_shots_18 = n()) %>%
  filter(n_shots_18 >= 100) %>%
  select(-season)

season_2019 <- pbp_data %>%
  filter(season == 20182019) %>%
  group_by(event_player_1, season) %>%
  summarise(n_goals_19 = sum(event_type == "GOAL"),
            n_xGs_19 = sum(shot_prob),
            n_shots_19 = n()) %>%
  filter(n_shots_19 >= 100) %>%
  select(-season)
```

Player metrics

```
season_combine <- season_2018 %>% inner_join(season_2019)
head(season_combine)
```

```
## # A tibble: 6 x 7
## # Groups:   event_player_1 [6]
##   event_player_1 n_goals_18 n_xGs_18 n_shots_18 n_goals_19 n_xGs_19
##   <chr>               <int>    <dbl>      <int>      <int>    <dbl>
## 1 AARON.EKBLAD           16    12.3         283         13     9.94
## 2 ADAM.HENRIQUE          24    23.2         212         18    17.2
## 3 ADAM.LARSSON            4     4.01        130          3     3.87
## 4 ADAM.PELECH             3     4.03        150          5     4.58
## 5 ADRIAN.KEMPE           16    11.8         161         12    11.3
## 6 ALEC.MARTINEZ           9     4.96        152          4     4.48
## # ... with 1 more variable: n_shots_19 <int>
```

```
library(corrplot)
```

Player metrics

```
cor_players <- cor(season_combine[,2:7])
corrplot(cor_players, method = "number")
```

Player metrics, conclusions

Score effects

```
pbp_data <- pbp_data %>%
  mutate(score_diff = ifelse(event_team == home_team,
                             home_score - away_score,
                             away_score - home_score),
         score_diff_cat = case_when(score_diff <= -1 ~ "Down",
                                    score_diff == 0 ~ "Tied",
                                    score_diff >= 1 ~ "Up"),
         is_goal = event_type == "GOAL")

pbp_data %>%
  group_by(score_diff_cat) %>%
  summarise(ave_goal = mean(is_goal),
            ave_distance = mean(event_distance, na.rm = TRUE),
            ave_Xg = mean(shot_prob))
```

```
## # A tibble: 3 x 4
##   score_diff_cat ave_goal ave_distance ave_Xg
##   <chr>             <dbl>        <dbl>  <dbl>
## 1 Down             0.0611         36.4 0.0615
## 2 Tied             0.0639         36.2 0.0623
## 3 Up               0.0791         37.0 0.0752
```

Score effects, conclusions

Soccer data

```
library(RCurl)
library(tidyverse)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/St
wwc_shot <- read.csv(text = url)
names(wwc_shot)
```

```
##  [1] "period"                "minute"                "second"
##  [4] "possession"            "duration"              "possessio
##  [7] "play_pattern.id"       "play_pattern.name"     "player.id
## [10] "player.name"           "position.name"         "shot.stat
## [13] "shot.first_time"       "shot.technique.name"   "shot.outc
## [16] "shot.type.name"        "shot.body_part.name"   "match_id"
## [19] "location.x"            "location.y"            "location.
## [22] "location.y.GK"         "player.name.GK"        "DistToGoa
## [25] "DistToKeeper"          "AngleToGoal"           "aveveloci
## [28] "distance.ToD1"         "DefendersBehindBall"   "TimeInPos
```

Soccer data

```
wwc_shot_summary <- wwc_shot %>%
  group_by(match_id, possession_team.name) %>%
  summarise(n_goals = sum(shot.outcome.name == "Goal"))
wwc_shot_summary %>%
  head()
```

```
## # A tibble: 6 x 3
## # Groups:   match_id [3]
##   match_id possession_team.name   n_goals
##      <int> <fct>                    <int>
## 1    22921 France Women's               4
## 2    22921 Korea Republic Women's       0
## 3    22924 Nigeria Women's              0
## 4    22924 Norway Women's               2
## 5    22926 China PR Women's             0
## 6    22926 Germany Women's              1
```

Expected goals/link to future

Summarize: *Expected goals 2.0* (link)

Summarize: *Best predictor of future performance is expected goals* (link)

Expected goals, repeatability of finishing skill

Summarize: *Repeatability of finishing skill* (link)

Randomness and expected goals

Summarize: *12 shots good, 2 shots better* (link)

Expected goals and addition

Summarize: *Expected goals don't add* (link)