HW 7 solutions

Stats and sports class

Fall 2019

Part II: Implementation

We can access recent shot data here:

```
library(RCurl); library(tidyverse)
```

```
## Loading required package: bitops
```

```
## -- Attaching packages ---------------------------------------------------------- tidyverse 1.2.1 -
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------- tidyverse_conflicts() -
## x tidyr::complete() masks RCurl::complete()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
```

```
gitURL<- "https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/pbp_data_hockey.rds"
pbp_data <- readRDS(gzcon(url(gitURL)))
```

Question 2

Run the model below

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

```
## # A tibble: 9 x 5
##   term                    estimate std.error statistic   p.value
##   <chr>                      <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)              -1.16    0.0329      -35.4  6.14e-275
## 2 event_distance           -0.0416  0.000646    -64.3  0.
## 3 event_angle              -0.0144  0.000446    -32.3  4.38e-229
## 4 event_detailDeflected    -0.154   0.0556       -2.77 5.68e-  3
## 5 event_detailSlap          0.345   0.0423        8.15 3.49e- 16
## 6 event_detailSnap          0.397   0.0370       10.7  7.93e- 27
## 7 event_detailTip-In       -0.175   0.0400       -4.38 1.21e-  5
## 8 event_detailWrap-around  -0.447   0.101        -4.41 1.03e-  5
## 9 event_detailWrist         0.234   0.0312        7.51 5.86e- 14
```

Interpret the coefficient on `event_detailWrist`

Solution: The odds of a successful shot are `exp(0.234) = 1.26` times higher on wrist shots, relative to backhand shots, using a model accounting for distance and angle. Note – the reference group is backhand shots and should be mentioned

Question 4

For `game_id == 2017020324`, identify each participating team's goals and expected goals. Did the outcome of this game match the relative shot inputs?

Solutions

```
pbp_data %>% filter(game_id == 2017020324) %>%
  group_by(event_team) %>%
  summarise(ave_xg = sum(shot_prob),
            act_goal = sum(event_type == "GOAL"))
```

```
## # A tibble: 2 x 3
##   event_team ave_xg act_goal
##   <chr>       <dbl>    <int>
## 1 PIT          4.48        2
## 2 VAN          3.78        5
```

Pittsburgh outscored Vancouver on expected goals 4.48 to 3.78, but lost 5 to 2

Bonus

Find the one game across the last two seasons where the different between the observed goal differential was as different from the expected goal differential

```
pbp_data %>%
  group_by(game_id) %>%
  summarise(xg_home = sum(shot_prob*(event_team == home_team)),
            xg_vis = sum(shot_prob*(event_team == away_team)),
            home_goal = sum((event_type == "GOAL")*(event_team == home_team)),
            vis_goal = sum((event_type == "GOAL")*(event_team == away_team))) %>%
  mutate(goal_diff = home_goal - vis_goal,
         xg_diff = xg_home - xg_vis,
         abs_diff = abs(goal_diff - xg_diff)) %>%
  arrange(-abs_diff) %>%
  head(5)
```

```
## # A tibble: 5 x 8
##   game_id    xg_home xg_vis home_goal vis_goal goal_diff xg_diff abs_diff
##   <chr>        <dbl>  <dbl>     <int>    <int>     <int>   <dbl>    <dbl>
## 1 2018020141    3.38   2.75         1        9        -8  0.626      8.63
## 2 2018020974    3.12   2.51         1        8        -7  0.604      7.60
## 3 2018021046    3.22   3.23         1        8        -7 -0.0194     6.98
## 4 2017020275    3.64   4.55         8        2         6 -0.909      6.91
## 5 2017020010    4.01   1.89        10        1         9  2.12       6.88
```

Solutions: game_id == 2018020141 was expected to be a 3.38 to 2.75 win for the home team, but was actually a 9 to 1 win for the away team. That has the largest difference in this data set. There are *lots* of ways to code this – any is sufficient