Exam 2

Stats and sports class

Fall 2019

Preliminary notes for doing exams

1. All files should be knit and compiled using R Markdown. Knit early and often! I do not recommend waiting until the end of the HW to knit.
2. All questions should be answered completely, and, wherever applicable, code should be included.
3. You may not work with anyone else or seek help beyond the use of your notes, HW, labs
4. Copying and pasting of code is a violation of the Skidmore honor code
5. 11 AM is the deadline (Saturday)

Part I (20 total pts)

Current Brewers analyst Dan Turkenkpf wrote an article titled https://www.beyondtheboxscore.com/2008/4/5/389840/framing-the-debate, in which he outlines the idea of pitch framing in baseball.

1. Identify three of the best catchers at framing according to the article. In the first table, what do each columns refer to?
2. In the comment section, Dan identifies a correlation coefficient of 0.51 with respect to pitch framing (year over year). Given the size of the differences between pitchers, as well as this correlation coefficient, do you think pitch framing in 2008 was undervalued or overvalued by Major League Baseball officials?

Current Eagles analyst Namita Nandakumar proposed using a tool called survival analysis to assess the time that it takes NHL prospects to reach the NHL. See her slides at https://hockeygraphsdotcom.files.wordpress.com/2017/10/namita.pdf

3. Describe what's going on in the graph "Draft prospect survival curves, time until nth career NHL game". Specifically, identify:

- Why is the red line below the purple line?
- Roughly what fraction of players that were just drafted play the first game they possibly can?

4. Namita writes that North American players have approximately a 35 percent decrease in the hazard ratio. What does this mean as far as drafting players goes? Are teams generally overvaluing or undervaluing North American players, relative to European ones?

Part 2 (40)

The next part of the test will use our hockey shot data set.

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

See our hockey unit for a description of each variable

1. Identify the shooter who took the highest number of shots.

2. Identify the goalie who faced the highest number of shots.
3. Describe differences in the likelihood of a goal based on shot type, `event_detail`.
4. A coach is interested in using the first 500 games of the season, and to use shooting percentages to predict shooting percentages over the remainder of the season (as well as the next season). The code below splits games at `game_id == 2017020500`, which is the 500th game of the season in 2017-2018.

```
first_shots <- filter(nhl_shots, game_id <= 2017020500)
current_shots <- first_shots %>%
  group_by(event_player_1) %>%
  summarise(n_shots_past = n(),
            shot_p_past = mean(event_type == "GOAL")) %>%
  filter(n_shots_past >= 150)

future_shots <- nhl_shots %>%
  filter(game_id > 2017020500, event_player_1 %in% current_shots$event_player_1)%>%
  group_by(event_player_1) %>%
  summarise(n_shots_future = n(),
            shot_p_future = mean(event_type == "GOAL"))

nhl_players <- current_shots %>% inner_join(future_shots)
```

Provide the coach with the following:

i)

Two sets of estimates of the goal percentages for the remainder of the two seasons, the James-Stein estimate and the maximum likelihood estimate (MLE)

ii)

A comparison of the accuracy of each of your two sets of estimates. Use `shot_p_future` as the known truth regarding how well a player shoots.

iii)

The relative amount of shrinkage towards the overall league average that a shooter can expect after roughly 175 shots.

(Bonus, 5 points)

Visualize the James-Stein estimator with respect to past performance and eventual career performance for these players.

Part 3 (30 points)

```
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/sb_shot_data.csv")
wwc_shot <- read.csv(text = url)
names(wwc_shot)
```

1. Find the woman player for England (`possession_team.name == "England Women's"`) who had the best shooting performance of the tournament. That is, given each England shooters number of expected goals, which player overperformed the most?
2. Make a shot map for all shots from England's Ellen White (`player.name`), using a different symbol for whether or not each shot resulted in a goal.
3. A coach wants an analyst to measure the likelihood of a goal.

```
library(splines)
wwc_shot <- wwc_shot %>%
  mutate(is_goal = shot.outcome.name == "Goal")

fit1 <- glm(is_goal ~ avevelocity + minute, data = wwc_shot, family = "binomial")
fit2 <- glm(is_goal ~ ns(avevelocity, 5) + minute, data = wwc_shot, family = "binomial")
fit3 <- glm(is_goal ~ avevelocity + ns(minute, 5), data = wwc_shot, family = "binomial")
fit4 <- glm(is_goal ~ ns(avevelocity, 5) + ns(minute, 5), data = wwc_shot, family = "binomial")
```

Pick which of the models above makes the most sense to use to share with the coach.

4. What is the association between average velocity `avevelocity` and the likelihood of a goal? Several possible tools are usable here, including a look at the models above.
5. Use the Hosmer-Lemeshow test to assess `fit2` above. Is there any evidence of a lack of fit?

Part 4 (10 points)

Use the home ice Bradley Terry model to answer the following questions

```
library(broom)
library(BradleyTerry2)
head(icehockey)
dim(icehockey)
homeBT <- BTm(result,
              data.frame(team = visitor, home.ice = 0),
              data.frame(team = opponent, home.ice = home.ice),
              ~ team + home.ice,
              id = "team", data = icehockey)

tidy(homeBT)
tidy(homeBT) %>% tail()
head(BTabilities(homeBT), 10)
```

1. Assuming team strength is held constant, what are the increased odds that the home team wins?
2. Estimate the probability that Alabama Huntsville beating Air Force in

- a game with no home-ice advantage
- a game with Alabama Huntsville having home advantage
- a game with Air Force having home advantage