Exam 2

Stats and sports class

Fall 2020

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, and class recordings.
4. Copying and pasting of code is a violation of the Skidmore honor code
5. Submit (virtually) your exam by 5:00 PM EST on Friday, Nov 20th

Part I (15 total pts)

Read Derrick Yam's article at StatsBomb: https://statsbomb.com/2019/02/attacking-contributions-markov-models-for-football/

Then, answer the following questions.

1. What does *transient states* mean? And how does Derrick calculate 84 of them in the article?
2. Derrick's chosen Markov model is *memoryless*: What does that mean, and do you think that assumption is justified in this example?
3. Describe how Derrick uses the Markov model to calculate a players contribution.

Part II (20 points)

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

```r
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. The team estimate for Yale is 0.519. Interpret this number.

3. 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

Part III (35)

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. Look only at shooters with at least 100 shots. Which player has the highest average shot probability (shot_prob) within this group?

2. Describe differences in the likelihood of a goal based on shot type, event_detail.

3. 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, and calculates the James-Stein estimator for these players.

```
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)


p_bar <- mean(nhl_players$shot_p_past)
p_bar
p_hat <- nhl_players$shot_p_past
p_hat


N <- nhl_players$n_shots_past
N
sigma_sq <- sd(p_hat)^2 ##Rough approximation
sigma_sq
```

```
c <- (N/0.25)/(N/0.25 + 1/sigma_sq)
c
```

```
nhl_players$Shp_MLE <- nhl_players$shot_p_past
nhl_players$Shp_JS <- p_bar + c*(p_hat - p_bar)
head(nhl_players)
```

```
nhl_players %>%
  ungroup() %>%
  mutate(abs_error_mle = abs(Shp_MLE - shot_p_future),
         abs_error_js = abs(Shp_JS - shot_p_future)) %>%
  summarise(mae_mle = mean(abs_error_mle),
            mae_js = mean(abs_error_js))
```

Provide the coach with the following:

i) An overall estimate for the grand-mean of all players average shooting percentages

ii) For a given shooting percentage, the rough fraction that one should shrink back towards the overall estimate in (i)

iii) For a given player, provide an example of the above. That is, show how his shooting percent is pulled towards the league average

iv) For that same player – using his `shot_p_future` – which was a more accurate estimate? The James-Stein estimate or the Maximum Likelihood Estimate?

(Bonus, 5 points)

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

Part IV (30 points)

```
wwc_shot <- read_csv("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/sb_shot_da
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. Pick which of the models above makes the most sense to use to share with the coach. Justify using two reasons.

```
wwc_shot <- wwc_shot %>%
  mutate(is_goal = shot.outcome.name == "Goal",
         ave_velocity_sq = avevelocity^2,
         minute_sq = minute^2)
```

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

4. What is the association between game minute and the likelihood of a goal? Several possible tools are usable here, including a look at the models above. Your answer should include at least one graph, as well as examples cited from the statistical models. Respond in paragraph form, as if you were summarizing your results in 4 to 5 sentences to a coach.