Exam 2

Michael Lopez, Skidmore College

Note: Please submit the exam using RMarkdown. **Five** points of your exam will be awarded based on formatting and presentation. Take time to ensure that your answers are shown in R output and succinctly described.

Part 1 (16 total pts)

Beginning this postseason, the National Basketball Association is going to track *hustle stats*, as explained here. Read the article, and answer the following questions.

Question 1 (8 pts)

The article indicates that *trained personnel are going to count the hustle stats.* With reference to an example from our unit on referee bias, provide a 2-3 sentence explanation as to how the league's tracking of hustle statistics could go awry.

Question 2 (8 pts)

Which *hustle stats* are the league going to record? Of these, identify which metric you think is (i) most repeatable (ii) most important to team success (iii) most a function of individual talent, and not team strategy. Justify your answers in one sentence each. Note that you can use the same metric more than once.

Part 2 (24 total pts)

We are going to use the NBA's shot-level data to look at the **two-point shots**. Here's the data you'll need to start. The variable `dist.cat` splits two-point shots into four categories: 0 to 3 feet, 4 to 6 feet,

```r
library(RCurl)
library(mosaic)
url <- getURL("https://raw.githubusercontent.com/JunWorks/NBAstat/master/shot.csv")
nba.shot <- read.csv(text = url)
nba.shot <- nba.shot %>%
  na.omit() %>%
  filter(SHOT_DIST < 22 & PTS_TYPE==2)
nrow(nba.shot)
nba.two <- nba.shot %>%
  mutate(dist.cat = cut(SHOT_DIST, breaks = c(-100, 3, 6, 12, 100),
                        labels = c("D1", "D2", "D3", "D4")),
         late.clock = SHOT_CLOCK < 5)
```

```r
fit1 <- glm(FGM ~ dist.cat + SHOT_CLOCK, data = nba.two)
fit2 <- glm(FGM ~ dist.cat + late.clock, data = nba.two)
```

Question 1 (5 pts)

Interpret the coefficient for `dist.catD2` in `fit1`, using the odds ratio scale.

Question 2 (6 pts)

Compare the coefficients for `SHOT_CLOCK` and `late.clock` in `fit1` and `fit2`, respectively. What is each one suggesting about the chances of a two-point shot going in as a function of the shot clock?

Question 3 (6 pts)

For measuring the link between shot clock and success (given distance), would you prefer `fit1` or `fit2`? If you don't like either `fit1` or `fit2`, suggest an alternative model specification. *Note that you should not fit any additional models or provide any code here.*

Question 4 (6 pts)

Using `fit1`, estimate the expected point total for a 1 foot shot taken with 10 seconds left on the shot clock.

Part 3 (56 total pts)

The final part of the test will use `shots`, a newer hockey shot-level data set that is posted to the class site.

```r
library(mosaic)
library(RCurl)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/shots.csv")
nhl.shots <- read.csv(text = url)
nhl.shots <- nhl.shots %>%
  mutate(distance.sq = distance*distance, goal = (etype =="GOAL"))
```

Most variables are self-explanatory: note that `type` is the shot type and `etype` is the event outcome (shot or goal). The distance of the shot is measured in feet; we add a square term for distance, as well as an indicator variable, `goal`, which is a 1 if the shot goes in and a 0 otherwise. The variables `ev.player.1`, `away.G`, and `home.G` are numerical identifiers of the shooter and the two goalies on the ice during the time of the shot: the variable `Goalie` is the goalie facing the shot.

Question 1 (5 pts)

Identify the shooter who took the highest number of shots and the goalie who faced the highest number of shots.

Question 2 (5 pts)

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

Question 3

A coach is interested in using the first 200 games of the season to project performance of a goalie over the remainder of both that season and the following season. Here's code to get a subset of goalies that the coach wants to lookat.

```
first.shots <- filter(nhl.shots, seas == 20142015, gcode <= 22000)
first.players <- first.shots %>%
  group_by(Goalie) %>%
  summarise(n.shots = length(goal), n.saves = n.shots-sum(goal), save.p = n.saves/n.shots) %>%
  filter(n.shots > 1400, n.shots < 1800)
```

Consult the following code as guides, per HW 8 and Lab 9.

For the variance of the binomial distribution:

```
sigma.sq <- p.bar*(1-p.bar)/1600 ##Rough approximation
sigma.sq
```

To get career percentages:

```
all.players <- nhl.shots %>%
  group_by(Goalie) %>%
  filter(Goalie %in% first.players$Goalie) %>%
  summarise(n.shots.all = length(goal), n.saves.all = n.shots.all-sum(goal), save.p.all = n.saves.all/n
```

You can link the `all.players` and `first.players` data sets by using `inner_join()`: See Lab 9.

Provide the coach with the following:

3a (8 pts)

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

3b (8 pts)

A comparison of the accuracy of each of your two sets of estimates.

3c (6 pts)

The relative amount of shrinkage towards the overall league average that a goalie can expect after roughly 1600 shots

3d (8 pts)

A summary of your findings, in 2-3 non-technical sentences.

4 (5 pts)

Earlier in class we separated shots by position (forwards, defense) and we also advocated a similar strategy in baseball. Justify the decision to not attempt such a strategy in this analysis.

5 (6 pts)

Using concepts from our hockey readings as a guide, comment on save percentage in hockey with respect to both its repeatability and its importance to team success.

6 (5 pts)

Given your results above, do you think there would be more or less reversion to the league average when looking at the three point success rate of NBA shooters with a similar number of attempts (~ 1600)? Explain your answer.

(Bonus, 5 points)

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