Exam 1

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. Turn in a hard copy of your exam (stapled) to my mailbox by 12 noon on Tuesday.

Part I (16 total pts)

Beginning a few years ago, the National Basketball Association began to track *hustle stats*, as explained here. Read the article, and answer the following questions.

Question 1 (4 pts)

Which *hustle stats* are the league going to record?

Question 2 (8 pts)

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

Question 3 (4 pts)

In our basketball readings section is an article on Goodhart's law: https://www.vice.com/en_us/article/jp7xb3/moreyball-goodharts-law-and-the-limits-of-analytics. What is Goodhart's law? And why might analysts want to think about Goodhart's Law as the NBA begins to track hustle stats?

Part II (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,

```
library(RCurl)
library(tidyverse)
url <- getURL("https://raw.githubusercontent.com/JunWorks/NBAstat/master/shot.csv")
nba_shot <- read.csv(text = url)
nba_two <- na.omit(nba_shot)%>%
  filter(SHOT_DIST <=21 & PTS_TYPE==2, SHOT_DIST >= 0)
nrow(nba_shot)
nba_two <- nba_two %>%
  mutate(dist_cat = cut(SHOT_DIST, breaks = c(-100, 3, 6, 12, 100),
                        labels = c("D1", "D2", "D3", "D4")),
         late_clock = SHOT_CLOCK < 5)
```

```
fit1 <- glm(FGM ~ dist_cat + SHOT_CLOCK, data = nba_two, family = "binomial")
fit2 <- glm(FGM ~ dist_cat + late_clock, data = nba_two, family = "binomial")
```

Question 1 (4 pts)

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

Question 2 (4 pts)

What does `fit1` suggest about the chances of a two-point shot going in as a function of the shot clock?

Question 3 (4 pts)

What does `fit2` suggest about the chances of a two-point shot going in as a function of the shot clock?

Question 4 (4 pts)

Both terms for the shot clock in `fit1` and `fit2` are significant. Provide one possible explanation for the discrepency you find above.

Question 5 (4 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 6 (4 pts)

Using `fit1`, estimate the expected point total for a 10 foot shot taken with 16 seconds left on the shot clock.

Part III (4 pts each, 44 total)

In the `Lahman` package, the `Fielding` data set contains information about how players performed defensively in each season.

We're particularly interested in the repeatability of fielding percentage, defined the number of total putouts (`PO`) and assists (`A`) divided by the number of opportunities a player had to field a ball (putouts, assists, and errors (`E`)). The following code identifies players with at least 100 attempts to field a ball between 1970 and 2000.

```
library(Lahman)
Fielding_1 <- Fielding %>%
  mutate(fielding_attempts = PO + A + E,
         fpct = (PO + A)/fielding_attempts) %>%
  filter(fielding_attempts >= 100, yearID >= 1970, yearID <= 2000)
```

Question 1

Make a histogram of fielding percentage, and comment on its center, shape, and spread.

Question 2

Compare the distributions of fielding percentage by each position (`POS`). What does this suggest about certain positions in baseball?

Question 3

Assess the repeatability of fielding percentage from one year to the next. That is, for each player, calculate their fielding percentage in the following season. Call each players' fielding percentage in the following season `fpct_next`.

Question 4

Same as in **Question 3**, except calculate the repeatability of fielding percentage within each position.

Question 5

Revisit our baseball readings and labs on repeatability. Where does fielding percentage rank, relative to batting and pitching metrics?

Question 6

Imagine fielding percentage had instead been nearly 100% repeatable – that is, each player's fielding percentage stayed consistent across his or her career. Why might a baseball expert not neccesarily conclude that the players with the best fielding percentages were the players who were best defensively?

Question 7

Make a spaghetti plot of each player's fielding percentage, and facet by position. Can you identify any conclusions related to your findings in **Question 5**?

Question 8

Roughly, what is the mean absolute error when using a players' fielding percentage in one year to predict his fielding percentage in the next year?

Question 9

Fit a linear regression of `fpct_next` as a function of `fpct` and position. What is your estimated model?

Question 10

Field the player-season with the lowest residual. What position did that player play?

Question 11

Were the residuals from your fit normally distributed?

Part IV (5 pts each, 15 total)

Question 1

A coach is faced with a fourth down conversion attempt, 75 yards from his own goal. He looks at the following table of expected point totals and their conditional probabilities under two strategies - the coach goes for it or the coach kicks a field goal. Which decision will maximize this teams' expected points?

Go for it	Field Goal	Points
0.60	0.00	7
0.20	0.80	3
0.10	0.05	-3
0.10	0.15	-7

Question 2

Explain which strategy the team's coach should take under the minimax criterion, and why.

Question 3

Go back to one of our readings - the sabermetric manifesto. What about the sport of football makes it more difficult to achieve some of the general principles that the author discusses? In that regard, why are field goal kickers among the easiest group to study?

Write the Skidmore Honor Code (by hand, 1 point)

While taking this examination, I have not witnessed any wrongdoing, nor have I personally violated any conditions of the Skidmore College Honor Code.

Bonus (5 pts)

Recall our kicking data in the NFL

```r
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/nfl_fg.csv")
nfl_kick <- read.csv(text = url)
nfl_kick <- nfl_kick %>%
  mutate(Distance_sq = Distance^2)
head(nfl_kick)
fit_1 <- glm(Success ~ Distance_sq + Distance + Grass + Year,
             data = nfl_kick, family = "binomial")
```

Estimate the odds of a kick going in if it's 5-yards further away using `fit_1`.