Exam 1

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, Oct 2nd

Part I: Data wrangling and exploratory analysis (35 pts)

We'll start by using the `Fielding` data in the `Lahman` package, which calculates the fielding metrics in each season for each player.

We're particularly interested in 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)
```

For each of the following questions, please use the `Fielding_1` data set.

Question 1

Identify the player/year with the lowest fielding percentage in any season in this time frame.

Question 2

Identify the outfielder (`POS == "OF"`) with the lowest fielding percentage in any season in this time frame.

Question 3

A coach wants to identify *perfect* fielders – that is, those whose `fpct` is 100 percent. What percent of players at each position register as having perfect fielding percentages?

Question 4

Make a chart of fielding percentage (y-axis) versus fielding attempts (x-axis), faceted by position. Describe the general trend of what happens as attempts goes up. What does this indicate about fielding percentage?

Question 5

Describe the center, shape, and spread of the double plays (`DP`) variable

Question 6

Use a visualization to compare the distribution of double plays (`DP`) turned by players at each position.

Question 7

For each player, calculate his average (average of each season) fielding percentage across this time frame. Which 5 players have the lowest average fielding percentage?

Part II (35 pts)

Recall our kicking data in the NFL. An analyst fits a pair of logistic regression models.

```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")

fit_2 <- glm(Success ~ Distance + Grass + Year,
             data = nfl_kick, family = "binomial")
```

Question 1

Interpret the coefficient on Distance in `fit_2`, on the log odds scale

Question 2

Interpret the coefficient on Grass in `fit_2`, on the odds scale

Question 3

Which model would you recommend? Use two justifications

Question 4

Using each of the models, estimate the probability of a successful field given the following conditions:

- 50 yards
- not on Grass
- Kicked in 2013

How important is the squared term in terms of changing the probability of this successful field goal?

Question 5

Using the distance and surface info above, estimate the likelihood of the same field goal being made in 2030, and comment on the appropriateness of this estimate.

Question 6 (open ended)

Using a combination a code and intuition, assess “field goal success” as a yearly measure of kicker aptitude. Consider the three ways we’ve discussed measuring a metric.

Part III (30 pts)

Return to the Lahman data.

```
Batting_1 <- Batting %>%
  filter(yearID >= 1970, AB >= 500) %>%
  mutate(K_rate = SO/(AB + BB),
         BB_rate = BB/(AB + BB),
         BA = H/AB,
         HR_rate = HR/(AB + BB),
         X1B = H - X2B - X3B - HR,
         TB = X1B + 2*X2B + 3*X3B + 4*HR,
         RC = (H + BB)*TB/(AB + BB))

Batting_1 <- Batting_1 %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(RC_next = lead(RC)) %>%
  filter(!is.na(RC_next)) %>%
  ungroup()
```

The following code creates categories for hitters based on the number of stolen bases they record in a season.

```
Batting_1 <- Batting_1 %>%
  mutate(SB_category = case_when(SB > 25 ~ "Fast",
                                 SB > 5 ~ "Moderate",
                                 SB <= 5 ~ "Slow"))
```

Question 1

Fit a regression model of runs created as a function of stolen base category, and interpret the coefficient for `Moderate` speed

Question 2

What is the estimated difference in runs created between runners in the `Moderate` and `Slow` categories?

Question 3

Fit a series of multiple regression models trying to estimate the link between runs created in the future and runs created in the season. Consider the inclusion of three additional variables: `HR_Rate`, `K_rate`, and `BB_rate`, in addition to `RC`. Pick the best possible model using the AIC criterion.

Question 4

Using your model in Question 3, estimate the residual for the very first row of the data set (`Batting_1`). Did this player create more or fewer runs than the model expected?

Question 5

Check your assumptions for fitting a multiple regression model (using your model in Q3)

Question 6

For your best model, estimate the mean absolute error (MAE) when applying your model to the `Batting_1` data set. Interpret this number.