Exam 1

Michael Lopez, Skidmore College

Note: Please submit the homework using RMarkdown.

Part 1

The first part of the test will use the Lahman package in R. Our interest lies in modeling the relationship between salary and player specific characteristics. To start, we need to join two data frames. We also only look at roughly the last 15 years of players (salary data has not always been available), and restrict our sample to those players with at least 500 at bats in a season. In this data set, `salary` is measured in dollars - we divide by a million to get `salary2`, which is a bit easier to interpret.

```r
library(Lahman)
library(mosaic)
Salaries <- Lahman::Salaries
data("Batting")

Batting.1 <- left_join(Batting, Salaries)
Batting.1 <- Batting.1 %>%
      filter(yearID >=2000, AB > 500) %>%
      mutate(X1B = H - X2B - X3B - HR,
      TB = X1B + 2*X2B + 3*X3B + 4*HR,
      RC1 = (H + BB)*TB/(AB + BB), salary2 = salary/10^6)
```

Question 1 (5 pts)

Use an appropriate visualization to identify and describe the link between runs created (`RC1`) and salary (`salary2`).

Question 2 (4 pts)

Identify the median salary of players (in millions of dollars) in each of the American and National Leagues.

Question 3 (6 pts)

Using runs created (`RC1`), singles, doubles, home runs, stolen bases, and salary, make a correlation matrix. Which offensive statistics appears most strongly linked to salary?

Question 4 (6 pts)

Estimate a multiple regression model of `salary2` as a function of singles, doubles, triples, home runs, and stolen bases. Running the R-code is not sufficient - state your estimated model.

Question 5 (4 pts)

Using the model in (Q4), interpret your coefficient for home runs.

Question 6 (5 pts)

Use the appropriate specification checks to determine if your model in (Q4) meets the assumptions for linear regression.

Question 7 (6 pts)

What fraction of variability in salary can be explained by the model in (Q4). Would you consider this a relatively large or small percentage? And what does this suggest about the link between offensive metrics and player evaluation (as judged by salary)?

Question 8 (5 pts)

The coefficient for stolen bases from your fit in (Q4) is not significant. Drop this term and refit your model. Using the AIC criteria, which model makes for a stronger fit - the one with or without stolen bases?

Question 9 (4 pts)

A general manager asks you to identify the player-seasons which, given a players singles, doubles, triples, and home runs, were most undervalued and most overvalued in terms of salary. What is the statistial term that this manager is looking for?

Question 10 (5 pts)

Using information from our readings to guide you, make an educated guess if the link between performance and salary among pitchers is stronger, weaker, or similar to the one that we just observed with hitters. Do not perform any analysis.

Part II

In this section, we'll explore logistic regression and outcomes in the NFL.

Recall: to load the data, run the following code.

```r
library(RCurl)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Kickers2.csv")
nfl.kick <- read.csv(text = url)
head(nfl.kick)
nfl.kick[is.na(nfl.kick$Temp),]$Temp <- 60
```

The data set loaded is slightly updated from our last kicker data - it also contains the game's temperature, stored as `Temp`. Given that there's some missing temperature data on games played indoors, a line of code arbitrarily sets those games at 60-degrees.

Question 1 (5 pts)

Identify if there's a link between field goal distance and the game's temperature using a scatter plot. Which of these two variables is the explanatory, and which is the response?

Question 2 (4 pts)

Describe the distribution of temperature among both successful and unsuccessful field goals using the following code. Are there any noticeable differences?

```
bwplot(Temp ~ as.factor(Success), data = nfl.kick)
```

Question 3 (4 pts)

A researcher fits a logistic regression model of `Success` as a function of `Distance` and `Temp` using the code below.

```
fit.LR <- glm(Success ~ Distance + Temp, data = nfl.kick)
msummary(fit.LR)
```

You notice that the coefficient for distance is nowhere near what we had obtained earlier in the semester during labs and class. Can you identify what the researcher did wrong?

Question 4 (4 pts)

The researcher identifies the problem, and refits the model. Does temperature appear to be linked with field goal success, given kick distance? Why or why not?

Question 5 (6 pts)

(a) Using your model in (4), estimate the odds of a successful field goal given a 50 degree increase in temperature.

(b) It appears temperature is a significant predictor of field goal success - is this information practically significant?

Question 6 (5 pts)

Using your model in (4) estimate the probability of a successful 40-yard field goal in 60-degree weather.

Question 7 (4 pts)

Your answer to question 6 is averaged over the last 10 years of kicking data. Explain why this percentage might underestimate the likelihood of a successful 40-yard field goal in 60-degree weather in next season's games.

Question 8 (6 pts)

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
0.58	0.05	7
0.14	0.80	3
0.10	0.05	-3
0.18	0.10	-7

Question 9 (6 pts)

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

Question 10 (6 pts)

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?

Bonus 1 (2 pts)

Did you attend the baseball & stats lecture by Dan Turkekopf?

Bonus 2 (3 pts)

One assumption that we've made in our models of kicker success is that the effect of distance on the log-odds of a successful attempt are constant across all distance. This suggests that moving from a 20-yard field goal to a 30-yard field goal has the same impact on the log-odds of success as going from 50-yards to 60-yards. Interpret the code below in reference to this possibility.

```r
nfl.kick$Distance.sq <- nfl.kick$Distance^2
fit.extra <- glm(Success ~ Distance+ Distance.sq + Grass + Year + Temp, data = nfl.kick, family = binom
msummary(fit.extra)
```