Exam 1 solutions

Stats and sports class

Fall 2019

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)

- how often defenders contest 2- and 3-point shots
- deflections by defensive players
- charges taken
- which players recover loose balls
- and so-called "screen assists"

Question 2 (8 pts)

Answers will vary.

Question 3 (4 pts)

Players may play to the metrics listed above if they know they are being tracked. If you try to take more charges, maybe you'll get fewer steals and more fouls (as one example of things that could go wrong)

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)
fit2 <- glm(FGM ~ dist_cat + late_clock, data = nba_two)
```

Question 1 (4 pts)

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

On average, shots in the second distance category are `exp(-0.079) = 0.924` times as likely to go in as shots in the first distance category, using a model with shot clock. This is equivalent to a 7.6 percent drop in odds

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?

Shots taken with more time on the shot clock are more likely to go in – so the less time, the less likely a shot is to go in.

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?

Only shots taken in the last 4 seconds of the shot clock are less likely to go in, and by a significant margin (the coefficient is -0.07)

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.

One model assumes a linear association between shot clock time and log odds of a shot – the other takes all shots in the early part of the shot clock (first 20 seconds) and treats them the same

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.*

The AIC is *much* lower on `fit1`, so I'd prefer that of these two models. It'd also be worth trying a quadratic term for shot clock time

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.

```
df_predict <- data.frame(SHOT_CLOCK = 16, dist_cat = "D3")
predict(fit1, df_predict, type = "response")

p <- exp(0.5687145 - 0.2203293 + 0.0051173*16)/(1 + exp(0.5687145 - 0.2203293 + 0.0051173*16))
p
```

We expect about a 43 percent chance of the shot going in if you code using predict, or about 60 percent if you code by hand. The initial model was off – either answer above counts

Part III (4 pts each, 44 total)

```
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.

```
ggplot(Fielding_1, aes(fpct)) +
  geom_histogram()

mean(Fielding_1$fpct)
```

Fielding percentage is centered around 98 percent, with a strong left skew. The lowest value is around 85 percent, while no fielding percentage is larger than 100.

Question 2

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

```
ggplot(Fielding_1, aes(x = POS, y = fpct)) +
  geom_boxplot()
```

Fielding percentage is lowest for third basemen and highest for first basemen and catchers. There are noticeable between-position differences. Most positions are skewed left. Shortstops also have relatively lower fielding percentages. The suggestion is that not all positions should be treated equally

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`.

```
Fielding_2 <- Fielding_1 %>%
  arrange(playerID, yearID) %>%
  group_by(playerID) %>%
  mutate(fpct_next = lead(fpct, 1)) %>%
  filter(!is.na(fpct_next))

Fielding_2 %>%
  ungroup() %>%
  summarise(cor_fpct = cor(fpct, fpct_next))
```

The year-to-year correlation is about 0.43.

Question 4

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

```
Fielding_2 %>%
  ungroup() %>%
  group_by(POS) %>%
  summarise(cor_fpct = cor(fpct, fpct_next))
```

The year-to-year correlation is highest among shortstops and outfielders. For third basemen, there basically is no correlation. Shortstops are at 0.26, OF at 0.026, C at 0.18, 3B at 0.02, 2B at 0.201, and 1b at 0.100

Question 5

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

Low. This is not a metric that is relatively repeatable

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?

You can only field a ball if you get to it. Players who are able to get to more balls and make plays on them are more valuable, even if they have a slightly lower fielding percentage because of it

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**?

```
ggplot(Fielding_1, aes(yearID, fpct, group = playerID)) +
  geom_point() +
  geom_line() +
  facet_wrap(~POS)
```

Hard to tell given the overlap, but there is lots of noise from year to year in the 3B plot (which matches the low correlation)

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?

```
Fielding_2 %>%
  ungroup() %>%
  mutate(ae = abs(fpct-fpct_next)) %>%
  summarise(mae = mean(ae))
```

The MAE is 0.0118.

Question 9

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

```
fit_1 <- lm(fpct_next ~ fpct + POS, data = Fielding_2)
summary(fit_1)
```

The estimated model is:

$fpct_next - hat = 0.79 + 0.20 * fpct - 0.008 * POS2b - 0.015 * POS3b + 0.002 * POSC - 0.002 * POSOF + 0.016 * POSP - 0.012 * POSSS$

Question 10

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

```
Fielding_2$pct_hat <- predict(fit_1, Fielding_2)
Fielding_2 %>%
  ungroup() %>%
  mutate(resid = fpct_next - pct_hat) %>%
  arrange(resid) %>%
  slice(1) %>%
  print.data.frame()
```

An outfielder – who actually had a fielding percentage of 86 percent – was predicted to hav ea fielding percentage of 98.1 percent. PlayerID: `willike02`

Question 11

Were the residuals from your fit normally distributed?

```
qqnorm(resid(fit_1))
```

There's some skewness to the QQ plot – the residuals were not normall 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

Go for it: *7*0.6+0.2*3-3*0.1*-70.1 = 3.8** **Field goal: 0.8*3* -*3*0.05-7*0.15 = 1.2**

Going for it is the better way of maximizing expected points

Question 2

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

Minimax: avoid worst case scenario. To avoid worst case scenario, the coach would want to go for it (lower probability of giving up -7)

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?

Field goal kickers are primarily the players in charge of their outcomes. Less reliance on teammates

Bonus (5 pts)

Recall our kicking data in the NFL

```
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")

predict_df <- data.frame(Distance = c(30, 25), Grass = TRUE, Year = 1980) %>%
  mutate(Distance_sq = Distance^2)
predict_df$phat <- predict(fit_1, predict_df, type = "response")
p1 <- predict_df$phat[1]
p2 <- predict_df$phat[2]

(p1/(1-p1))/(p2/(1-p2))


predict_df <- data.frame(Distance = c(50, 45), Grass = TRUE, Year = 1980) %>%
  mutate(Distance_sq = Distance^2)
predict_df$phat <- predict(fit_1, predict_df, type = "response")
p1 <- predict_df$phat[1]
p2 <- predict_df$phat[2]

(p1/(1-p1))/(p2/(1-p2))
```

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

Odds of a kick 5-yards further away is about 40% lower, when given a model with Grass and Year