Homework 5 solutions

MA 276, Skidmore College

Overview

In this assignment, we'll practice implementing logistic regression to estimate the probability of successful NBA shots. We'll also link to shot-level probabilities and expected points. Before we do anything, we have to load and clean the data, as in Lab 6.

```
library(RCurl)
library(mosaic)
url <- getURL("https://raw.githubusercontent.com/JunWorks/NBAstat/master/shot.csv")
nba.shot <- read.csv(text = url)
nba.shot <- na.omit(nba.shot)
nba.shot <- filter(nba.shot, PTS <4, SHOT_DIST>=22 |PTS_TYPE==2)
nrow(nba.shot)
```

```
## [1] 193716
```

Expected Points

All else being equal, what's the most efficient shot in the NBA?

In our lab, we characterized by points type using the following code:

```
tally(SHOT_RESULT ~ PTS_TYPE, data = nba.shot, format = "proportion")
```

```
##             PTS_TYPE
## SHOT_RESULT         2         3
##      made   0.4872612 0.3588893
##      missed 0.5127388 0.6411107
```

Of course, all two-point shots are not created equal. Using the cut command, we split two-pointers by distance into different groups, labeled `D1` to `D7`, in order from shortest to longest and grouped by shot type (2 or 3 points). The two data sets, `nba.two` and `nba.three` contain the two and three-pointers, respectively.

```
nba.two <- nba.shot %>%
  filter(PTS_TYPE == 2) %>%
  mutate(dist.cat = cut(SHOT_DIST, breaks = c(-100, 3, 6, 12, 100),
                        labels = c("D1", "D2", "D3", "D4")))

nba.three <- nba.shot %>%
  filter(PTS_TYPE == 3) %>%
  mutate(dist.cat = cut(SHOT_DIST, breaks = c(0, 23, 25, 100),
                        labels = c("D5", "D6", "D7")))


tally(SHOT_RESULT ~ dist.cat, data = nba.two, format = "proportion")
```

```
##            dist.cat
## SHOT_RESULT        D1        D2        D3        D4
##      made   0.6466731 0.5590257 0.4058810 0.4016135
##      missed 0.3533269 0.4409743 0.5941190 0.5983865
```

```
tally(SHOT_RESULT ~ dist.cat, data = nba.three, format = "proportion")
```

```
##            dist.cat
## SHOT_RESULT        D5        D6        D7
##      made   0.3911755 0.3651993 0.3274834
##      missed 0.6088245 0.6348007 0.6725166
```

Question 1

In order from best (highest expected points) to worst (lowest), order the categories D1 to D7.

```
tally(SHOT_RESULT ~ dist.cat, data = nba.two, format = "proportion")[1,]*2
```

```
##        D1        D2        D3        D4
## 1.2933461 1.1180514 0.8117621 0.8032271
```

```
tally(SHOT_RESULT ~ dist.cat, data = nba.three, format = "proportion")[1,]*3
```

```
##        D5        D6        D7
## 1.1735264 1.0955978 0.9824501
```

D1 is worth the most (short two's, 1.29 EP), followed by D5, D2, D6, D7, D3, D4 (long two's, 0.80 EP)

Question 2

Using code from our last lab, identify of expected points are higher on two or three point shots taken by Rajon Rondo.

```
tally(SHOT_RESULT ~ PTS_TYPE, data = filter(nba.shot, playerName=="Rajon Rondo"), format= "proportion")
```

```
##            PTS_TYPE
## SHOT_RESULT         2         3
##      made   0.4438202 0.3243243
##      missed 0.5561798 0.6756757
```

```
0.444*2
```

```
## [1] 0.888
```

```
0.324*3
```

```
## [1] 0.972
```

Rondo is slightly better from three (in terms of EP) than from two.

Question 3

Here's are two models of shot success (note that we re-bind all of the shots together).

```
nba.shot2 <- rbind(nba.two, nba.three)

fit.1 <- glm(SHOT_RESULT == "made" ~ SHOT_DIST + TOUCH_TIME +
             DRIBBLES + SHOT_CLOCK + CLOSE_DEF_DIST,
              data = nba.shot2, family = "binomial")

fit.2 <- glm(SHOT_RESULT == "made" ~ dist.cat + TOUCH_TIME +
             DRIBBLES + SHOT_CLOCK + CLOSE_DEF_DIST,
              data = nba.shot2, family = "binomial")

AIC(fit.1);AIC(fit.2)
```

```
## [1] 256475.4
```

```
## [1] 255898.9
```

Using the AIC criteria, which is the preferred fit of shot success? Is it close?

It's not close; the second model is much preferred (the second model categorizes distance)

Question 4

Using `fit.2`, estimate the increased odds of a made shot given a one-unit increase in closest defender distance. Then, estimate the increased odds of a made shot given a ten-unit increase in closest defender distance.

```
exp(0.0906)
```

```
## [1] 1.094831
```

```
exp(0.0906*10)
```

```
## [1] 2.474405
```

The odds of a successful shot go up about 9.4% for a one unit increase in defender distance, and about 147% for a 10 unit increase.

Question 5

Add game location (`LOCATION`) to `fit.2`. Does this improve the fit? Is the coefficient for this term statistically and/or practically significant? What does that suggest?

```
fit.3 <- glm(SHOT_RESULT == "made" ~ dist.cat + TOUCH_TIME +
             DRIBBLES + SHOT_CLOCK + CLOSE_DEF_DIST + LOCATION,
              data = nba.shot2, family = "binomial")
summary(fit.3)
```

```
##
## Call:
## glm(formula = SHOT_RESULT == "made" ~ dist.cat + TOUCH_TIME +
##     DRIBBLES + SHOT_CLOCK + CLOSE_DEF_DIST + LOCATION, family = "binomial",
##     data = nba.shot2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4754  -1.0261  -0.8628   1.1921   1.8961
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.2041825  0.0205654   9.928  < 2e-16 ***
## dist.catD2     -0.3207490  0.0168415 -19.045  < 2e-16 ***
## dist.catD3     -0.9154825  0.0177404 -51.605  < 2e-16 ***
## dist.catD4     -1.1318980  0.0171878 -65.855  < 2e-16 ***
## dist.catD5     -1.3866261  0.0273928 -50.620  < 2e-16 ***
## dist.catD6     -1.4632110  0.0202127 -72.391  < 2e-16 ***
## dist.catD7     -1.5580083  0.0234208 -66.523  < 2e-16 ***
## TOUCH_TIME     -0.0355983  0.0044736  -7.957 1.76e-15 ***
## DRIBBLES        0.0128513  0.0037917   3.389 0.000701 ***
## SHOT_CLOCK      0.0133471  0.0008658  15.415  < 2e-16 ***
## CLOSE_DEF_DIST  0.0906125  0.0022216  40.787  < 2e-16 ***
## LOCATIONH       0.0370002  0.0093952   3.938 8.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 266912  on 193715  degrees of freedom
## Residual deviance: 255861  on 193704  degrees of freedom
## AIC: 255885
##
## Number of Fisher Scoring iterations: 4
```

```
AIC(fit.3)
```

```
## [1] 255885.3
```

As judged by AIC, we have a stronger fit. This suggests court location may improve our model- perhaps players shoot better at home (or take easier shots).

Question 6

Does it make sense to add if the shooter's team was victorious (variable `W`) or margin of victory (`FINAL_MARGIN`) to the model? Why or why not? You do not need to run any code to answer this.

No: Both of these variables are calculated **after** the players have shot.

Question 7

Using Seth's article here and referencing the charts shown, explain Goodhart's law as it applies to statistics in the NBA.

Answers will vary: primarily, more short three point shots and fewer longer two-point shots indicates the sign of a good offensive team, and that forcefully taking more three point shots and removing the long two-point shots will not neccessarily make an offense better.