# Lab 10: Penalties in football

*Michael Lopez, Skidmore College*

## Overview

In this lab, we are going to work in pairs to answer interesting questions on penalties in the NFL. As you start, think back to our lessons on referee behavior. Where do you expect to find differences in NFL penalty rates?

First, the data:

```
library(mosaic)
library(RCurl)
x <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/penalties.csv")
nfl.data <- read.csv(text = x)
head(nfl.data)
```

```
##    gid    pid off def type qtr min sec ptso ptsd dwn ytg yfog zone  ptm
## 1 3990 652367 PIT  NE KOFF   1  15   0    0    0   0   0    0    0 <NA>
## 2 3990 652368 PIT  NE RUSH   1  15   0    0    0   1  10   20    1 <NA>
## 3 3990 652369 PIT  NE PASS   1  14  21    0    0   1  10   38    2 <NA>
## 4 3990 652370 PIT  NE RUSH   1  14   4    0    0   2   1   47    3 <NA>
## 5 3990 652371 PIT  NE PASS   1  13  26    0    0   1  10   51    3 <NA>
## 6 3990 652372 PIT  NE RUSH   1  12  42    0    0   1  10   65    4 <NA>
##   desc seas day   v  h temp
## 1 <NA> 2015 THU PIT NE   65
## 2 <NA> 2015 THU PIT NE   65
## 3 <NA> 2015 THU PIT NE   65
## 4 <NA> 2015 THU PIT NE   65
## 5 <NA> 2015 THU PIT NE   65
## 6 <NA> 2015 THU PIT NE   65
```

```
dim(nfl.data)
```

```
## [1] 44976    21
```

The data set `nfl.data` contains play-level data for the last full season of the NFL. It will take a moment to load - that's okay, it's roughly 44,000 rows.

Some variables have straight-forward interpretations, while others are clarified below.

| Variable | Description |
|---|---|
| gid | Game ID |
| pid | Play ID |
| off | offensive team |
| def | defensive team |
| type | play type |
| ptso | points for offense |
| ptsd | points for defense |
| dwn | 1-4 |

| Variable | Description |
|---|---|
| yfog | yards from own goal |
| zone | quintile of field (1-5) |
| ptm | team penalized |
| desc | penalty description |
| day | day of the week |
| v | visiting team |
| h | home team |

You may have noticed that the first six variables have some missing data. That's fine; on those plays, no penalties were called. We can get a sense of what rows with penalties look like by sampling other rows.

```
nfl.data[6:10,]
```

```
##     gid    pid off def type qtr min sec ptso ptsd dwn ytg yfog zone  ptm
## 6  3990 652372 PIT  NE RUSH   1  12  42    0    0   1  10   65    4 <NA>
## 7  3990 652373 PIT  NE PASS   1  12   5    0    0   1  10   76    4 <NA>
## 8  3990 652374 PIT  NE NOPL   1  11  20    0    0   2  18   68    4  PIT
## 9  3990 652375 PIT  NE RUSH   1  10  53    0    0   2  28   58    3 <NA>
## 10 3990 652376 PIT  NE PASS   1  10  28    0    0   3  22   64    4 <NA>
##                  desc seas day   v  h temp
## 6                <NA> 2015 THU PIT NE   65
## 7                <NA> 2015 THU PIT NE   65
## 8  Offensive Holding 2015 THU PIT NE   65
## 9                <NA> 2015 THU PIT NE   65
## 10               <NA> 2015 THU PIT NE   65
```

On the 8th play of the game against New England (`def`), Pittsburgh (`off` and `ptm`) was called for an offensive holding (`desc`). This play occured on 2nd down and 18, with Pittsburgh 68 yards from its own goal. Note that this play is officially recorded as `NO PLAY`, and not a run or a pass.

We can get a sense of what penalties were called by using `tally()`.

```
tally(~desc, data = nfl.data)
```

```
##
##                           12 On-field
##                                    67
##                            Chop Block
##                                    16
##                              Clipping
##                                     8
##                     Defensive Holding
##                                   315
##                     Defensive Offside
##                                   240
##           Defensive Pass Interference
##                                   270
##                        Delay of Game
##                                   167
##                      Disqualification
##                                     4
```

```
##                           Encroachment
##                                     45
##                              Face Mask
##                                    114
##                  Fair Catch Interference
##                                      7
##                             False Start
##                                    589
##                            Horse Collar
##                                     15
##                  Illegal Blindside Block
##                                      8
##           Illegal Block Above the Waist
##                                    154
##                          Illegal Contact
##                                     92
##                         Illegal Crackback
##                                      4
##                         Illegal Formation
##                                    103
##                       Illegal Forward Pass
##                                      9
##                            Illegal Motion
##                                     19
##                         Illegal Peelback
##                                      3
##                            Illegal Shift
##                                     48
##                       Illegal Substitution
##                                     12
##                         Illegal Touch Kick
##                                      2
##                         Illegal Touch Pass
##                                     11
##                       Illegal Use of Hands
##                                    168
##                 Ineligible Downfield Kick
##                                     11
##                 Ineligible Downfield Pass
##                                     33
##                     Intentional Grounding
##                                     35
## Interference with Opportunity to Catch
##                                      3
##                 Invalid Fair Catch Signal
##                                      2
##                     Kickoff Out of Bounds
##                                      1
##                                  Leaping
##                                      3
##                                Leverage
##                                      1
##                               Low Block
##                                      3
```

```
##                           Neutral Zone
##                                    160
##                       Offensive Holding
##                                    883
##                       Offensive Offside
##                                      4
##             Offensive Pass Interference
##                                    135
##                     Offside on Free Kick
##                                     17
##                            Personal Foul
##                                     12
##           Player Out of Bounds on Punt
##                                     15
##                      Roughing the Kicker
##                                      2
##                      Roughing the Passer
##                                    107
##                  Running Into the Kicker
##                                     19
##                                Taunting
##                                     23
##                                Tripping
##                                     12
##                    Unnecessary Roughness
##                                    259
##                  Unsportsmanlike Conduct
##                                     95
##                                   <NA>
##                                  40651
```

## Tasks for today

Using whatever penalty outcome(s) that you'd like, come up with both an interesting visualization (mosiac plot, histogram, scatter plot, bar plot, etc) and a logistic regression model that look at the likelihood of your outcome(s) as a function of game and/or play-specific characteristics. For example, you could look at likelihood of a penalty as it relates to the game's temperature, the minute of each play, the down and/or distance, if the guilty team was the home team, or the spot on the field where the play occurred.

Here's some code that can help. Once you've chosen your penalty outcome(s), you can create a new variable that reflects a `TRUE` if your specific penalty was called and a `FALSE` if that penalty was not called.

```
nfl.data1 <- nfl.data %>%
  mutate(DPI = (desc =="Defensive Pass Interference"),
         DPenalty = (desc =="Face Mask")|(desc == "Horse Collar"))
```

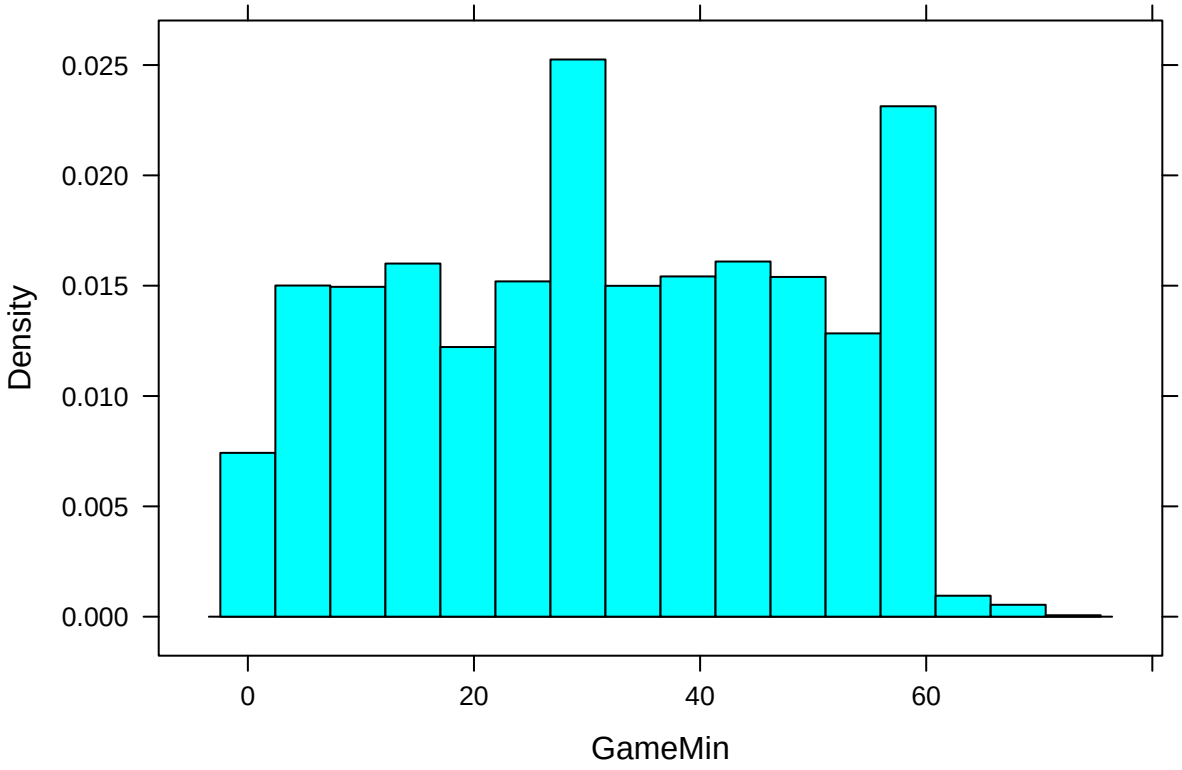
The first variable, `DPI`, is TRUE if there was a defensive pass interference. The second variable, `DPenalty` is a TRUE if there was either a face mask or a horse collar penalty.

You can check this by coding the following table, and ensuring that your outcomes line up with the penalty descriptions.

```
tally(desc ~ DPI, data = nfl.data1)
tally(desc ~ DPenalty, data = nfl.data1)
```

One other variable that takes a bit of care in coding is the game's minute (1-60+). Here's one attempt which roughly gets it right.

```
nfl.data2 <- mutate(nfl.data1, GameMin = 15*(qtr-1) + (15-min))
histogram(~GameMin, data = nfl.data2)
```




The first play of the game is always coded as Minute 0; nearly all games end at minute 60 (other than overtime ones).

Finally, when you are done, knit your file in RMarkdown and we'll share out class' work in the last fifteen minutes.

Good luck - and please see me for help coding!