Aggressiveness

Andrew Lim

2017-07-08

Initialization:

```
# Loading libraries:
suppressMessages(library(tidyverse))
```

```
# Reading data:
data <- suppressMessages(read_csv("../data/AllActionsperMatchUSAMajorLeagueSoccer2016.csv"))
match_data <- suppressMessages(read_csv("../data/TeamDataMLS2016.csv"))
```

```
## Warning: Duplicated column names deduplicated: 'goals' => 'goals_1' [49]
```

```
# Renaming variable:
match_data <- match_data %>%
  rename(match_result = Result)
```

```
# Joining the two data tables:
data <- data %>%
  left_join(match_data, by=c("Match" = "Matchid", `Team in possession` = "Team"))
```

Data wrangling

```
# Calculating number of aggressive turnovers per team per match:
aggressive_df <- data %>%
  mutate(next_team = lead(team)) %>%
  mutate(turn_over = team != next_team) %>%
  mutate(aggressive_turnover = LocX > 50 & turn_over == TRUE) %>%
  mutate(aggressive_turnover_by_team = aggressive_turnover == TRUE & `Team in possession` == next_team)
  select(`Team in possession`, next_team, aggressive_turnover, aggressive_turnover_by_team, everything()
  group_by(Match, `Team in possession`) %>%
  summarize(num_aggressive_turnovers = sum(aggressive_turnover),
            match_result = unique(match_result),
            HomeAway = unique(HomeAway),
            num_passes = unique(`#Passes`),
            num_goals = unique(goals)) %>%
  mutate(is_home = ifelse(HomeAway == "T", TRUE, FALSE)) %>%
  mutate(is_win = match_result == "W")
```

Creating models

Model showing teams play more aggressive at home

```
# Teams tend to play more aggressive when they are at home:
model <- lm(num_aggressive_turnovers ~ is_home, data=aggressive_df)
summary(model)
```

```
##
```

```
## Call:
## lm(formula = num_aggressive_turnovers ~ is_home, data = aggressive_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.147 -14.538  -1.538  13.462  76.462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  130.147      1.102 118.126  < 2e-16 ***
## is_homeTRUE   10.391      1.558   6.669 5.35e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.32 on 678 degrees of freedom
## Multiple R-squared:  0.06156,    Adjusted R-squared:  0.06018
## F-statistic: 44.48 on 1 and 678 DF,  p-value: 5.346e-11
```

The result above shows that teams, on average, have 10.391 more aggressive turn overs when they are playing at home. In other words, teams play more aggressive at home.

Model showing if it is a good idea to play aggressive or not

Now, we can create a model to see whether or not playing aggressive can lead to a higher chance of winning either HOME or AWAY.

```
# Creating logistic model:
model2 <- glm(is_win ~ num_aggressive_turnovers + is_home + is_home*num_aggressive_turnovers, data=aggr

# Displaying results:
summary(model2)
```

```
##
## Call:
## glm(formula = is_win ~ num_aggressive_turnovers + is_home + is_home *
##     num_aggressive_turnovers, data = aggressive_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7817  -0.2872  -0.1567   0.4177   0.9751
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.651766   0.152234   4.281 2.13e-05
## num_aggressive_turnovers            -0.003562   0.001156  -3.082  0.00214
## is_homeTRUE                          0.883470   0.225062   3.925 9.54e-05
## num_aggressive_turnovers:is_homeTRUE -0.003826   0.001643  -2.329  0.02016
##
## (Intercept)                          ***
## num_aggressive_turnovers             **
## is_homeTRUE                          ***
## num_aggressive_turnovers:is_homeTRUE *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 0.1887578)
##
##     Null deviance: 153.16  on 679  degrees of freedom
## Residual deviance: 127.60  on 676  degrees of freedom
## AIC: 801.99
##
## Number of Fisher Scoring iterations: 2
```

Great, now that we've created a model, we can see that playing aggressive both at HOME and AWAY may not be a good idea. There is a negative, although very slightly, correlation between winning and `num_aggressive_turnovers`.

The model is further interpretted below:

$$P(win) = \frac{1}{1 + e^{-x}}$$

Where:

$x = 0.65 - 0.0036 *$(number of aggressive turnovers) + 0.883 * (is at home) - 0.033 * (number of aggressive turnovers * is at h

```
1/(1 + exp(-0.65))
```

```
## [1] 0.6570105
```

The above is the chances of winning under the following conditions:

- Away
- Zero aggressive turn overs

```
1/(1+exp(-(0.65 + 0.883)))
```

```
## [1] 0.8224448
```

The above is the chances of winning under the following conditions:

- Home
- Zero aggressive turn overs

NOTE: There are caveates that go along with this interpretation

- There may be confounders at play
- Many unaccounted variables
- There are other caveates

Model that includes passess as a proxy for possession time

```
# Creating logistic model:
model3 <- glm(is_win ~ num_aggressive_turnovers + is_home + num_passes + is_home*num_aggressive_turnover

# Displaying results:
summary(model3)
```

```
##
## Call:
## glm(formula = is_win ~ num_aggressive_turnovers + is_home + num_passes +
##     is_home * num_aggressive_turnovers * num_passes, data = aggressive_df)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8054  -0.2809  -0.1620   0.3996   0.9980
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                   8.761e-02  6.598e-01
## num_aggressive_turnovers                      2.190e-03  5.019e-03
## is_homeTRUE                                   1.772e+00  1.068e+00
## num_passes                                    1.400e-03  1.824e-03
## num_aggressive_turnovers:is_homeTRUE         -9.369e-03  7.814e-03
## is_homeTRUE:num_passes                       -2.375e-03  2.710e-03
## num_aggressive_turnovers:num_passes          -1.453e-05  1.362e-05
## num_aggressive_turnovers:is_homeTRUE:num_passes  1.529e-05  1.965e-05
##                                               t value Pr(>|t|)
## (Intercept)                                     0.133   0.8944
## num_aggressive_turnovers                        0.436   0.6628
## is_homeTRUE                                     1.659   0.0977 .
## num_passes                                      0.768   0.4430
## num_aggressive_turnovers:is_homeTRUE           -1.199   0.2310
## is_homeTRUE:num_passes                         -0.876   0.3812
## num_aggressive_turnovers:num_passes            -1.067   0.2865
## num_aggressive_turnovers:is_homeTRUE:num_passes   0.778   0.4368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1864095)
##
##     Null deviance: 153.16  on 679  degrees of freedom
## Residual deviance: 125.27  on 672  degrees of freedom
## AIC: 797.44
##
## Number of Fisher Scoring iterations: 2
```

There are a couple things that are troubling about the above model:

- Because of the additional interaction term, this model is a lot less intuitive and more difficult to interpret
- None of the coefficients are "significant"
- Johann, over the phone I said that passes were negatively correlated with winning, this can be shown below:

```
# Creating logistic model:
model4 <- glm(is_win ~ num_passes, data=aggressive_df)

# Displaying results:
summary(model4)
```

```
##
## Call:
## glm(formula = is_win ~ num_passes, data = aggressive_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4236  -0.3508  -0.3198   0.6330   0.7358
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4792620  0.0856048   5.599 3.14e-08 ***
## num_passes  -0.0003548  0.0002172  -1.633    0.103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2250192)
##
##     Null deviance: 153.16  on 679  degrees of freedom
## Residual deviance: 152.56  on 678  degrees of freedom
## AIC: 919.49
##
## Number of Fisher Scoring iterations: 2
```

The above result is super counter-intuitive. Teams that pass a lot are more likely to lose. This seems wrong to me. Perhaps this is because crappier teams require more passes to get it up the field? Maybe its because they are passing it between their defenders a lot of the time. The passes may not be up field passes. The passes may not have even been completed.

As a result, I would be very careful with saying anything to do with causality. We shouldn't say whether or not passing is good or bad. One suggestion may be to try to find the posession times for each team of each game instead of using passing as a proxy.

Given more time:

If given more time to work, I would try the following:

- See if I can tell a more compelling story or make a better model by "binning" each team in a match as an "aggressive" or "passive" team. We can do this by setting a threshold on the number of aggressive turnovers. All occassions that are under this threshold are considered "passive" - otherwise, they are considered aggressive. Then, with this new classification of a team, we can see whether or not an aggressive team has a better chance of winning by doing another `glm`
- Make visualizations for models. This would involve showing the predicted value of the `glm` in one color as well as the actual values in a different color.

Good luck everyone. Wish I could be there.