FC Cincinnati Attendance Forecasting

Kuan-Cheng Fu

2/26/2020

Introduction

The marketing department is looking to plan giveaways to help boost attendance, so they hope to figure out that they should focus on weekday or weekend games more. Meanwhile, they are also curious if opponents, the time the game starts, or win percentage at the time of the game have an effect on the home attendance. Therefore, we will be trying to analyze the Scores & Fixtures data of 2019 FC Cincinnati to provide some suggestions for the marketing department.

Data Source

In this case study, we will be primarily working with the Scores & Fixtures data of 2019 FC Cincinnati from https://fbref.com/. Considering that the marketing department is curious whether opponents have an effect on the home attendance, we will also prepare the Scores & Fixtures data of each opponent which the FC Cincinnati faced at home in 2019 season.

```r
Raw_Data <- data.frame(read.table(file = "data/Cincinnati_Data.txt", header = T, sep = ","))
Portland <- data.frame(read.table(file = "data/Portland_Data.txt", header = T, sep = ","))
Philadelphia <- data.frame(read.table(file = "data/Philadelphia_Data.txt", header = T, sep = ","))
SportingKC <- data.frame(read.table(file = "data/SportingKC_Data.txt", header = T, sep = ","))
RealSaltLake <- data.frame(read.table(file = "data/RealSaltLake_Data.txt", header = T, sep = ","))
Montreal <- data.frame(read.table(file = "data/Montreal_Data.txt", header = T, sep = ","))
NYRedBulls <- data.frame(read.table(file = "data/NewYorkRedBulls_Data.txt", header = T, sep = ","))
LAGalaxy <- data.frame(read.table(file = "data/LAGalaxy_Data.txt", header = T, sep = ","))
Houston <- data.frame(read.table(file = "data/Houston_Data.txt", header = T, sep = ","))
DCUnited <- data.frame(read.table(file = "data/DCUnitedStats_Data.txt", header = T, sep = ","))
NewEngland <- data.frame(read.table(file = "data/NewEngland_Data.txt", header = T, sep = ","))
Vancouver <- data.frame(read.table(file = "data/Vancouver_Data.txt", header = T, sep = ","))
NYCFC <- data.frame(read.table(file = "data/NYCFC_Data.txt", header = T, sep = ","))
Columbus <- data.frame(read.table(file = "data/Columbus_Data.txt", header = T, sep = ","))
TorontoFC <- data.frame(read.table(file = "data/TorontoFC_Data.txt", header = T, sep = ","))
Atlanta <- data.frame(read.table(file = "data/Atlanta_Data.txt", header = T, sep = ","))
Chicago <- data.frame(read.table(file = "data/Chicago_Data.txt", header = T, sep = ","))
OrlandoCity <- data.frame(read.table(file = "data/OrlandoCity_Data.txt", header = T, sep = ","))
```

Creating Predictors

In this section, we will create several predictors that will be used in our model. First, we will create a predictor called "Weekend" by mapping (weekend, weekday) to (1,0). Then, we will create a predictor called "Night.Game" by mapping (game time after 17:30, game time before 17:30) to (1,0). Finally, we will create a predictor called "Win.Percentage" by calculating our win percentage before the game.

```
Weekend <- ifelse(Raw_Data$Day == "Sat" | Raw_Data$Day == "Sun", 1, 0)
Night.Game <- ifelse(Raw_Data$Time == "13:00" | Raw_Data$Time == "15:00" | Raw_Data$Time == "17:00", 0,
Win.Percentage <- vector()
for (i in 1:nrow(Raw_Data)) {
  if (i == 1){
    Win.Percentage[i] = 0
  } else {
    Win.Percentage[i] = sum(Raw_Data[1:(i-1),]$Result == "W")/nrow(Raw_Data[1:(i-1),])
  }
}
```

Creating the Analyzed Data

In the beginning of this section, we will combine the Scores & Fixtures data of 2019 FC Cincinnati with the predictors that we created in the previous section. After that, we will create the analyzed data of game-by game home attendance for 2019 FC Cincinnati including “Opponent”, “Night.Game”, “Weekend”, “Win.Percentage”, and “Attendance”.

```
Raw_Data <- cbind(Weekend, Night.Game, Win.Percentage, Raw_Data)
Analyzed_Data <- Raw_Data[Raw_Data$Venue == "Home", c("Opponent","Night.Game","Weekend","Win.Percentage
rownames(Analyzed_Data) <- 1:nrow(Analyzed_Data)
Analyzed_Data
```

```
##          Opponent Night.Game Weekend Win.Percentage Attendance
## 1        Portland          0       1      0.0000000      32250
## 2    Philadelphia          1       1      0.5000000      25867
## 3      Sporting KC          0       1      0.4000000      26023
## 4  Real Salt Lake          1       0      0.2857143      26416
## 5        Montreal          0       1      0.1818182      26258
## 6    NY Red Bulls          1       1      0.2307692      28290
## 7       LA Galaxy          1       1      0.1875000      32250
## 8         Houston          1       1      0.1666667      26276
## 9      D.C. United          1       0      0.2500000      28774
## 10    New England          1       1      0.2380952      25095
## 11      Vancouver          1       1      0.2173913      27106
## 12          NYCFC          1       1      0.2000000      27273
## 13       Columbus          1       1      0.1923077      30611
## 14     Toronto FC          1       1      0.1785714      25339
## 15        Atlanta          1       0      0.2000000      24774
## 16        Chicago          1       1      0.1935484      26466
## 17   Orlando City          0       1      0.1875000      25652
```

However, after observing the analyzed data, we find that FC Cincinnati only face each opponent once at home in the 2019 season. Therefore, we decide to consider the effect of each opponent’s win percentage berfore the game on home attendance instead of directly using opponents as factors. Hence, we will update the analyzed data by adding a new predictor called “Opponent.Win.Percentage” by calculating each opponent’s win percentage before the game.

```
Opponent.Win.Percentage <- vector()
z = list(Portland,Philadelphia,SportingKC,RealSaltLake,Montreal,NYRedBulls,LAGalaxy,Houston,DCUnited,Ne
for (i in 1:length(z)) {
  n = which(z[[i]]$Venue == "Away" & z[[i]]$Opponent == "FC Cincinnati")
  Opponent.Win.Percentage[i] = sum(z[[i]][1:(n-1),]$Result == "W")/nrow(z[[i]][1:(n-1),])
}
Analyzed_Data <- cbind(Opponent.Win.Percentage, Analyzed_Data[2:5])
Analyzed_Data
```

```
##    Opponent.Win.Percentage Night.Game Weekend Win.Percentage Attendance
## 1                0.0000000          0       1      0.0000000      32250
## 2                0.2500000          1       1      0.5000000      25867
## 3                0.5000000          0       1      0.4000000      26023
## 4                0.2857143          1       0      0.2857143      26416
## 5                0.5000000          0       1      0.1818182      26258
## 6                0.3846154          1       1      0.2307692      28290
## 7                0.5625000          1       1      0.1875000      32250
## 8                0.4705882          1       1      0.1666667      26276
## 9                0.3809524          1       0      0.2500000      28774
## 10               0.3333333          1       1      0.2380952      25095
## 11               0.1666667          1       1      0.2173913      27106
## 12               0.4347826          1       1      0.2000000      27273
## 13               0.2500000          1       1      0.1923077      30611
## 14               0.3571429          1       1      0.1785714      25339
## 15               0.5172414          1       0      0.2000000      24774
## 16               0.2903226          1       1      0.1935484      26466
## 17               0.2812500          0       1      0.1875000      25652
```

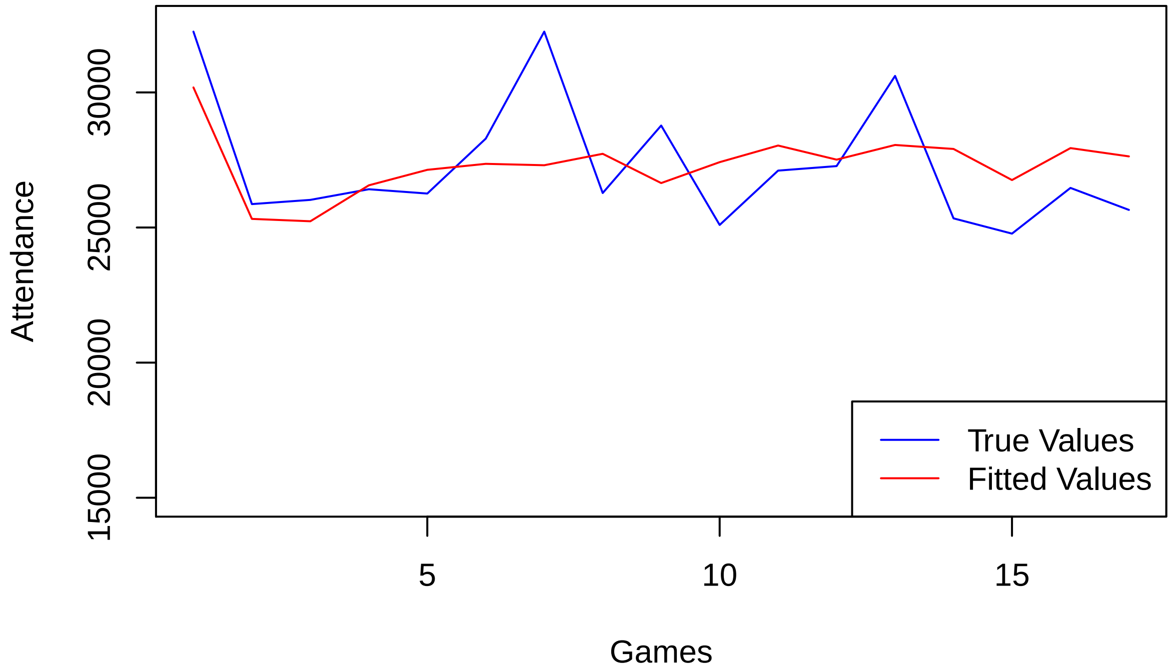
Building and Evaluating the Poisson Regression Model

In this case, we are going to regard the home attendance as count data. Therefore, we will use Poisson Regression Model to fit the home attendance. In this model, "Attendance" will be the response while "Opponent.Win.Percentage", "Night.Game", "Weekend", and "Win.Percentage" will be the predictors.

```
fit.poisson <- glm(Attendance ~ as.factor(Night.Game) + as.factor(Weekend) + Win.Percentage + Opponent.
```

From the fitted values versus true values plot, the trend of the home attendance could basically be captured by our Poisson Regression Model. However, it's obvious that our model underestimates the attendance of about three games. We suppoose that there might be some nice giveaways provided in those games, which does not be considered in our model but could be a potential variable that might has an effect on the home attendance.

```
plot(1:nrow(Analyzed_Data), Analyzed_Data$Attendance, type="l", xlab = "Games", ylab="Attendance", col=
lines(1:nrow(Analyzed_Data), fit.poisson$fitted.values, col="red")
legend("bottomright", c("True Values", "Fitted Values"), col = c("blue", "red"), lty = c(1, 1))
```

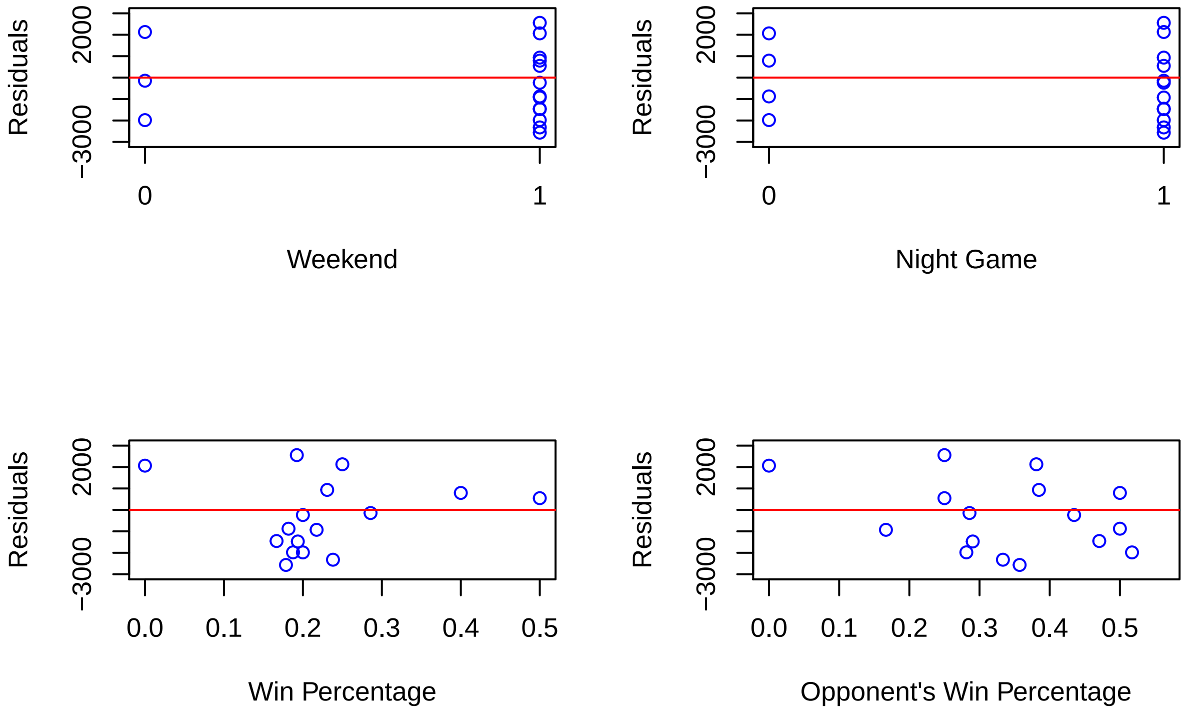


Furthermore, we will look at the deviance goodness of fit test for our model. Generally, if the model is correct, and since the model is a sub-model of the saturated model, then the deviance of the model has a chi-squre distribution with n-p degrees of freedom. In our case, the p-value is approximately equal to zero. This suggests that we have evidence to reject the null hypothesis which is that our model fits well. On the basis of the test result, Poisson Regression Model might not be the most appropriate model for our dataset. Nevertheless, we still want to figure out what the possible problems are.

```
pvalue <- pchisq(q = fit.poisson$deviance, df = nrow(Analyzed_Data)-5, lower.tail = F)
```

Therefore, we decide to look at the residual plots against each predictor. From the residual plots, the residuals basically bounce around the zero line, and are loacted from -3000 to 3000. This suggests that our model is somehow reasonable, although several residuals seem to be too far from the zero line. Given only 17 observations in our analyzed data, we think that the Poisson Regression Model is acceptable for modeling the home attendance at this stage.

```
par(mfrow=c(2,2))
plot(Analyzed_Data$Weekend,Analyzed_Data$Attendance-fit.poisson$fitted.values, xlab = "Weekend", ylab="
axis(side=1, at=seq(0, 1, by=1))
abline(h=0, col="red")
plot(Analyzed_Data$Night.Game,Analyzed_Data$Attendance-fit.poisson$fitted.values, xlab = "Night Game", 
axis(side=1, at=seq(0, 1, by=1))
abline(h=0, col="red")
plot(Analyzed_Data$Win.Percentage,Analyzed_Data$Attendance-fit.poisson$fitted.values, xlab = "Win Percen
abline(h=0, col="red")
plot(Analyzed_Data$Opponent.Win.Percentage,Analyzed_Data$Attendance-fit.poisson$fitted.values, xlab = "
abline(h=0, col="red")
```



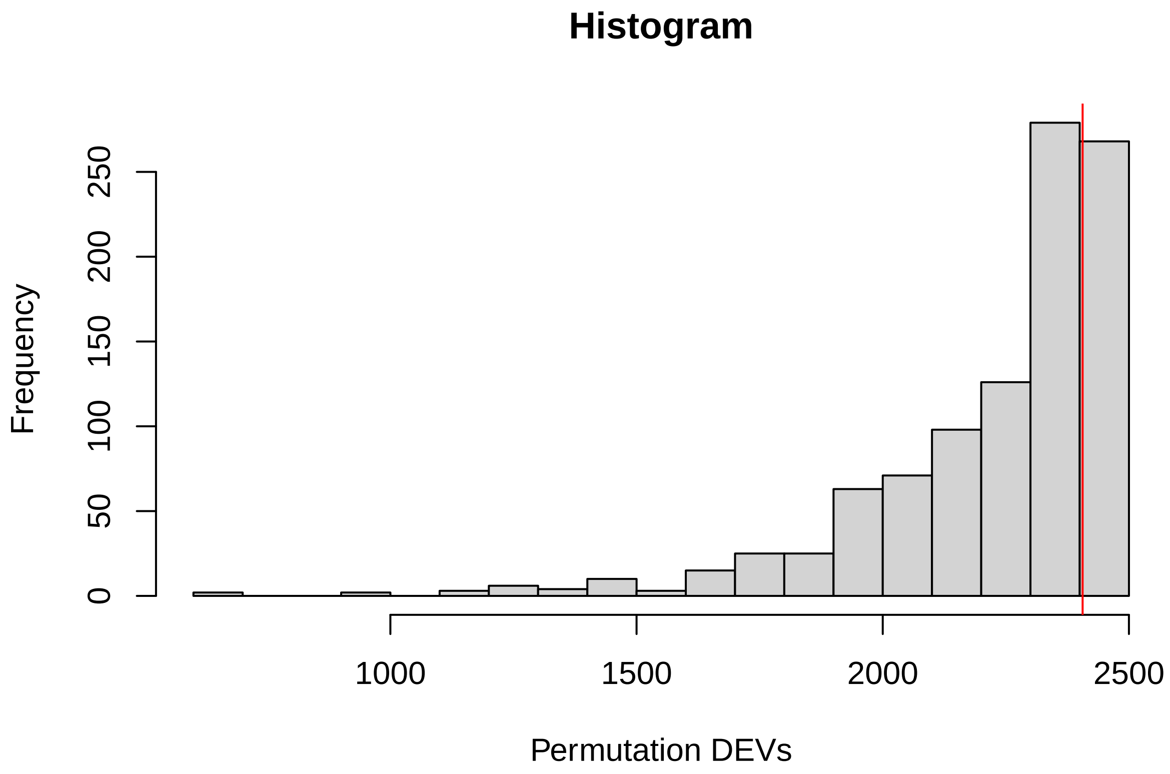
Determining the Impact of Weekend Games

In this section, we will apply permutation test to determine whether weekend games have a siginificant impact on the home attedance or not. In this case, we will regard the deviance as test statistic. First, we could compute the original deviance by using the original model in the previous section. Then, we shuffle the order of the predictor "Weekend" and recompute a new deviance called permuted deviance. After permuting the order of the predictor "Weekend" 1000 times, we will get 1000 permuted deviances. Finally, we could look at the distribution of those permuted deviances. If the orginal deviance isn't lower than most permuted deviances, then the original model isn't doing better with true values of "Weekend", which means weekend games don't have a siginificant impact on the home attedance.

```
set.seed(1)
DEV = vector()
for (i in 1:1000) {
  Analyzed_Data[1:nrow(Analyzed_Data),6] = sample(Analyzed_Data$Weekend, size = nrow(Analyzed_Data), re
  colnames(Analyzed_Data)[6] = "Weekend_new"
  model <- glm(Attendance ~ as.factor(Night.Game) + as.factor(Weekend_new) + Win.Percentage + Opponent.
  DEV[i] = model$deviance
}
```

Therefore, on the basis of the conception of permutation test and from the histogram below, the orginal deviance (red line) isn't lower than most permuted deviances. Hence, we could conclude that weekend games don't have a statistically significant impact on the home attendance given our model and the analyzed date we created.

```
hist(DEV, breaks = 25, main = "Histogram", xlab = "Permutation DEVs")
abline(v=fit.poisson$deviance, col = "red")
```



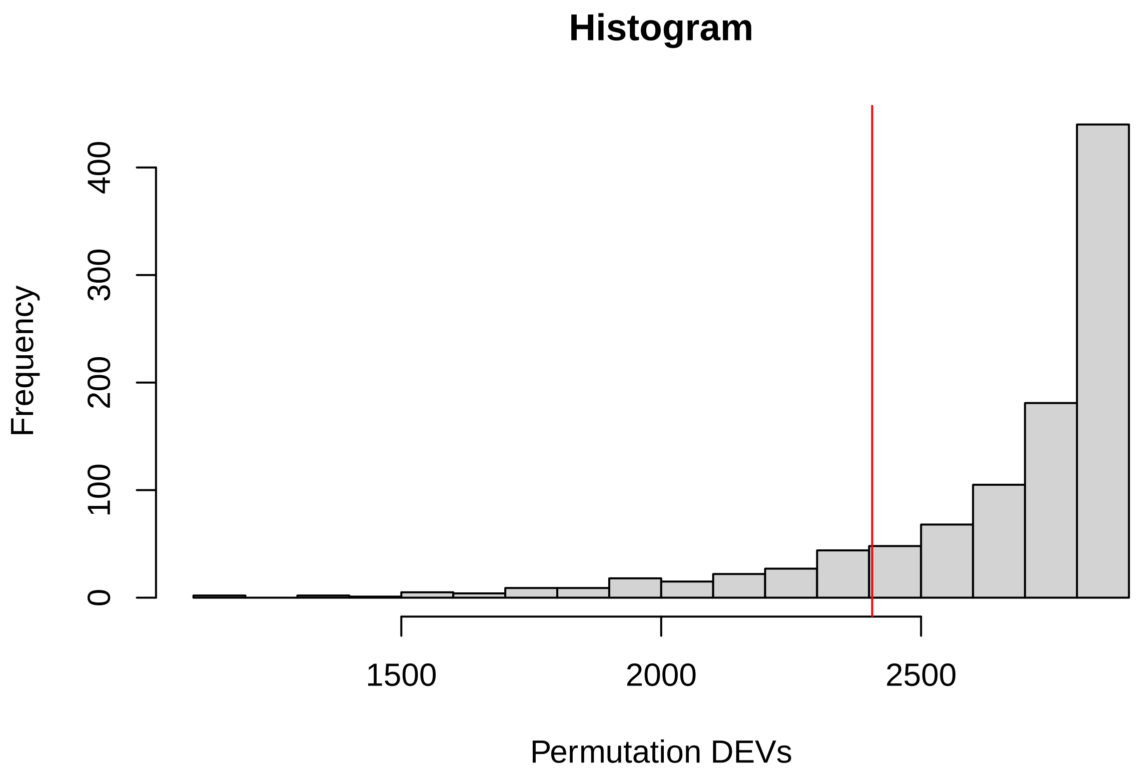
Determining the Impact of Win Percentage

In this section, we will also apply permutation test to determine whether win percentage has a siginificant impact on the home attedance or not. In this case, we will also regard the deviance as test statistic. The process is similar as the previous section. The only difference is that we will shuffle the order of the predictor “Win.Percentage”. After permuting the order of the predictor “Win.Percentage” 1000 times, we will also get 1000 permuted deviances. In the end, we would also look at the distribution of those permuted deviances. If the orginal deviance isn’t lower than most permuted deviances, then the original model isn’t doing better with true values of “Win.Percentage”, which means win percentage doesn’t have a siginificant impact on the home attedance.

```
set.seed(1)
DEV = vector()
for (i in 1:1000) {
  Analyzed_Data[1:nrow(Analyzed_Data),6] = sample(Analyzed_Data$Win.Percentage, size = nrow(Analyzed_Dat
  colnames(Analyzed_Data)[6] = "Win.Percentage_new"
  model <- glm(Attendance ~ as.factor(Night.Game) + as.factor(Weekend) + Win.Percentage_new + Opponent.W
  DEV[i] = model$deviance
}
```

Therefore, from the histogram below and the comparison with the histogram in the previous section, the orginal deviance (red line) is lower than most permuted deviances. Hence, we could conclude that win percentage has an impact on the home attendance given our model and the analyzed date we created. As a result, our win percentage is more important than either weekday or weekend games for helping boost the home attendance.

```r
hist(DEV, breaks = 20, main = "Histogram", xlab = "Permutation DEVs")
abline(v=fit.poisson$deviance, col = "red")
```



Further Analysis

We definitely need more data to discuss this problem. Since the FC Cincinnati just joined the league in 2019, we believe that more data will be collected through the games with each opponent in the next few years. After that, we could apply ANOVA to figure out whether there is a difference between the average home attendance for all opponents. Moreover, we might get a better forecasting result by utilizing KNN because the prediction only depends on the data points which have similar conditions (e.g. game time, opponent, weekend/weekday) instead of being influenced by all the data points.