MLBAttendance

Nate Crumbaker

3/8/2021

## Packages

require(xlsx)

## Loading required package: xlsx

require(ggplot2)

## Loading required package: ggplot2

require(knitr)

## Loading required package: knitr

require(UsingR)

## Loading required package: UsingR

## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

##   
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':  
##   
## cancer

require(dplyr)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Hmisc':  
##   
## src, summarize

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

require(stargazer)

## Loading required package: stargazer

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

require(car)

## Loading required package: car

## Warning: package 'car' was built under R version 4.0.5

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## Abstract

<<<<<<< HEAD The problem that this project seeks to address is whether weather factors have an affect on attendance to Major League Baseball games. It is important to determine this because the MLB relies heavily on attendance. Other researchers have already conclude that some factors due have affect on however we will expand on those factors (Ge, 2020). Our data comes from the websites Baseball Reference and Wunderground. This project concluded that we were unable to determine that weather has a significant affect on attendance and we recommend that more research is done into the area.

## Introduction

Weather has always been seen as a determinant to attendance of a sports game such as the Major League Baseball. Logically it makes sense to believe that line of thinking because who would want to go and sit through a game while it is raining. However this study aims to address the that line of thinking and see if it is actually backed by empirical data and evidence. This study will utilize multiple linear regression analysis to test select factors commonly believed to affect on weather to see if they statistically support the claims.

## Literature Review

The Major League Baseball(MLB) is a professional sports league located in North America that plays baseball games through out the months of April to October. Naturally we a lot of games being played outside it can be believed that weather might have an effect on attendance. Attendance is an important factor for the MLB because attendance will help determine ticket prices and whether to relocate teams to a different city. Due to the importance of attendance in the MLB there is already some studies done in this field. Qi Ge ran a study in which he found factors such as average temperature and precipitation due have an affect on attendance (Ge, 2020). This study seeks to expand on that idea by adding in more weather factors such as Wind Speed, Dew Point and Humidity.

## Theory

For this project our Hypothesis that this study seeks to answer is that Weather does have a signifigicant affect on determing Attendance.

## Data

The Data that we will be using to test this hypothesis will be the Atlanta Braves Home games attendance from 2017 to 2019 seasons in the Truist Park baseball field. First we import the attendance data from Sports Reference, Next we import the weather data from Wunderground.

myfile1 <- "ATL2019Attendance.xlsx"  
myfile2 <- "ATL2019Weather.xlsx"  
ATL2019Attendancedata <- read.xlsx2(myfile1, sheetName="ATL2019Attendance")  
ATL2018Attendancedata <- read.xlsx2(myfile1, sheetName="ATL2018Attendance")  
ATL2017Attendancedata <- read.xlsx2(myfile1, sheetName="ATL2017Attendance")  
ATL2017Weatherdata <- read.xlsx2(myfile2, sheetName="2017")  
ATL2018Weatherdata <- read.xlsx2(myfile2, sheetName="2018")  
ATL2019Weatherdata <- read.xlsx2(myfile2, sheetName="2019")

After importing the data from excel files next we will merge the Weather Data and Attendance data in the yearly data. Following that we then merge all the yearly data into a the total data. Finally we then remove all none home games due to those games being the ones not located in Atlanta

ATL2019 <- merge(ATL2019Weatherdata, ATL2019Attendancedata, by= "Date")  
ATL2018 <- merge(ATL2018Weatherdata, ATL2018Attendancedata, by= "Date")  
ATL2017 <- merge(ATL2017Weatherdata, ATL2017Attendancedata, by= "Date")  
ATL <- rbind(ATL2019, ATL2018, ATL2017)  
names(ATL) <- c("Date", "MaxTemp", "AvgTemp", "MinTemp", "MaxDewPoint", "AvgDewPoint", "MinDewPoint", "MaxHumidity", "AvgHumidity", "MinHumidity", "MaxWindSpeed", "AvgWindSpeed", "MinWindSpeed", "TotalPrecipitation", "Game", "X", "Team", "Away", "Opp", "Attendance")  
ATLHome <- ATL[!(ATL$Away==1),]  
i <- c(2,3,4,5,6,7,8,9,10,11,12,13,14,20)  
ATLHome[ , i] <- apply(ATLHome[ , i], 2, function(x) as.numeric(as.character(x)))

After all the data cleaning is done here is what the data set currently looks like

stargazer(ATLHome, type = 'html')

Statistic

N

Mean

St. Dev.

Min

Pctl(25)

Pctl(75)

Max

MaxTemp

243

84.634

7.313

53

80

90

99

AvgTemp

243

75.392

7.040

43.100

72.250

80.100

87.500

MinTemp

243

67.416

7.863

36

65

72

78

MaxDewPoint

243

66.593

8.039

26

64

72

76

AvgDewPoint

243

62.935

8.959

23.100

59.550

69.150

73.000

MinDewPoint

243

58.802

10.459

21

54.5

67

71

MaxHumidity

243

86.733

8.870

55

82

93

100

AvgHumidity

243

67.824

11.953

37.500

58.800

76.950

96.000

MinHumidity

243

46.539

12.815

17

38

54

84

MaxWindSpeed

243

14.551

5.176

6

12

16

40

AvgWindSpeed

243

7.429

2.622

2

5.6

8.7

20

MinWindSpeed

243

1.749

2.638

0

0

3

15

TotalPrecipitation

243

0.142

0.379

0

0

0.1

3

Attendance

243

31,752.890

7,177.981

16,049

25,145

37,871

43,619

## Methodology

Now that the data is cleaned we can begin the linear regression analysis in this project. First we create a linear regression model using the lm function which contains all the variables we will be using in the model for testing. Then we use the vif function to check for multicollinearity using the vif function from the car package. Our goal is to get every variable to have less than a value of 10.

AttendanceRegression <- lm(Attendance ~ MaxTemp + AvgTemp + MinTemp +   
MaxDewPoint + AvgDewPoint + MinDewPoint +   
MaxHumidity + AvgHumidity + MinHumidity +   
MaxWindSpeed + AvgWindSpeed + MinWindSpeed + TotalPrecipitation, data = ATLHome)  
vif(AttendanceRegression)

## MaxTemp AvgTemp MinTemp MaxDewPoint   
## 32.226059 255.808091 34.800739 39.337054   
## AvgDewPoint MinDewPoint MaxHumidity AvgHumidity   
## 386.742825 34.872750 8.476051 124.091444   
## MinHumidity MaxWindSpeed AvgWindSpeed MinWindSpeed   
## 18.327854 2.313665 6.055700 3.703357   
## TotalPrecipitation   
## 1.361271

Now we remove the variables will a high multicollinearity due to the interconnectedness between the variables. For the first round of removal we will be removing AvgTemp AvgDewPoint and AvgHumidity

AttendanceRegression <- lm(Attendance ~ MaxTemp + MinTemp + MaxDewPoint + MinDewPoint + MaxHumidity + MinHumidity + MaxWindSpeed + AvgWindSpeed + MinWindSpeed + TotalPrecipitation, data = ATLHome)  
vif(AttendanceRegression)

## MaxTemp MinTemp MaxDewPoint MinDewPoint   
## 16.739677 15.206244 19.233502 23.336592   
## MaxHumidity MinHumidity MaxWindSpeed AvgWindSpeed   
## 4.706081 12.122967 2.107162 5.969236   
## MinWindSpeed TotalPrecipitation   
## 3.687330 1.245022

Next we remove the variable with the highest value still left which is MinDewPoint

AttendanceRegression <- lm(Attendance ~ MaxTemp + MinTemp + MaxDewPoint + MaxHumidity + MinHumidity + MaxWindSpeed + AvgWindSpeed + MinWindSpeed + TotalPrecipitation, data = ATLHome)  
vif(AttendanceRegression)

## MaxTemp MinTemp MaxDewPoint MaxHumidity   
## 9.936532 10.639175 17.570762 3.901923   
## MinHumidity MaxWindSpeed AvgWindSpeed MinWindSpeed   
## 4.314822 2.091700 5.888557 3.655344   
## TotalPrecipitation   
## 1.244079

Finally we remove MaxDewPoint which has a value still above 10

AttendanceRegression <- lm(Attendance ~ MaxTemp + MinTemp + MaxHumidity + MinHumidity + MaxWindSpeed + AvgWindSpeed + MinWindSpeed + TotalPrecipitation, data = ATLHome)  
vif(AttendanceRegression)

## MaxTemp MinTemp MaxHumidity MinHumidity   
## 7.161299 7.162586 1.867367 3.382237   
## MaxWindSpeed AvgWindSpeed MinWindSpeed TotalPrecipitation   
## 2.091438 5.769957 3.653226 1.243736

Now that every variable has a value less than 10 we are able to continue to view the results of the regression analysis

## Results

Down below is the results of the linear regression analysis

summary(AttendanceRegression)

##   
## Call:  
## lm(formula = Attendance ~ MaxTemp + MinTemp + MaxHumidity + MinHumidity +   
## MaxWindSpeed + AvgWindSpeed + MinWindSpeed + TotalPrecipitation,   
## data = ATLHome)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16892.9 -6185.9 694.9 6054.1 12157.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35465.5109 9832.0558 3.607 0.000378 \*\*\*  
## MaxTemp -266.2004 166.2518 -1.601 0.110684   
## MinTemp 382.2749 154.6318 2.472 0.014142 \*   
## MaxHumidity 0.7885 69.9891 0.011 0.991021   
## MinHumidity -158.0658 65.1974 -2.424 0.016092 \*   
## MaxWindSpeed 170.1545 126.9234 1.341 0.181348   
## AvgWindSpeed -333.7594 416.2042 -0.802 0.423417   
## MinWindSpeed 291.7903 329.1940 0.886 0.376324   
## TotalPrecipitation -1225.0779 1338.1006 -0.916 0.360853   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7067 on 234 degrees of freedom  
## Multiple R-squared: 0.06269, Adjusted R-squared: 0.03064   
## F-statistic: 1.956 on 8 and 234 DF, p-value: 0.05284

The first thing to look at is the Multiple R-Squared which is just 0.063 which means that the weather variables explains only 6.3% of the variance in attendance. Next is we look at the two variables that are significant on the 0.01 level which is MinTemp which means that for every 1 degree Fahrenheit increase of minimum temperature the amount of people attending will increase by 382 people. Another significant variable is MinHumidity which means for every 1% increase in the Humidity level the amount of people in attendance will decrease by 158. There are a lot of non-significant results in this model however that does not mean that the factors do or do not affect attendance it simply means that the model could not conclude that the factors have an affect on weather.

## Implications

The implications of this project means that more research will be needed to conclude if weather does have an affect attendance due to the high amount of non significant results and the low multiple R-Squared.

## Conclusions

The conclusion from this research project is that we were unable to determine whether weather has a significant affect on attendance towards MLB games. The two variables MinTemp and MinHumidity however did seem to have a significant affect on attendance while the Model as a whole did not.

## References

Atlanta, GA Weather History | Weather Underground. Wunderground.com. (2021). <https://www.wunderground.com/history/monthly/KATL/date/2019-4>.

2019 Atlanta Braves Schedule. Baseball Reference. (2021). <https://www.baseball-reference.com/teams/ATL/2019-schedule-scores.shtml>.

Ge, Q., Humphreys, B. R., & Zhou, K. (2020). Are fair weather fans affected by weather? Rainfall, habit formation, and live game attendance. Journal of Sports Economics, 21(3), 304-322.