Andrea Contreras, Jay Irby, Samuel Wright, Benjamin Lang

Professor Harrington

MAD2502

3 May 2023

Exploring the Impact of Ballpark Factors and Atmospheric Conditions on MLB Offensive Output

Abstract

Altitude, temperature, humidity, and a variety of other climatic factors affect air density, which in turn affects the flight of a baseball. We investigated the extent to which these climatic factors and a variety of other ballpark factors such as dimensions have an effect on the offensive output of a Major League Baseball Team in the 2022 season. Offensive output can be evaluated through a plethora of statistics; home runs, total runs scored, and the advanced metric wOBA (weighted on-base average) is helpful when evaluating total offensive production. Statistical analyses in Python including multiple regressions and Mann-Whitney U T-test were used to evaluate which ballpark metric impacted offensive metrics the most in the 2022 MLB season. We found no to little correlation for the relationship between all of our measured independent and dependent variables. We determined that using our data we cannot completely and accurately measure a correlation without a controlled environment. This generates various implications regarding the future of studies regarding the climatic impacts of offensive output in baseball, such as its translation into an experimental setting.

Introduction

Coors Field, a Major League Baseball field in Denver, Colorado, has earned a reputation as being a hitters' dream and a pitchers' nightmare. There is a widely held assumption that due to

the high altitude of the park (5280 feet above sea level), baseballs travel ten percent farther. To understand the backing behind this assumption, we must discuss the physics of baseball. In a perfect world, gravity would be the only force influencing the flight of a baseball. This presumption would make the flight of a baseball precise regardless of the atmospheric conditions associated with the location of a hit. However, we live in a world where atmospheric conditions exert aerodynamic forces on the flight of a baseball. Drag force introduces a negative acceleration on a baseball, and as the drag coefficient decreases, the baseball speed increases (Eaton 6). Terry Bahill of the University of Arizona highlights a mathematical equation that directly connects the air density of an atmosphere to the drag force on a baseball (Bahill 118). Additionally, air density varies inversely with altitude, meaning that altitude also varies inversely with the drag force of a ball (Bahill 118).

In understanding the effect of atmospheric conditions on the flight of a baseball, it is important to acknowledge the role that climatic factors play in reducing drag and increasing the speed of a baseball. However, is this resulting increase in speed enough to affect the offensive output of a baseball team?

Humidity shares a physics based relation with the flight of a baseball, where the humid air contains more water vapor/water molecules which push heavier molecules out of the way of the path of the ball, giving it a clearer flight through the air (Lynn). Therefore, humidity has a positive correlation with the ball flying further, which indirectly could lead to better hitting statistics. Humidity, however, can only matter if the ball is being hit outside, and there are a few stadiums that play indoors, which would heavily affect the air and its moisture content.

Across Major League Baseball, there are only loose requirements regarding the stadiums themselves. This stems from the beginnings of the sport where the stadiums had to be placed in

the middle of population centers to facilitate easy access to the ballparks. The stadiums were built in the middle of city blocks, or wherever the builders could find a place for them (Vickars). This caused stark differences between stadiums, whether it be fence distance, fence height, or even having a roof or not.

The fence distances are something specific we looked at as something that could affect offensive output. This variable is cut and dry, whereas the further the wall is, the harder you have to hit it to get it over the fence, and vice versa, the closer the wall is, the easier it is to get the ball over the fence.

Review of Previous Works

Many research studies have examined the relationship between altitude and home runs hit, which is a critical factor in offensive production. However, these research studies have contradictory results. Richardson of Penn State University, used a least-squares best fit to the runs per game to generate the results that "there are more home runs hit at high altitude than low altitude" (Richardson). Jason Bloch, on the other hand, states that "the variation in ballpark elevation cannot explain the variation in the number of home runs hit" and highlights Coors Field as the only notable exception (246).

Using statistical analysis and regression modeling in Python, we aim to build off of previous research regarding climatic conditions and ballpark factors and their variance with total runs, home runs, wOBA, and park factor to answer the question:

To what extent do specific atmospheric conditions and stadium qualities impact total offensive production in a ballpark?

Home runs are essential for successful offensive production and are a great metric for evaluating total offensive production. However, we will also use other dependent variables that

measure offensive output such as wOBA, total runs, and park factor to not only measure an absolute offensive statistic but also measure the offensive statistics of a team relative to where they play, which may help answer questions as to whether the climatic and ballpark factors of a baseball field play a notable role in a team's success. We will evaluate the extent each of the climatic and ballpark factors plays (both individually and as a whole) using multiple regression in Python. Additionally, we will use matplotlib to generate graphs to visualize the results.

Methods

Independent and Dependent Variables

We used statistical analyses such as multiple and linear regression along with the Mann-Whitney U t-test to derive the relationships between our independent and dependent variables.

The independent variables of climatic and ballpark factors that we considered include altitude (in feet), stadium dimensions (in feet), which include left field length away from home plate, center field length away from home plate, and right field length away from home plate, the average attendance per game, humidity (%), stadium type (indoor/outdoor), and average temperature (in Fahrenheit). We chose these specific independent variables because, in his research paper, Bloch highlights the relationship between air density and altitude, humidity, and temperature, which has implications for the speed of a baseball when hit. Additionally, we chose to include attendance as a dependent variable to analyze the interplay between this significant home-field advantage compared to climatic conditions. In a research article by Erin Smith of the University of Utah, she uses MLB data from the early 21st century to confer that "increased attendance has a significant effect on game statistics… leading to an increased likelihood of winning for the home team" (18), meriting a place for attendance as one of our dependent

variables. This can be assumed anyways, as attendance has ties to the loudness of the stadium, which directly affects the performance of both the batter and the pitcher, tying to possible boosts in hitting output. Additionally, attendance is tied to revenue, as the more people that go to games, the more money they pay to go to those games, and that money will be reinvested into the team, indirectly leading to better hitting performance.

In addition, the majority of previous literature we examined, including articles by Smith, Bloch, and Richardson, mention how dimensions of certain ballparks can cause differences in results, so they utilize methods to control for ballpark dimension differences, which is why we felt it necessary to include ballpark dimensions as an independent variable.

We used several different dependent variables to evaluate a team's offensive output. Home runs are an absolute measurement of a team's success at hitting the ball out of the park. We also used wOBA, which is an advanced metric that weights different ways of reaching base. For example, a home run is worth more than a single, so wOBA gives more weight to home runs in its calculation. wOBA measures a player's overall offensive contribution per plate appearance, and when compiled for a team, it provides a more comprehensive picture of their offensive performance than just looking at home runs scored.

Park factor is another metric we used to evaluate a team's performance. It compares the offensive performance of a team at home versus away games to determine the home-field advantage of a specific MLB team. Each stadium's park factor is segmented into a score that measures individual offensive outcomes. We would later massage this data to get an overall park factor score for each stadium.

Total runs is a relative metric that measures a team's offensive performance compared to other teams. For this study, we used data that is average based on a 100 scale. For example, a

total runs scored metric of 114 means that a stadium's total runs scored is 14% above the league average, while a metric of 90 means it's 10% below the league average.

Overall, our study used a combination of absolute and relative measures to evaluate a team's offensive output, including home runs, wOBA, park factor, and total runs.

Data Scraping and Extraction

We compiled data for our dependent and independent variables from websites including the National Centers for Environmental Information website, BaseballSavant, and OnlyHomers, and used previously compiled data from research articles such as Bloch and Smith's research. We ran into issues trying to automatically scrape off of these websites. Using beautiful soup, we were unable to specify the exact tables we wanted to scrape from using the table's id or class. To resolve this issue we manually scraped data into a CSV file including the names of MLB teams, their home stadium, location, average temperature and humidity (in afternoons March-October, since that is when the majority of MLB games are played), altitude, dimensions, attendance, home runs, total runs, wOBA, park factor, stadium type (indoor or outdoor), and whether or not the stadium is at sea level. We were, however, successful in automatically scraping from FantasyPros using BeautifulSoup. This allowed us to pull a table of all the Park Factor scores for each stadium in the league.

Data Massaging

After loading in the CSV file that we manually created along with the scraped table from FantasyPros we began massaging the data. We started by averaging the homerun data from the past two seasons into a single column to get a more accurate measure of homerun trends. We also took the various park factor scores for different offensive events and averaged them together to create a single overall park factor score for each stadium. Some of the column names were

adjusted for readability and certain columns had their data types changed due to incorrect type classification when we imported the tables. In the park factor table we also had to separate the team name and stadium which had a comma delimiter. Once both tables had identical team name columns we were able to merge the two data frames into a single frame containing the manually and automatically scraped data. The data frames were merged by linking their respective team name columns. In this new data frame we specified the columns we wanted to keep and use later in our statistical analysis. We also sorted the entire data frame in descending order by the homeruns column. This increased readability and allowed us to try to find a visual correlation in the data between our independent variables and the homerun total.

Statistical Analyses - Multiple Regression

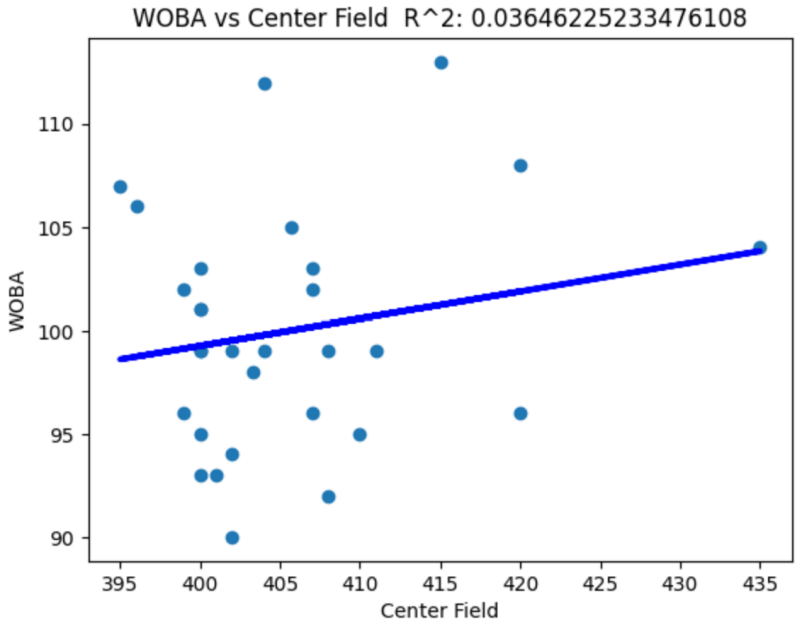
Our primary method of statistical analysis was multiple regression. We felt that multiple regression would be the best way to measure correlation between the independent variables and our three dependent variables. Python's statsmodel.api package came in very useful in the construction of or regression models. The package allowed us to easily create tables showing the coefficients, p-values, and R squared value of the model. The following is an example of one of the multiple regression results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Park Factor   R-squared:                       0.455
Model:                            OLS   Adj. R-squared:                  0.313
Method:                 Least Squares   F-statistic:                     3.205
Date:                Thu, 27 Apr 2023   Prob (F-statistic):            0.0195
Time:                        15:21:30   Log-Likelihood:                 29.606
No. Observations:                  30   AIC:                            -45.21
Df Residuals:                      23   BIC:                            -35.40
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            -0.0378      1.432     -0.026      0.979      -2.999       2.924
Altitude       7.981e-05   2.53e-05      3.154      0.004    2.75e-05       0.000
Attendance    -1.286e-06   2.04e-06     -0.630      0.535   -5.51e-06    2.94e-06
Humidity          0.0002      0.003      0.083      0.935      -0.005       0.005
Left (feet)       0.0003      0.001      0.221      0.827      -0.003       0.003
Right (feet)     -0.0023      0.002     -1.073      0.294      -0.007       0.002
Center (feet)     0.0041      0.002      1.668      0.109      -0.001       0.009
==============================================================================
Omnibus:                        1.134   Durbin-Watson:                   1.973
Prob(Omnibus):                  0.567   Jarque-Bera (JB):                0.924
Skew:                           0.157   Prob(JB):                        0.630
Kurtosis:                       2.199   Cond. No.                     2.16e+06
==============================================================================
```

In this model summary we see the only independent variable with a statistically significant p-value is altitude. Unfortunately all of the other p-values are too high to be significant which combined with a low R-squared value of .455 shows us that the multiple regression model fails to model or identify any meaningful correlation between variables.

Statistical Analyses - Linear Regression

In addition to the multiple regression, we also ran linear regression on each individual independent variable. We used the same statsmodel.api package, but also plotted the data points to give a visual representation of the distribution of data and the regression line of best fit. This was accomplished using the matplotlib package for the graph along with the sklearn package to create the regression line. The following is an example graph from a linear regression model.



In the graph it is apparent that the data does not show any kind of meaningful trend. The distribution of points is too scattered from the line of best fit. In the regression summary we can confirm these observations as there is a p-value of over 0.3 indicating that Center Field distance does not have a statistically significant effect on the ball park's wOBA score.

OLS Regression Results

Dep. Variable:	WOBA	R-squared:	0.036
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	1.060
Date:	Thu, 27 Apr 2023	Prob (F-statistic):	0.312
Time:	18:49:06	Log-Likelihood:	-93.633
No. Observations:	30	AIC:	191.3
Df Residuals:	28	BIC:	194.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	47.1620	51.309	0.919	0.366	-57.940	152.264
Center (feet)	0.1303	0.127	1.029	0.312	-0.129	0.389

Omnibus:	1.300	Durbin-Watson:	1.741
Prob(Omnibus):	0.522	Jarque-Bera (JB):	1.159
Skew:	0.446	Prob(JB):	0.560
Kurtosis:	2.637	Cond. No.	2.01e+04

Statistical Analyses - Mann-Whitney U t-test

As seen in the regression of homeruns and altitude, there is a slight positive correlation between the two variables, albeit weak correlation. Initially, we looked at the altitudes of the stadiums and realized that a significant majority of the stadiums are located at sea level with essentially no significant altitude. Since a decent majority of the stadiums have no altitude, we wanted to run a t-test with 2 groups: sea level and non sea level stadiums. We wanted to investigate whether the stadiums with altitude yield more offensive production than the stadiums without it. Since our data is not normally distributed and contain differing variances, we decided to utilize the Mann-Whitney U t-test, which is nonparametric, meaning it does not require specific parameters such as normal distribution in the 2 groups. Additionally, the Mann-Whitney U is well-suited for small sample sizes. We iterated through the 'Altitude' column in the scraped csv file to find the indexes where the altitude is at least 200 feet and less than 200 feet. We used if statements and appended the respective indexes to empty lists. With the 2 lists with the indexes of which stadiums are at sea level and the ones that are not, we were able to further iterate through the lists to obtain desired dependent variable values for each stadium. We used scipy.stats with the mannwhitneyu() function to calculate the test statistic and p-values. The syntax is in our code and here is an example of a result, comparing sea level and non sea level stadiums with homerun values:

```
#use mannwhitneyu() with group1 and group2 as parameters
statistic, p_value = stats.mannwhitneyu(group1, group2)
print(f'Mann-Whitney U statistic: {statistic}') #print test statistic
print(f'p-value: {p_value}') #print p-value
```

```
Mann-Whitney U statistic: 131.5
p-value: 0.4278683341581393
```

A similar methodology was used to calculate the t-test between indoor and outdoor stadiums. The only difference in the code is that ‘0’s and ‘1’ values were attached to each stadium in the csv file under the ‘Stadium Type’ column. We again iterated through the stadium type column, extracting the indexes that are ‘0’ or ‘1’. The rest of the code is the same.

Results

After running the statistical analyses in Python, we were able to deduce that the differences in climatic and ballpark factors do not explain any variation in the offensive output of teams. While we did get a handful of lower p-values in our linear and multiple regression models, none of these models successfully described correlation between variables. In terms of the t-tests, we also did not obtain any statistically significant p-values to show correlation between the variables at hand. For each Mann-Whitney U t-test, the p-values were greater than the three usual p-values of 0.01, 0.05, 0.1. Therefore, we fail to reject the null hypothesis that there is no significant difference between sea level and non sea level stadiums, and indoor and outdoor stadiums, respectively.

Conclusions

Implications

While our statistical analysis did not find a significant correlation between our measured independent and dependent variables, our study provides valuable insights for future research in this area. Further studies using a larger and more diverse dataset could help validate or add to our findings. Moreover, implementing an experimental method to measure the effect of each independent variable on offensive output in a controlled environment could provide more accurate results. However, given the limitations of time and resources, we were unable to

conduct such an experiment and acknowledge that confounding variables can impact the results of such a study.

Our research has important implications for the field of baseball and sports facility construction. While we could not identify specific factors that are more likely to contribute to a high offensive output, our study highlights the need for further research in this area to assist scouts, coaches, and managers in the selection and mentoring of players who will compete in stadiums with a climatic advantage. Furthermore, identifying these factors and their relationship with offensive output could pave the way for the development of a Python probability analysis using packages such as Sci-kit Learn. This would enable construction companies that specialize in sports facility construction to input ballpark factors and climatic conditions and determine the potential offensive advantage of a ballpark in a specific location. Such analyses would be valuable for decision-making in the construction of new ballparks and the renovation of existing ones.

Limitations

In baseball, like any sport, much of the variability within the dependent variables stems from a plethora of outside factors that are difficult to account for. For example, a good offensive baseball team is going to hit homeruns, thus majorly impacting the overall offensive production for each stadium, especially the home stadiums of such teams. It is very difficult to accurately determine the effect of ballpark factors and climatic conditions on offensive output because we are only running multiple regressions and t-test, which can only speak for correlation, not causation, and attempts to relate factors which are not mutually exclusive. Most importantly, the majority of our factors ignore that regardless of whether a ball is hit hard enough to be a homerun, the effectiveness of the defense of the opposing team plays the biggest factor in

limiting our measurement of whether offensive output would be considered successful in any controlled situation.

Additionally, there is not much variability in the dependent variables that we measured. For example, the altitude of MLB stadiums are mostly within a thousand foot range, with the exception of Coors Field in Denver, which is 4000 feet higher above sea level than the second-highest Major League Ballpark. This limited sample size and little difference in certain climatic conditions and ballpark factors of the majority of major league ballparks creates a gap in our data that, had it been filled, would have provided us with a clearer picture on the effects of these independent variables on offensive output.

Final Remarks

While our statistical analyses did not reveal any significant correlations between the variables we measured, we know that baseball is more complex than what can be captured by numbers alone. In a way, this may be a good thing. If baseball could be explained by climatic factors and data sets, this would remove value from the importance of practice in the development of a skill set for this sport. At this point, the only thing we can confidently conclude is that Major League Baseball exists because it cannot be fully explained or predicted. Otherwise, it would not be entertainment.

Works Cited

Bloch, Jason M., et al. "Park Elevation and Long Ball Flight in Major League Baseball." *Journal of Recreational Mathematics*, vol. 34, no. 4, 2006, pp. 243-246. *ProQuest*, https://login.lphscl.ufl.edu/login?url=https://www.proquest.com/scholarly-journals/park-elevation-long-ball-flight-major-league/docview/89070489/se-2.

Bahill, Terry, et al., "Effects of Altitude and Atmospheric Conditions on the Flight of a Baseball." *International Journal of Sports Science and Engineering,* vol 3, Jan 2009, pp 109-128. *ResearchGate,* https://www.researchgate.net/publication/228668381_Effects_of_altitude_and_atmospheric_conditions_on_the_flight_of_a_baseball.

Current Results. "Average Annual Temperatures for Large US Cities." 2022, https://www.currentresults.com/Weather/US/average-annual-temperatures-large-cities.php

Eaton, Steven. *Modeling the Flight of a Baseball.* 1998. Carroll College, Honors Thesis. *Carroll Scholars,* https://scholars.carroll.edu/handle/20.500.12647/3435.

FantasyPros. "MLB Park Factors." 2022, https://www.fantasypros.com/mlb/park-factors.php.

Lynn, Sarah, "How hot weather helps baseballs fly farther. Heat helps baseball fly farther." 2019, https://spectrumlocalnews.com/mo/st-louis/weather/2022/06/14/heat-impacts-on-baseball-#:~:text=Humid%20air%20can%20hold%20more,than%20cold%20and%20dry%20games.

National Centers for Environmental Information. "Average Relative Humidity: Morning (M), Afternoon (A)." 2022, https://www.ncei.noaa.gov/pub/data/ccd-data/relhum20.dat

National Centers for Environmental Information, "U.S. Climate Normals." 2022 https://www.ncei.noaa.gov/pub/data/ccd-data/relhum20.dat

OnlyHomers. “MLB Stadium Home Run Leaders,” 2022, *OnlyHomers,* https://www.onlyhomers.com/ballparks

Smith, Erin E, and Groetzinger, Jon D., "Do Fans Matter? The Effect of Attendance on the Outcomes of Major League Baseball Games" *Journal of Quantitative Analysis in Sports*, vol. 6, no. 1, 2010. https://doi.org/10.2202/1559-0410.1192

Richardson, Eliza. “High Altitude Offense: An Empirical Examination of the Relationship Between Runs Scored and Stadium Elevation.” *Fall 2014 Baseball Research Journal*, 2014, *Society for American Baseball Research*, https://sabr.org/journal/article/high-altitude-offense-an-empirical-examination-of-the-relationship-between-runs-scored-and-stadium-elevation/

Vickars, Sam. “The irregular outfields of baseball.” 2019, *The DataFace*. https://thedataface.com/2019/04/sports/baseballs-irregular-outfields