# Modeling Variables Affecting the 2019 Attendance for the West Virginia Black Bears

By Preston Klaus, Sport Analytics, B.S.

December 9, 2020

**Introduction:**

This paper will discuss the effects of multiple variables on the 2019 home attendance for the West Virginia Black Bears, the Short Season Single A affiliate for the Pittsburgh Pirates. This paper will go through my model, my selection process of choosing my specific variables, my model correction process, my summary of regression results, and wrap it up with the conclusions I came to after running the attendance model. Because of the COVID-19 pandemic, owners lost the 2020 season and potentially beyond, leading to huge revenue losses for each club. The ultimate goal is for management of the West Virginia Black Bears to learn what factors influence attendance numbers at their games, in order to hopefully improve their numbers in the next few seasons, once there is a solution to getting fans within stadiums amidst the COVID-19 pandemic.

**Data Used:**

I collected a wide variety of data from different sources from the 2019 MiLB season. While not all of it was used within my final model, I collected data for a few distinct categories: promotions, time of day, weather, home and away team performance, and game time statistics. For the promotions category, I created four binary variables including the four most common types of promotions I saw on the Black Bears' schedule: Giveaways, Kids Activities, Fireworks, and Auctions. While I

considered creating more categories, I decided there was not enough of that particular promotion to justify inclusion. I also considered a miscellaneous category for all other promotions, but concluded that the genres were too broad for any model to provide an accurate prediction of a miscellaneous promotions category. For time of day, I focused on the day of the week and the month the game was played on. I grouped day of the week into ‘Weekday’, for games Monday-Thursday and ‘Weekend’, for games Friday-Sunday. However, I ended up using day of the week as a dummy variable in my model, instead of the ‘Weekend/Weekday’ grouping because of the ambiguity of “the weekend” classification. For weather statistics, I collected average temperature and humidity data for Morgnatown, WV. For home and away team performance I collected post game winning percentages for both the home and away team, and transformed the win/loss category for each individual game into a binary variable, in hopes of lagging both variables for my model. Lastly, for game time statistics, I collected multiple box score stats, such as game delay, which was made into a binary variable, data for game start time, and duration of the game. I transformed the game start time variable into an afternoon/evening binary dummy variable, with any start time after 3:00 EST being considered ‘Evening’, and any time before 3:00 EST being considered ‘Afternoon’. Also, to make it easier to run my model, I put the duration of the game into minutes played, instead of an hour:minute format.

**Model Overview:**

Before coming to a final model, I ran a series of ANOVA tests to test for the significance of certain variable alterations, notably for temperature squared and

humidity squared. While humidity squared proved to be a better fit, it still wasn't statistically significant to justify inclusion within the model. I also tested a series of lag terms. For example, I tested a lag of 1 on cumulative opponent and home team winning percentage, which makes sense because I collected the record of the team after they finished their game, as well as delay, and length of game. Ultimately, I settled with the inclusion of a lagged term of one for W/L in the previous home game, because I, and my initial model, didn't believe W/L% was a good measure to lag by any amount in such a small dataset. I also created an interaction term between temperature and humidity because when I tested it in with an ANOVA test, it was more statistically significant than each variable by itself, and I believe that both variables depend on each other and have a joint factor on attendance. After long consideration on what variables to include, I decided to run my model with the promotion variables of Giveaways, Kids Activities, excluding Fireworks and Auctions because of multicollinearity concerns. I also included the dummy variables for Time of Game, excluding Evening because Afternoon didn't have enough variables in order to exclude it, Day, excluding Sunday because most afternoon games were on Sundays, and Month, excluding June because August and July were both statistically significant in my model. Lastly I included the W/L lag term and the Temperature:Humidity interaction terms that I previously mentioned.

**Results of the Regression Model:**

After deciding on all of the variables that I wanted to

**West Virginia Black Bears Attendance Model Regression Results**

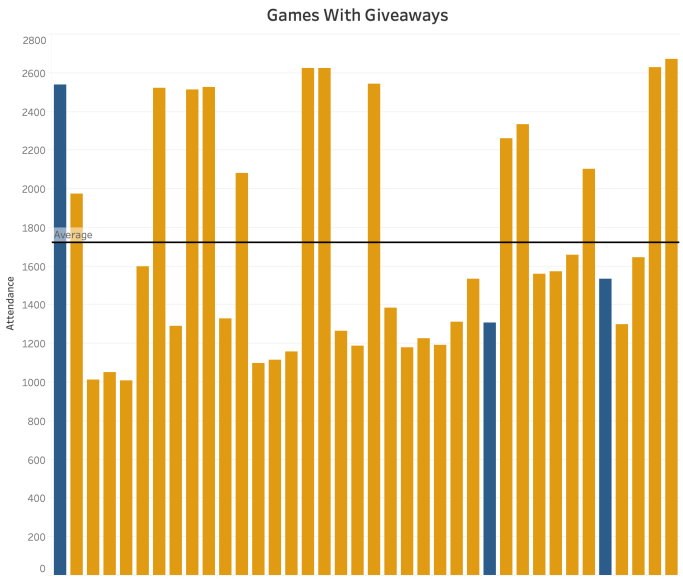
| Variable | Coefficient | Standard Error | t-statistic | p-value | Significance |
|---|---|---|---|---|---|
| Intercept | 439.791 | 642.0976 | 0.685 | 0.500234 | |
| Giveaways | -317.3687 | 282.6236 | -1.123 | 0.27305 | |
| Kids Activities | 258.2445 | 212.0893 | 1.218 | 0.235713 | |
| Lagged Black Bears Win/Loss | -16.9449 | 116.2511 | -0.146 | 0.88538 | |
| Afternoon Game | 418.656 | 246.0593 | 1.701 | 0.102342 | |
| Monday | -219.185 | 278.915 | -0.786 | 0.439978 | |
| Tuesday | -89.2439 | 248.9478 | -0.358 | 0.72325 | |
| Wednesday | -142.194 | 242.6749 | -0.586 | 0.563619 | |
| Thursday | 31.4845 | 219.0128 | 0.144 | 0.886945 | |
| Friday | 1097.362 | 241.397 | 4.546 | 0.000144 | *** |
| Saturday | 1144.3053 | 282.764 | 4.047 | 0.000501 | *** |
| July | 290.4163 | 151.8131 | 1.913 | 0.06828 | . |
| August | 457.3459 | 151.634 | 3.016 | 0.006153 | ** |
| AvgTemp:Humidity Interaction Term | 0.1145 | 0.1009 | 1.135 | 0.26817 | |

Significance codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

$R^2$ = 0.8047 Adjusted $R^2$ = 0.6943 p-value = 0.00002

F-Statistic: 7.29 on 13 and 23 degrees of freedom

include in my model, these are the results I got. From my model, some variables definitely carry more weight than others. Friday and Saturday had the most significance, with p-values that are statistically significant at any alpha level. July and August weren't too far behind with low p-values as well. The interaction term did not really have much of an impact on fans, showing that weather isn't much of a factor on how many people show up. Surprisingly, lagged win/loss and giveaways were shown to have a negative impact on attendance, albeit both had high p-values. Logically, one would think that both would increase attendance, but when looking at the data it makes sense. Many minor league baseball fans are casual fans who might not care about if they won or lost in their last home game. Additionally, out of the three giveaways this season, only the first game attracted a higher than average attendance. Combined with that being opening day and a game that included another fireworks promotion, giveaways didn't do well overall. All other signs on the coefficients make logical sense in the model. Lastly, the adjusted r-squared value, which is the corrected explanatory power of the model, is .6943. Therefore, we can conclude that 69.43% of the variation in attendance is explained by the regression variables, making my model a decent predictor of attendance. These are all important insights for the West Virginia Black Bears when looking to optimize attendance, and therefore profits. This final model



needed to be checked for multicollinearity, autocorrelation, and heteroscedasticity before the results could be statistically significant in conjunction with the Ordinary Least Squares (OLS) Assumptions of Regression. I assured these were satisfied by running a series of tests.

**Diagnostics:**

To satisfy the OLS assumptions, I had to test for multicollinearity, heteroscedasticity, and autocorrelation in my model. This was done at the same time I was correcting my model and excluding/including certain variables.

| Variable | VIF |
|---|---|
| Giveaways | 1.495391 |
| Kids Activities | 1.214432 |
| Lagged Black Bears Win/Loss | 1.924852 |
| Afternoon Game | 2.137437 |
| Monday | 2.122188 |
| Tuesday | 2.652017 |
| Wednesday | 2.929562 |
| Thursday | 2.386117 |
| Friday | 3.272826 |
| Saturday | 3.421433 |
| July | 1.984789 |
| August | 1.918615 |
| AvgTemp:Humidity Interaction Term | 1.938471 |

Multicollinearity is measured by looking at the variance inflation factors (VIF) of each variable. A VIF under 5 for each variable suggests that there shouldn't be much concern for multicollinearity for that variable. A VIF between 5-10 is more concerning, but not necessarily indicative of multicollinearity. A VIF of over 10 is very concerning and suggests that multicollinearity is an issue. When testing models, I had issues with multicollinearity for Auctions and Fireworks, which make sense because those promotions took place on weekends. When looking at the VIFs for my model, there are no multicollinearity issues with my model.

| Bruesch-Pagan Test | |
|---|---|
| BP | 9.2259 |
| degrees of freedom | 13 |
| p-value | 0.7557 |

| White's Test | |
|---|---|
| Statistic | 9.2874 |
| Parameter | 26 |
| p-value | 0.9989 |

Heteroskedasticity, where error terms are not independent and identically distributed, is another assumption that needs to be checked for within the model. For this I used two tests, a Bruesch-Pagan Test and White's Test, to completely verify that there is no evidence of heteroskedasticity. Both of these tests have assumptions that at a significant alpha level, e.g. 0.05 or 0.10, there is evidence of heteroskedasticity. Based on my very high p-values, there is no clear statistically significant evidence of heteroskedasticity within my model. Similarly to multicollinearity, I had heteroskedasticity issues in my previous models, but that was corrected when I addressed the previous multicollinearity issues.

Lastly, autocorrelation, or correlation between the same variables across different observations, was tested for within the model. Since this model includes time series variables, there is a larger threat of autocorrelation. I used both the Durbin-Watson Test and the Breusch-Godfrey Test to test for such autocorrelation within my data. Similar to the heteroskedasticity tests, a p-value of *over* a standard alpha level of 0.1, means that the significance test for autocorrelation fails and autocorrelation is not present. Based on my p-values, I did not have any evidence of autocorrelation within my model. Unlike multicollinearity and heteroskedasticity, I did not encounter issues with autocorrelation.

| Durbin-Watson Test | |
|---|---|
| lag | 1 |
| D-W Statistic | 2.2025 |
| p-value | 0.836 |

| Bruesch-Godfrey Test | |
|---|---|
| LM test | 0.5675 |
| degrees of freedom | 1 |
| p-value | 0.4513 |

**Conclusion:**

While my model can be useful for predicting attendance numbers for the West Virginia Black Bears, all readers of the results must keep in mind that the data was severely limited in calculating these variables, especially in terms of promotions. The West Virginia Black Bears were very creative in their promotions over the course of the season and, as previously mentioned, grouping the unique promotions together would hinder the cause of their success or failure. Having one, five, ten, or even twenty more years of promotional data would allow me to get a better sense of the attendance trends as it relates to promotions. Additionally, when correcting the model, I removed some of my limited data to give it more degrees of freedom, but the tradeoff is that the model may be slightly underfit. Although my 38 observations (37, when you exclude one variable because of the lag) meets the general rule of thumb for a good statistical sample size, acquiring more data would have been helpful.

With that being said, the main takeaway from the model in terms of the bottom line has to be the limited success of giveaways. Replacing giveaways with a different promotion, such as Kids Activities, or just scrapping that promotion all together would definitely save the staff time, effort, and money in the long run. The team should also push to schedule home games when they can on Fridays and Saturdays, especially in the months of July and August, as they drastically increase fan attendance at their games. Combining this with a promotion, possibly, could make these the biggest revenue days of the year. Lastly, the Black Bears had limited success with afternoon games. In the future, they should experiment with scheduling more games earlier in the day. Even with some inherent model limitations, combining these suggestions should lead to higher attendance numbers for the West Virginia Black Bears.

**Data Sources:**


"2019 Promotional Schedule." *Milb.com*, West Virginia Black Bears- MiLB, 2019.

Baseball Reference Writers. "2019 West Virginia Black Bears Statistics." *Baseball*, 2019, www.baseball-reference.com/register/team.cgi?id=032dc54e.

Weather Underground Team. "Morgantown, WV Weather Historystar_ratehome." *Weather Underground*, 2019, www.wunderground.com/history/monthly/us/wv/morgantown/KMGW/date/2019-8.

"West Virginia Black Bears Schedule: Schedule." *MiLB.com*, 2019, www.milb.com/west-virginia-black-bears/schedule/2019/fullseason.