Using Regression for The New York Mets

Matthew Barlow

2023-05-17

Executive Summary

For this report, I will analyze data for the New York Mets and compare their team performance for the 1993 to 2023 seasons to the Miami Marlins. Since the New York Mets play in the same conference as the Miami Marlins, doing a competitor analysis would help the Mets gain a better understanding of its competitive landscape. Therefore, the two objectives of the report are (1) to predict the Met's average home game attendance and (2) to predict whether the Mets have a better chance of performing better than the Marlins in a team stat percentage.

Building a model using intentional walks, team name, and the interaction variable, which considers how the two coefficients might depend on each other, results in the Mets having a higher predicted average home game attendance than the Marlins. However, though the overall results for the model are unlikely due to chance, having a p-value of 0.00, the model might lack practical significance. The p-value for the interaction coefficient is 0.138, which means that it brings little unique information to the model. In addition, the model has an adjusted R-Squared value of 0.471, indicating that the model explains about 47.1% of the variation in average home game attendance for the season but fails to explain the other 52.9%. The model also has a typical miss of 7,377 people when predicting average home game attendance, which could result in uncertainty in the coefficient's predictions.

On the other hand, building out a second model, using BABIP to predict whether the team is the Mets or the Marlins, results in a model that could be of more use in decision-making. Checking the regression model results in the p-values of 0.592 and 0.96 for the method one and method two test's respectively, so the scientists who wrote the tests would agree to use the model. The model also has a 64.52% accuracy, an improvement of 14.52% from the naïve model, which would predict the team using the majority of occurrences.

The first key takeaway from the report is that to predict the average home game attendance for the Mets, the team could add additional variables to make more confident predictions. For instance, the team might consider how Miami's local weather and unemployment rate could impact home game attendance. By predicting home game attendance more accurately, management could better estimate the amount of merchandise that needs to be stocked and hire enough staff to be ready for the season.

Second, by knowing that the Miami Marlins are more likely to have a better season BABIP than the Mets, the coaching staff can plan their team plays accordingly when facing the Marlins. The coaching staff can also take corrective action and examine what has helped its rival have a better BABIP as the Mets look to progress in the 2023 season.

Analysis/Findings

First, I will build a model to predict home game attendance using the all-model approach, which will consider all possible x-variables in the data set. However, I will simplify the model by limiting the number of variables to two, making the model easier to interpret and understand. The model with the lowest AIC might be the best model to use, given the data provided and the limited number of predicting variables.

The table below shows the AIC for three possible models that predict average home game attendance based on the team. The model with the lowest AIC, 1,108.6, uses the interaction variable "Team Name: New York Mets" and "Intentional Walks" as the predicting variables. In addition, the models shown in the second and third rows also use the "Team Name: New York Mets" variable. However, the second lowest AIC model uses "Base on Balls" as its second variable, while the model with the third lowest AIC uses "Ground Outs."

Since the models in the second and third rows have an AIC that is less than two higher, one might consider using any of the three models to predict average home game attendance since they would likely produce similar results. However, for my analysis, I will use the model with the lowest AIC shown in the first row that uses the "Team Name: New York Mets" and intentional walks variables as the predictors. I will also add the Miami Marlins to the model as a reference for predicting the Met's home game attendance.

AIC	Number of Variables	Terms
1,108.6	2	Team Name: New York Mets, Intentional Walks
1,110.1	2	Team Name: New York Mets, Base on Balls
1,110.4	2	Team Name: New York Mets, Ground Outs

The table on the next page summarizes the model that uses the team name and intentional walks variables to predict home game attendance and uses the interaction variable to consider how the number of intentional walks in a season might depend on the team. The interaction between the two variables is shown in the last row, labeled "Team Name: New York Mets (Intentional Walks)."

While the intercept and intentional walks variables have a p-value below 0.05, indicating that they bring new information to the model, the "Team Name: New York Mets" and the interaction variable "Team Name: New York Mets (Intentional Walks)" do not. Thus, the team name and interaction variables, having p-values of 0.696 and 0.138, respectively, might bring some redundant information that is already explained by intentional walks.

The model estimates that the New York Mets are predicted to have a higher average home game attendance than the Miami Marlins, as indicated by an estimate of 2,140.47 for the "Team Name: New York Mets" coefficient. Adding the estimate of 188.53 for the interaction

variable to the 21,140.74 estimate for the "Team Name: New York Met" results in an estimated positive slope of 2,329. Because of the positive interaction between the team name and intentional walks coefficients, the model predicts that the fitted regression line for the New York Mets will gradually increase at a higher incline than the Marlins as the coefficient for intentional walks increases.

Term	Estimate	Standard Error	Statistic	P-Value
Intercept	12,027.56	3,518.95	3.42	0.001
Team Name: New York Mets	2,140.47	5,451.61	0.393	0.696
Intentional Walks	174.74	86.82	2.01	0.049
Team Name: New York Mets (Intentional Walks)	188.53	125.42	1.50	0.138

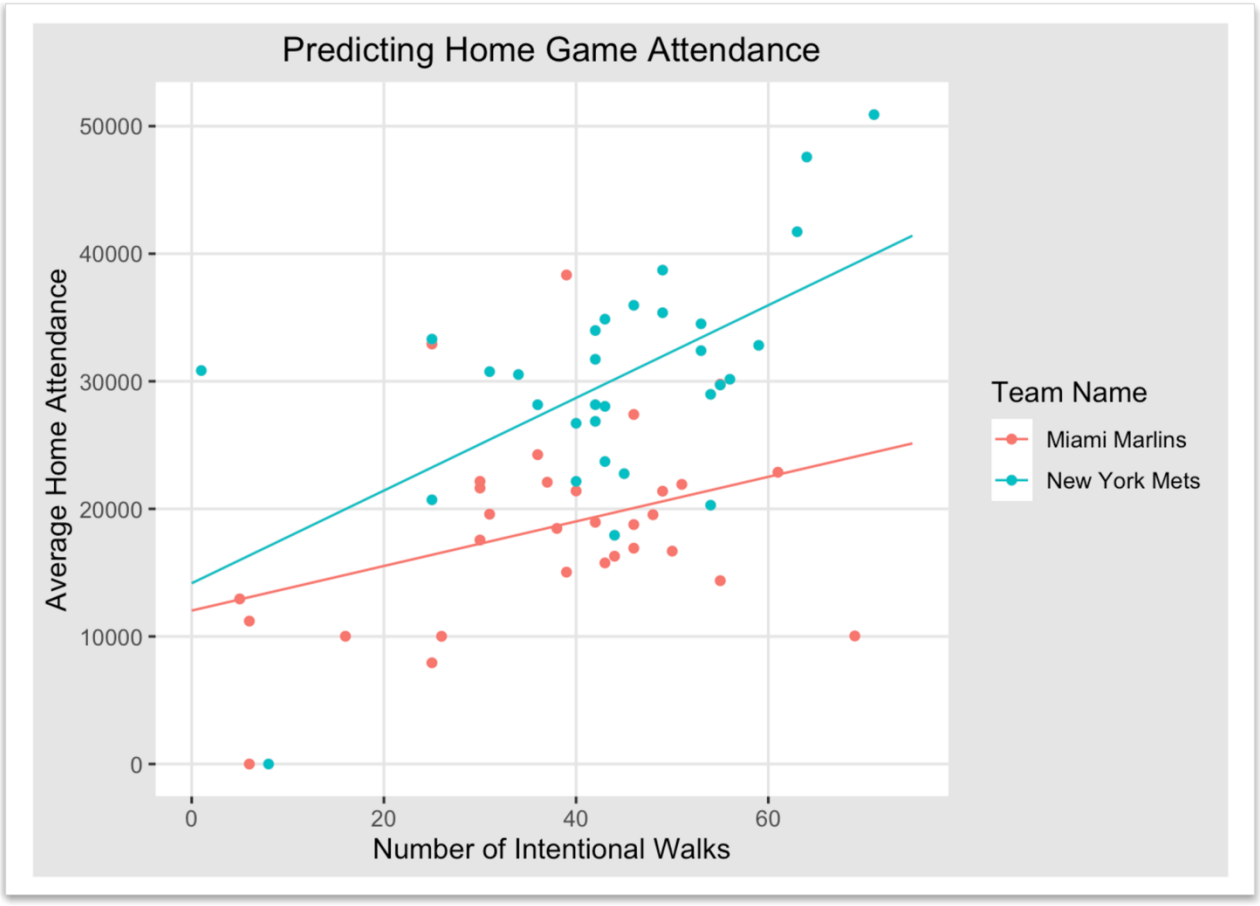
For instance, if the number of intentional walks for the season is 20, the model predicts the average home game attendance for the Marlins and the Mets to be 15,522.35 and 21,433.40 respectively, given they had otherwise identical seasons. Thus, at 20 intentional walks, the model predicts the difference in attendance to be 5,911.05 between the two teams. On the other hand, if the two teams had an otherwise identical season with 60 intentional walks, the model predicts the Marlins to have an average home game attendance of 22,511.92 while predicting the Mets to have an attendance of 35,964.14. Because the model factors in a dependency between the team name coefficient and the number of intentional walks in its predictions, the difference in the estimated home game attendance between the two teams is noticeably higher at 13,452.22.

Intentional Walks	Predicted Home Attendance Mets	Predicted Home Attendance Marlins	Difference in Attendance
20	21,433.40	15,522.35	5,911.05
60	35,964.14	22,511.92	13,452.22

The table on the next page shows that the typical miss the model makes when predicting average home game attendance is 7,377 people. The adjusted R-Squared value of 0.471 means that the model explains about 47.1% of the variation but fails to explain the other 52.9%. The p-value for the effect test is not statistically significant, having a value of 0.138. This is likely because it brings some redundant information already found in the two other variables. The team name variable, one of the variables used in the interaction, does not bring new information to the model either. Nevertheless, the P-value for the model shows that the variation explained as a whole is unlikely due to chance, having a P-value of 0.000.

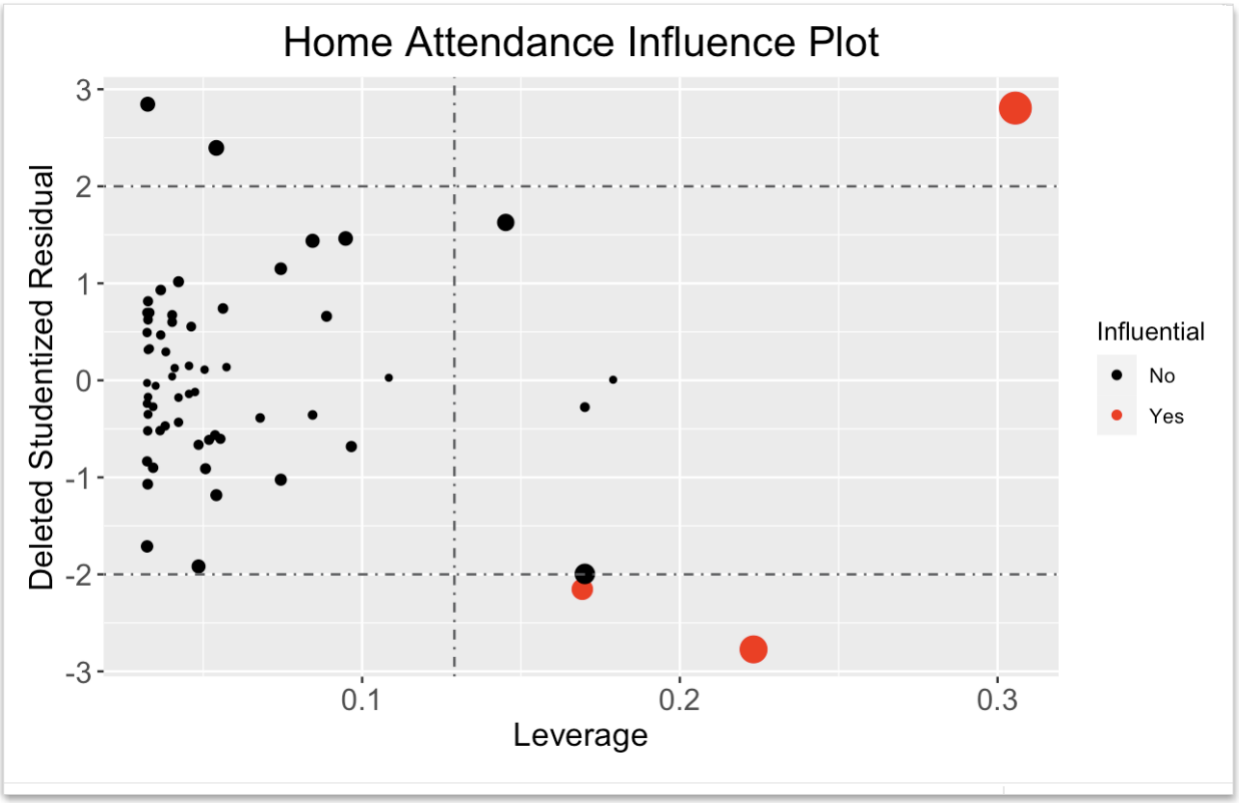
Residual Standard Error	Adjusted R-Squared	Effect Test for Interaction has P-Value	P-Value for Model
7,377.00	0.471	0.138	0.000

The following visualization shows the predicted average home game attendance, shown on the y-axis, for the two teams based on the number of intentional walks displayed on the x-axis. The blue line representing the New York Mets gradually steepens relative to the Miami Marlins from left to right because of the interaction coefficient. There might be some outliers in the data that could throw off the coefficient's predictions, resulting in more uncertainty when making predictions. In addition, the data points appear to have some slightly unequal vertical spread rather than fitting snuggly around the regression line, which could also negatively impact the model.



Examining the influential plot on the next page shows three influential points in the model. The point must have enough leverage to be influential, meaning the x-value is far from the model's mean x-variables. In addition, the y-value must also be far away from the model's estimated y-value with a deleted studentized residual above 2 or below -2.

The first influential point has a deleted studentized residual value of 2.806 on the top right and exerts powerful leverage with a value of 0.306. In addition, two other points on the bottom of the graph have a deleted studentized residual of less than -2 and leverage values of 0.169 and 0.223. A point also slightly overlaps with the influential point with a leverage value of 0.169, but fails to have a low enough deleted studentized residual. This variable might represent the 2020 season for the Miami Mets, where they had 6 intentional walks but an average attendance of 0 people due to the pandemic.



The table below shows the three influential points. The first influential point represents the 2022 season for the Miami Marlins, which had a poor average home game attendance of 10,039 people yet a staggering 69 intentional walks. The second row shows that the Mets had an average attendance of 0 people during the 2020 season, likely due to the pandemic, while still having eight intentional walks for the season. Finally, the last row shows that the New York Mets had an average attendance of 30,843 people for the 2023 season and only one intentional walk, but the Mets will likely have more as the 2023 season progresses.

Team Name	Season	Attendance	Intentional Walks
Miami Marlins	2002	10,039	69
New York Mets	2020	0	8
New York Mets	2023	30,843	1

Next, I will examine the coefficient's variance inflation to gain a better understanding of how much unique information each variable brings to the model. The table shows that the Team Name variable, displayed on the first row, has a variance inflation of 8.46, which is relatively high, and the interaction variable also has a high variance inflation of 10.44. However, the intentional walks variable brings in a considerable amount of unique information with a value of 1.999 (the best variance inflation value would be one). Overall, the variance inflation for the three variables confirms that while the intentional walks variable brings unique information to the model, the team name and interaction variables likely bring redundant information that is already explained with intentional walks.

Coefficient	Variance Inflation
Team Name: New York Mets	8.465
Intentional Walks	1.999
Team Name: Intentional Walks	10.441

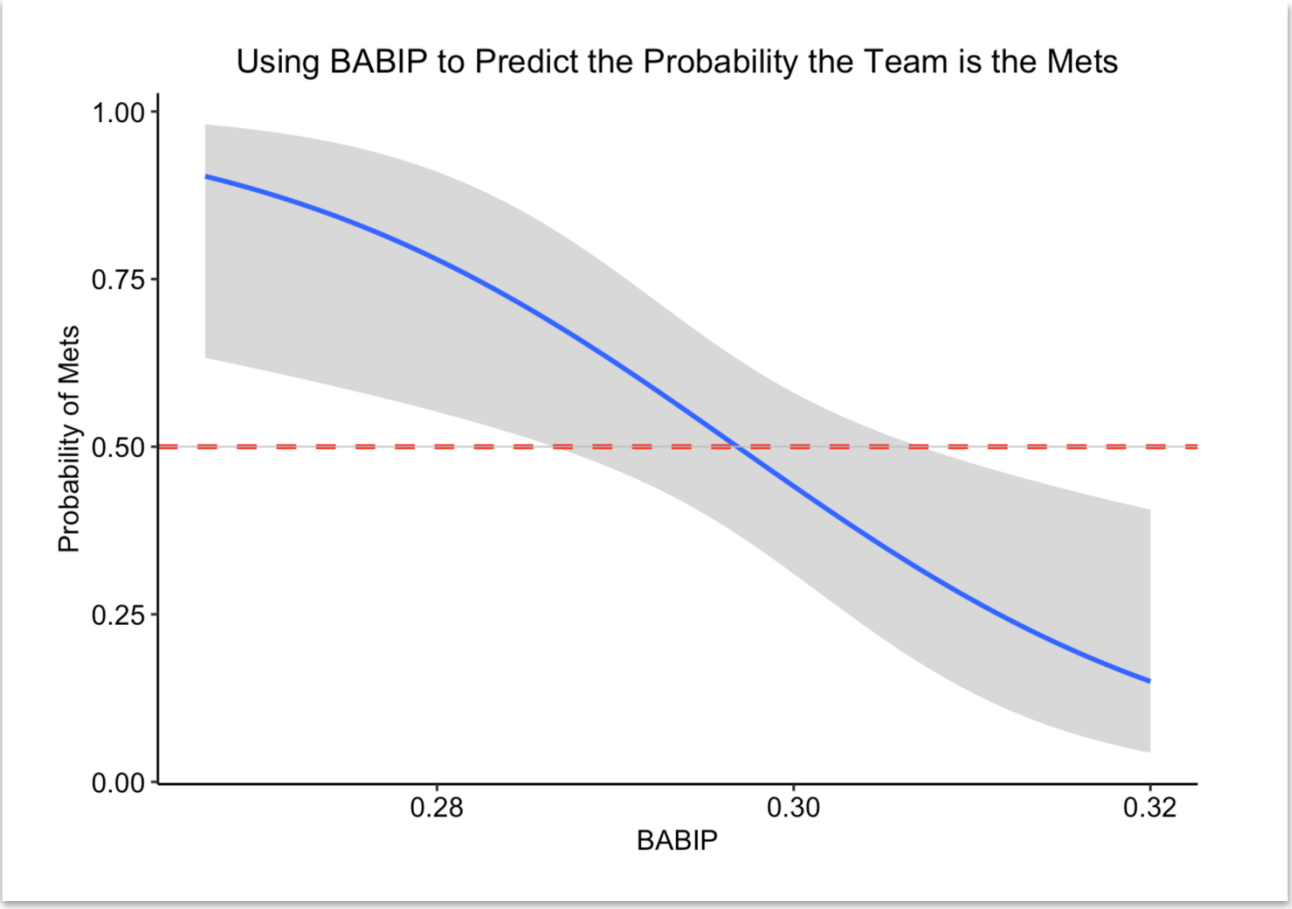
I will next consider a second model that predicts the probability of the team being the New York Mets based on their season performance for a given stat. I will build the model using the all-model approach as done previously to consider the best-predicting variables. However, I will limit the number of predicting variables to one. The table below shows the top three variables that predict whether the team is the New York Mets or Miami Marlins.

AIC	Number of Variables	Terms
-90.9	1	BABIP
-87.2	1	Ground Outs to Airouts
-87.0	1	Catcher's Interference

The model with the lowest AIC to predict the team is BABIP, having an AIC of -90.9. The second and third x-variable models with the lowest AIC are "Ground Outs to Airouts" and "Catcher's Interference," with a value of -87.2 and -87.0, respectively. Therefore, the model will use BABIP as the predicting variable since it has the lowest AIC by more than 2.

The visualization on the following page displays the probability that the team is the Mets, shown on the y-axis, based on BABIP, labeled on the x-axis. The visualization shows that the estimates that fall above the dashed red line are predicted to be the Mets, while the points below the line are estimated to be the Marlins. Since the model's S-curve trends

downward, there appears to be a negative relationship between the probability of being the New York Mets and BABIP.



The following table predicts the team's probability of being the Mets based on a BABIP of 0.26, 0.28, 0.30, and 0.31. Having a 0.26 BABIP for the season results in a 94.05% predicted probability of the team being the Mets and, consequently, a 5.95% chance of being the Marlins. On the other hand, having 0.28 BABIP results in a 77.93% predicted chance of being the Mets and a 22.07% predicted probability of being the Marlins. Finally, a BABIP of 0.30 and 0.31 results in a 44.10% and 27.16% predicted chance of being the Mets, respectively, so a BABIP of 0.30 and 0.31 would result in the model predicting the team to be the Miami Marlins.

BABIP	Predicted Mets
0.26	94.05%
0.28	77.93%
0.30	44.10%
0.31	27.16%

Checking the regression model results in the scientist who wrote the method one and the scientist who wrote the method two test both agreeing that the logistic model can be used. The method one test results in a p-value of 0.592, and the second test results in an even higher, near perfect, p-value of 0.96. Since both p-values are above 0.05, both scientists who wrote the method one and method two tests would say that using the model would be a generally "good" idea in predicting the probability that the team is the Mets based on the season BABIP.

Method One	Method Two
0.592	0.96

To confirm the results from checking the regression model, I will also examine a confusion matrix to see how often the model accurately predicts the team to be the New York Mets. The "Predicted Miami Marlins" column shows that the model accurately predicted the team to be the Miami Marlins 19 out of 29 times, while the "Predicted New York Mets" column indicates that the New York Mets were accurately predicted 21 out of 33 times.

Overall, the model made accurate predictions 64.52% of the time (40 out of 62) and inaccurate predictions 35.48% of the time (22 out of 62). In a naïve model, where one would use the majority level to make the predictions, which in this case would use the team with the highest number of observations, one would only make accurate predictions 50% of the time (31 out of 62). Therefore, our model could improve one's accuracy by 14.52%, which could benefit decision-making.

Terms	Predicted Miami Marlins	Predicted New York Mets	Total
Actual Miami Marlins	19	12	31
Actual New York Mets	10	21	31
Total	29	33	62