# The Shohei Ohtani Effect: Statistical Analysis and Insights

Dan Fuentes (dwf64), Maggie Cui (mmc273)
Olivia Zhang (okz3), Xavier Park (xsp2)

ORIE 3120: Practical Tools for Operations Research, Machine Learning, and Data Science
Dr. Frazier



## *I. Introduction*

Shohei Ohtani is a remarkable professional baseball player who was born and raised in Oshu Iwate, Japan. Since his debut as a two-way player in March 2018, he has become a household name in the world of baseball. Ohtani is currently a designated hitter and pitcher for the Los Angeles Dodgers, showcasing his incredible talent in both positions. In December 2023, the Dodgers acquired Ohtani, from the Los Angeles Angels, for a groundbreaking 10-year, $700 million contract, making him the most valuable player in baseball history and creating the biggest deal in sports history. It is essential to note that the other most substantial contracts in the sports world, in descending order, belong to Lionel Messi with a 4-year $674 million contract,

Cristiano Ronaldo with a 2.5-year $536 million contract, and Patrick Mahomes with a 10-year $450 million contract.

Shohei Ohtani's monumental talent and fame have made him a transcendent figure in Major League Baseball. His performances on the mound and at the plate draw immense attention, altering the landscape for his team. Intrigued by Ohtani's impact, we sought to investigate how his presence influences various factors within the organization. This curiosity led us to pose a key question: Does game attendance depend on whether Ohtani pitches or not?

While this question may seem trivial to the average fan or other MLB franchises, the answer carries significant weight for Dodgers owner Mark Walter. As a businessman, Walter understands that he doesn't merely own a baseball team; he oversees a multifaceted enterprise driven by profitability. Consequently, comprehending Ohtani's drawing power and his financial implications is crucial for informed decision-making and maximizing the organization's profits.

## *II. Analysis Question*

***Is audience attendance influenced by whether or not Ohtani pitches?***

## *III. Describing Dataset*

The final dataset comprises records for each regular season game played by the Los Angeles Angels between 2021 and 2023. It was collected from BaseballReference, which has risen in popularity over the past several years due to its reliable and accessible data. With MLB seasons spanning 162 games each, our dataset contains 486 rows and includes 11 columns, providing comprehensive information on factors we believe will influence stadium attendance. The names of the key columns for this analysis and what they represent are as follows:

- **date**: Specifies the date on which the game was played. Serves as the index, with the most recently played games at the end of the dataset.
- **Gtm**: Represents the game number of that particular season (ex. 46 represents the $46^{th}$ game of that season). Also handles indexing tiebreaks in the event of a doubleheader (two games played on the same day).
- **season**: Indicates the MLB season in which the game occurred.
- **homegame**: A binary indicator (1/0) denoting whether the game was played at the home stadium of the Los Angeles Angels.
- **nightgame**: A binary variable (1/0) indicating whether the game was played during nighttime.
- **Attendance**: Reflects the number of tickets sold for the respective game, providing a measure of stadium attendance.
- **pitched**: A binary flag (1/0) denoting whether Shohei Ohtani was the starting pitcher for the Los Angeles Angels for that game.
- **dayofweek**: Specifies the day of the week on which the game was played.

This comprehensive dataset enables an in-depth analysis of various factors influencing attendance at Los Angeles Angels' regular season games, ranging from temporal aspects such as day of the week and time of day to team performance indicators like championship leverage index and Ohtani's involvement. Given the record-breaking contract just signed by Shohei Ohtani, we will examine whether his new team, the Los Angeles Dodgers, was justified in spending such a large amount of money.

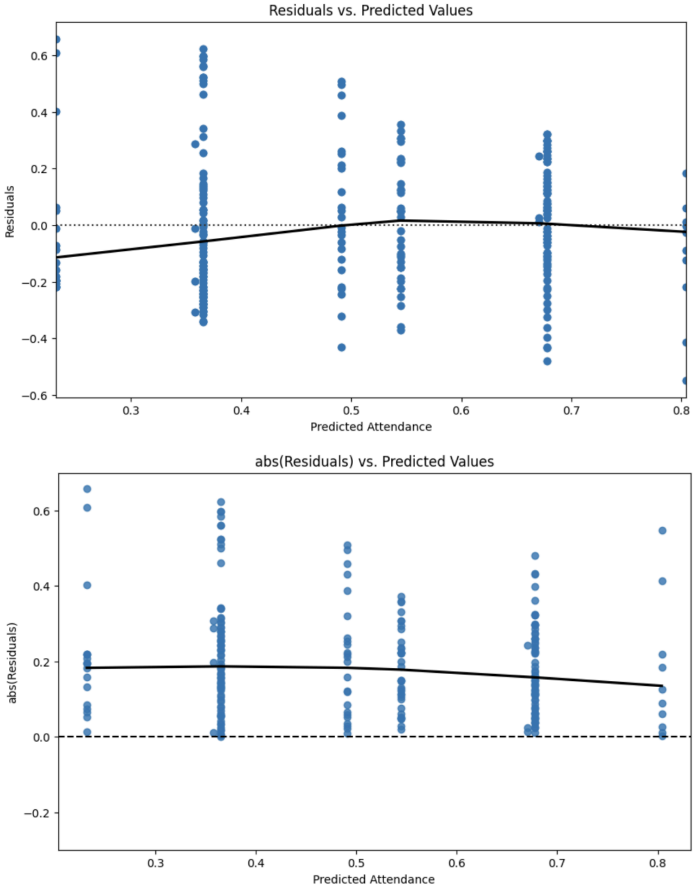
# *IV. Data Analysis*

## ***Testing Assumptions of Linear Regression***

Before continuing with a linear regression model, we first test the assumptions of linear regression to validate our analysis. The following plots test these assumptions:

1. **Linearity in the parameters and constant variance:**

***Figure 1. Residual Plots***



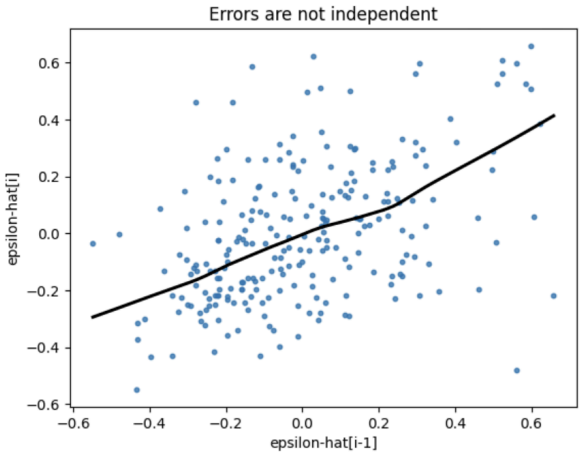

The residuals seem to be evenly scattered around the horizontal zero line across the range of predicted values. There is no clear pattern indicating that the variance of the residuals increases or decreases as the predicted value increases, which is a good sign for linearity. As we have fewer data points, we tolerate more variation in the LOWESS line and even with fewer data points, our LOWESS line is quite flat.

When checking for constant variance in errors, we look at the LOWESS line of the absolute value of the residuals. This line is pretty flat, indicating that there is constant variance in our errors.

As the lines in both plots are relatively flat, this first assumption is met.

## 2. Independence

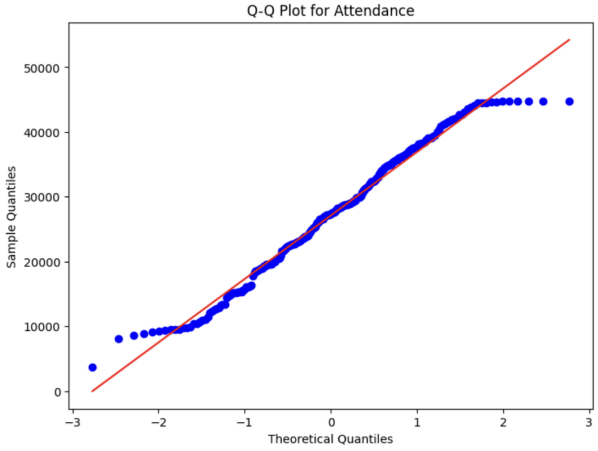
***Figure 2. Correlation***




Based on this correlation plot, it seems that the errors are not independent. However, the residuals are sorted on date due to how the dataset was created. Because of the cyclical nature of pitching rotations and days of the week, there can be certainty on whether Ohtani will pitch the next day based on the current day, and whether the next day will be a weekend. We will continue on the assumption that, outside of these factors, the attendance of one game does not impact the attendance of the next.

## 3. Normal Distribution

***Figure 3. QQ-Plot***




Looking at this QQ-plot, we see that it roughly follows the linear red line and has a slight "light-tailed" distribution. This behavior indicates that compared to the normal distribution there is a little more data located at the extremes of the distribution and less data in the center of the distribution. However, this light-tailed distribution is still strong enough to continue on.

# ***Linear Regression***

In order to examine Shohei Ohtani's effect on stadium attendance, we first fit a multiple linear regression model. To control for variability across different sports markets/fanbases, our data was filtered to only include home games at Angel Stadium. After examining Figure 4, we found that the average stadium attendance was different for every season. We then decided on percentile ranking normalization to standardize attendance for each season to a value within the range of [0, 1]. Our dependent variable therefore represents the percentile of the observed attendance for a game in relation to its corresponding season, and we selected *pitched*, *dayofweek*, and *nightgame* as the regressors. *Dayofweek* was coded such that Friday games were the control category.

***Figure 4. Box Plot***

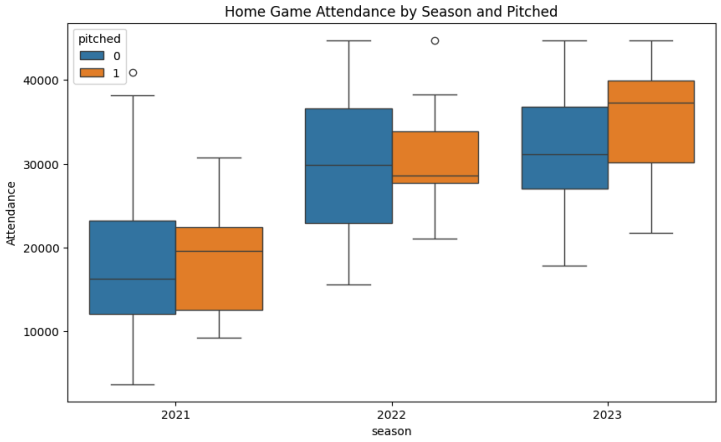



***Table 1. Linear Regression Model 1 Features***

| **Features** | **Estimated Coefficient** (rounded to 2 sig digits) | **Standard Error** | **T-Score** | **P-Value** |
|---|---|---|---|---|
| const | 0.50 | 0.068 | 7.283 | 0.000 |
| pitched | 0.11 | 0.045 | 2.558 | 0.011 |
| nightgame | 0.16 | 0.057 | 2.774 | 0.006 |
| day_Monday | -0.31 | 0.059 | -5.286 | 0.000 |
| day_Tuesday | -0.30 | 0.055 | -5.355 | 0.000 |
| day_Wednesday | -0.26 | 0.060 | -4.319 | 0.000 |
| day_Thursday | -0.25 | 0.068 | -3.623 | 0.000 |
| day_Saturday | 0.052 | 0.054 | 0.966 | 0.335 |
| day_Sunday | 0.052 | 0.078 | 0.659 | 0.510 |

These preliminary results shown in Table 1 are promising. Based on the coefficients and p-values, both Ohtani pitching and the game taking place at night are correlated with higher stadium attendance. Upon examining the dummy variable coefficients for the days of the week, it seems there may be a pattern with attendance based on whether the game took place on a weekday or weekend. Note that the coefficients for *day_Saturday* and *day_Sunday* are shown to not be statistically significant. In context, this means that Saturday and Sunday attendance is not significantly different from Friday attendance. The weekday coefficients are also all negative, statistically significant, and similar in magnitude. For these reasons, we will refit the model using a new variable indicating whether the game took place over the weekend, including Friday in this grouping.

***Table 2. Linear Regression Model 2 Features***

| **Features** | **Estimated Coefficient** (rounded to 2 sig digits) | **Standard Error** | **T-Score** | **P-Value** |
|---|---|---|---|---|
| const | 0.23 | 0.038 | 6.095 | 0.000 |
| pitched | 0.13 | 0.042 | 2.986 | 0.003 |
| nightgame | 0.13 | 0.037 | 3.632 | 0.000 |
| is_weekend | 0.31 | 0.032 | 9.803 | 0.000 |

This new model, using the indicator *is_weekend*, yields several improvements. Our F-statistic increased from 13.08 to 34.58. While both of these scores are statistically significant, the joint significance of the variables in our second model seems to be more extreme. AIC decreased from 14.23 to 6.424, supporting that the new model might be more robust against overfitting. This is also supported by our $R^2$ values; while multiple-$R^2$ decreased from 0.308 to 0.302, our adjusted-$R^2$ increased from 0.285 to 0.293, which corroborates the change we saw in AIC. In context, this means that roughly 30% of the variation in home game attendance is explained by

the variation in our predictors. Although this value may seem low, an $R^2$ value in this range is not uncommon in social sciences and other fields where randomness of human behavior is involved.

All of the coefficients in the new model are also statistically significant at a very high confidence level. Interpretations for the coefficients in Table 2 are as follows:

1) For daytime weekday games where Ohtani is not the starting pitcher, we expect stadium attendance to be around the 23rd percentile of that season.
2) Holding all other variables constant, we expect to see normalized attendance 13 points higher for games where Shohei Ohtani pitched, in comparison to those where he did not. Ohtani pitching is therefore positively correlated with home game attendance, but this model alone is not enough to conclude causality. We will examine this later.
3) Holding all other variables constant, we expect to see normalized attendance 13 points higher for games that take place at night, in comparison to day games. This positive correlation could be attributed to fans being busier during the day with a variety of responsibilities like work or familial obligations, making night games more accessible.
4) Holding all other variables constant, we expect to see normalized attendance 31 points higher for weekend games than weekday games. Similar to night games, we could reasonably hypothesize that fans may have more difficulty in attending weekday games as a result of work, family obligations, and other factors that would affect their availability throughout the week. Culturally, weekends are widely associated with rest and leisure, so this result is not unexpected.

While we must note the caveat that correlation does not imply causation, our model does predict higher attendance for weekend games that take place at night, with the distinguished Ohtani as starting pitcher. In order to maximize ticket sales, it may be beneficial for Ohtani's team owner/general manager to try and schedule his pitching starts for games that are at night and on the weekend.

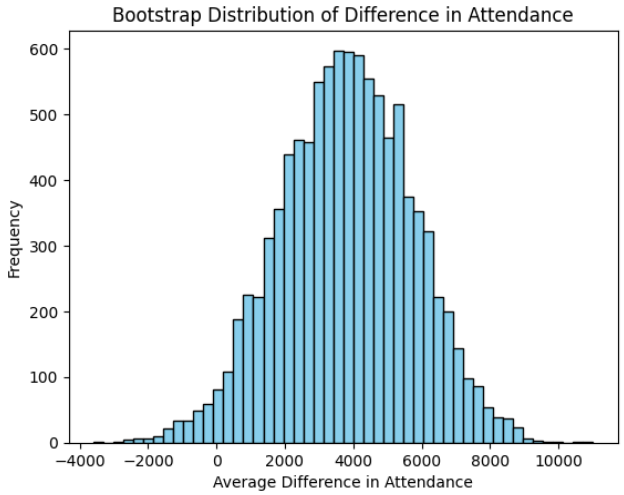
## ***A/B Testing with Bootstrapping***

To further examine the causality of Ohtani pitching on game attendance, we employed bootstrapping methods to assess and quantify statistical significance and estimate confidence intervals. We sampled with replacement to generate distributions of the differences in attendance between games where Ohtani pitched and those where he did not. This bootstrapping method enables us to estimate statistical metrics such as p-values and confidence intervals more precisely, mitigating the impacts of working with our smaller-sized dataset.

Given that Shohei Ohtani was just traded to the Los Angeles Dodgers in 2024, we will perform bootstrapping on only the 2023 season to ascertain Ohtani's most recent causal effect on attendance. We believe the recency of this subset will ultimately be the most relevant to his effect moving forward as past years may have underrepresented his current potential while he was still rising to stardom. Although we cannot randomly assign Ohtani to pitch certain games, MLB

structure and tendencies of baseball coaches and decision-makers work to our benefit. In Major League Baseball, it is standard for teams to use a rotation of starting pitchers. Starting pitchers typically play in every four to five games in order to ensure recovery between games. MLB schedules so that teams play the same opponent in a series that typically lasts around three to four games. As a result, each pitcher on a given team plays under reasonably similar conditions to one another throughout a given season: the amount of home and away games remain fairly consistent, the days of the week they pitch changes due to unevenness in the length and size of a weekly rotation, and each pitcher gets to play against the same opposing teams. We therefore justify the use of A/B methodology because our sample of when Ohtani pitched in 2023 resembles that of a systematic random sample.

***Figure 5. Bootstrap Distribution Plot***




The results from the bootstrap analysis demonstrated a statistically significant influence, with a p-value of 0.024. This indicates that there is less than a 2.4% chance that the observed difference in attendance between games where Ohtani pitched and those where he did not could be due to random variation alone.

The average difference in attendance when Ohtani pitched was notably higher, with an observed difference of 3785.785. This suggests a strong positive effect of his appearances on game attendance. The 95% confidence interval for the difference in attendance ranged from 24.73 to 7499.62, affirming the robustness of these findings. This interval does not include zero, which further supports the conclusion that Ohtani's pitching had a positive impact on attendance numbers.

## *V. Discussion and Conclusion*

Our statistical analysis focused on assessing the impact of Shohei Ohtani pitching on game attendance for the Los Angeles Angels. We followed a structured approach starting with preparing data, breaking down and comparing variables across days of the week, and whether or

not the game was at night. We tested assumptions necessary for linear regression, allowing us to move forward with linear regression and bootstrapping methodologies to gauge Ohtani's influence on audience attendance.

1. Linear Regression Insights:
    - The regression analysis indicated a statistically significant positive correlation between Ohtani pitching and increased game attendance. Games where Ohtani pitched saw an increase in attendance percentile rank, confirming the pulling power of Ohtani as a major draw for fans.
    - Additional factors such as whether the game occurred at night or over the weekend also showed significant effects on attendance, highlighting times when games are likely more accessible to larger audiences.
2. Bootstrapping Results:
    - The bootstrap analysis, reinforced by A/B testing, provided compelling statistical evidence that attendance significantly increased when Ohtani pitched, with a noteworthy p-value of 0.024. This robustly suggests a causal relationship, indicating that Ohtani's presence as a pitcher directly influences spectator numbers, surpassing typical attendance fluctuations.
    - Importantly, these findings complement our insights from linear regression, which initially highlighted a positive correlation between Ohtani's pitching performances and attendance patterns. Together, these analyses not only affirm the correlation but also establish a causal link, solidifying the impact of Ohtani's pitching on drawing larger crowds to Angels games.

Limitations and Further Research:

- While our analysis confirms the correlation between Ohtani pitching and increased attendance, it remains cautious of claiming outright causation due to limitations in observational studies. Future studies might explore a more granular dataset or employ advanced statistical techniques with a larger dataset to solidify these findings.
- Further research could also evaluate the long-term impact of Ohtani on the Dodgers' brand value and other business metrics, potentially integrating fan engagement and merchandise sales data.

Ultimately, Shohei Ohtani’s influence on game attendance is both statistically significant and financially meaningful. The results from our analysis support the decision of Dodgers management to acquire Ohtani as his star status has had a noticeable effect on fanbase attendance during his time with the Angels. We suggest that the Dodgers management team optimizes game scheduling and marketing strategies around Ohtani's outings to increase profits. Prioritizing his starts in accordance with our findings may maximize ticket sales for home games and, subsequently, revenue.