Analyzing how the Houston Astros benefitted from cheating

Daniel Opdahl, Ryan Manternach

Introduction

In November of 2019, an article was published in The Atlantic describing a scheme that the Houston Astros of the MLB implemented for the 2017, 2018, and 2019 seasons that allowed the team to steal the opposing team's signs and relay them to the Astros' hitters.

The MLB records 144 different offensive metrics for each player over the course of a season. Examples of these metrics are at-bats, runs, RBIs, steals, zone swing rate, barrel percentage, line-drives, etc. More traditional statistics combined with advanced metrics allow baseball nerds to quantify a players offensive prowess.

In our project, we will be looking at Houston Astros players pre 2017, and post 2017 and comparing them with a baseline comprised of the rest of the league to see if the Astros' hitters demonstrably increased their offensive production as a result of their cheating. Comparing the Astros' players against their pre 2017 statistics as well as comparing their post 2017 statistics against the rest of the league will allow us to account for factors like player development (as the Astros' players are typically young, an increase in offensive production could be a result of simple skills and athletic improvement), and league-wide trends (for example, a "juiced ball" will benefit not only the Astros' hitters, but hitters all over the league).

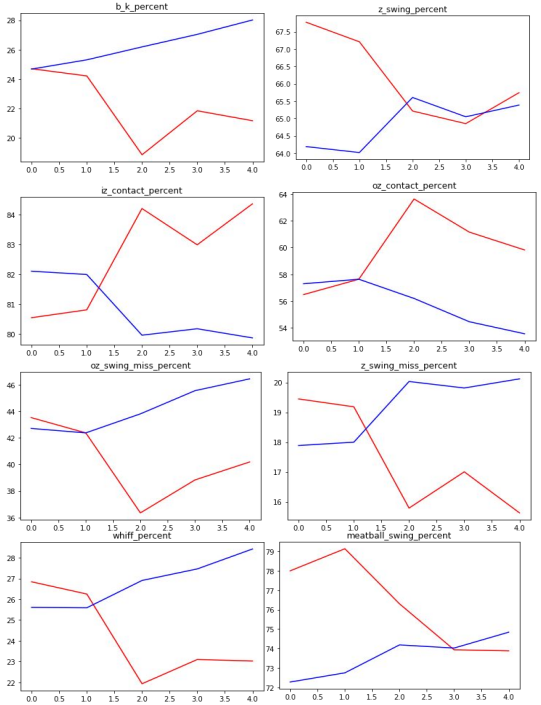
Seeing what metics, if any, show a statistically significant "jump" during the 2017 season for Astros' hitters when compared to the rest of the league and previous player trends will give us insight into how players benefitted from the cheating, and to what degree.

Selecting Data

We took our data from MLB Savant, the official source of statistics for the MLB. To produce a more holistic picture of how the Astros may have benefitted from cheating, we wanted to use as much available data as possible, but for practical and computational reasons, it was necessary to drop some of the stats that intuition says would be unimpacted by prior knowledge of what pitch was coming.

After cleaning and filtering the data, we generated charts that compare the MLB's average stats and the Astros' hitters stats over the years 2015-2019 visually. The x-axis is years, but the scale is incorrect. The y-axis is the respective stat being visualized.

Astro Hitters' Performance Compared With Average MLB Performance



From the charts, it appears the Astros' benefit from cheating was reflected most in the following stats: K-percentage, zone swing percentage, zone swing and miss percentage, out of zone swing and miss percentage, out of zone contact percentage, meatball swing percentage, in zone contact percentage, whiff percentage.

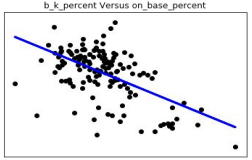
Individual Analysis

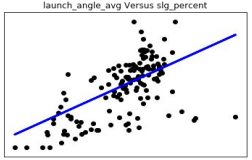
In order to determine how individual players benefitted, and not just how the team as a whole benefitted, we selected players that had been a part of the Astros pre-cheating and post-cheating. This would allow us to establish a baseline of at least one year so that we can measure their post-cheating stats against the baseline and come up with figures of benefit. We decided not to include players who had joined the Astros after 2017 who were already in the league. An argument can be made for finding those players' baselines by using their stats while they were playing for their previous team, but controlling for variables is an issue then. Maybe that player benefited more from simply changing teams or leagues and that is why they saw an increase in offensive production, and not because of cheating.

What we largely found was that some players benefited greatly, and very few players were harmed by the cheating.

Linear Regression - Was The Cheating Worth It?

We used a linear regression model to associate changes in some stats with changes in other stats. By using a linear regression model and the MLB's league-wide averages, we can find relationships between selected stats, and then determine how much the benefit the Astros hitters saw in certain stats should have translated to other stats. For example, how much should Alex Bregman's 10% decrease in K-percentage be worth in terms of on base percentage?





Although the coefficients of determination do not seem very high for these fits, they are better fits than the R^2 value would lead you to believe. The variability of certain stasts is very high in the MLB. This leads to many data points that fall far from the line of regression, but as long as that line of regression follows the prevalent trend of the data, I believe that it can be trusted for the most part.