Bang the Can Slowly: An Investigation into the 2017 Houston Astros

October 2, 2020

Abstract

This manuscript is a statistical investigation into the 2017 Major League Baseball scandal involving the Houston Astros, the World Series championship winner that same year. The Astros were alleged to have stolen their opponents' pitching signs in order to provide their batters with a potentially unfair advantage. This work finds compelling evidence that the Astros on-field performance was significantly affected by their sign-stealing ploy and quantifies the effects. The three main findings in the manuscript are: 1) the Astros' odds of swinging at a pitch were reduced by approximately 27% (OR: 0.725, 95% CI: (0.618, 0.850)) when the sign was stolen, 2) when an Astros player swung, the odds of making contact with the ball increased roughly 80% (OR: 1.805, 95% CI: (1.342, 2.675)) on non-fastball pitches, and 3) when the Astros made contact with a ball on a pitch in which the sign was known, the ball's exit velocity (launch speed) increased on average by 2.386 (95% CI: (0.334, 4.451)) miles per hour.

Keywords: Baseball, sports statistics, generalized linear mixed model

1 Introduction

During the 2017 Major League Baseball (MLB) season, the Houston Astros are alleged to have implemented a elaborate sign-stealing scheme. In 2019, Mike Fiers, a pitcher for the Houston Astros during their 2017 World Series championship run, claimed that his former team was stealing signs by using a camera in center field, Walstein (2019). The information in these stolen signs was relayed to players by banging a baseball bat against a trash can (Rosenthal & Drellich 2019), referred to here as a "bang". In this particular scheme, a bang indicated to the batter that the upcoming pitch would be an off-speed pitch such as a curveball or a slider. The absence of a bang is inconclusive; it could indicate a fastball or that they simply could not decode the sign.

Ultimately, MLB punished the Houston Astros by suspending their manager, A.J. Hinch, and general manager, Jeff Lunhow, for one year. Additionally, the Astros were fined $5 million and their first and second round draft picks were taken away for the 2020 and 2021 amateur drafts. This was a substantial penalty, and meant to dissuade future teams from impacting their games in a similar manner.

However, not everyone agrees on the effects of stealing pitching signs during an MLB game. In one particularly bizarre exchange during a press conference on February 13, 2020, the owner of the Astros, Jim Crane, was quoted as saying "**Our opinion is this didn't impact the game**. We had a good team. We won the World Series and we'll leave it at that." In that same press conference, less than a minute later, he is also quoted as saying, "I didn't say it didn't impact the game." See Axisa (2019) for more information on this press conference.

Others have found many striking examples of different aspects of the game that appear to show that that Houston benefited from sign stealing such as Sawchik (2019), Stark & Sarris (2020), and Arthur (2019) whereas Lindbergh (2019) found litte evidence the Astros gained much.

While there has been quite a bit of analysis on the on-field effects of the Astros' sign stealing, it is still largely an open question as to whether or not there were on-field improvements because of the sign stealing. In addition, if stealing signs did lead to improvement, what types of improvements were observed and can the magnitude of these improvements be quantified. In this paper, we attempt to erase any ambiguity related to the efficacy of the Astros' sign-stealing scheme during the 2017 season. In other words, we address whether or not their scheme affected on-field performance during the 2017 season and quantify their impacts where appropriate.

2 Data

In the analysis that follows, we rely on three data sources: Statcast, Pitch Info, and Bangs. First, Statcast is Major League Baseball's ball and player tracking system that has been in every MLB park since the 2015 season. A new version of Statcast, Version 2.0, is set to be released in 2020. The Statcast V1.0 system has two data collection components: (1) Trackman Doppler Radar that tracks baseball events and (2) Chryon Hego Cameras that track player movements. In the first three seasons alone, 2.1 MM pitches and 400K balls in play were tracked. Examples of variables that are available in Statcast include a hit's launch speed (exit velocity), pitch classification, pitch spin rate, among a host of additional measurements. See Major League Baseball (2020) for additional information related to MLB's Statcast application programming interface (API).

Although pitch classification is available in Statcast, we relied on Pitch Info (Pavlidis & Brooks 2020) data for classifications rather than Statcast. Pitch Info is regarded as the most accurate classification system in terms of pitch group classification in the industry. In addition, we utilize Pitch Info's derived variable, referred to as called strike probability (CSP) as a covariate in our analysis. CSP is an estimate of the probability that a pitch will be called a strike. Full details related to the CSP model can be found here (Judge & Brooks 2015).

Finally, we merge the previous two data sets with the so-called Bangs data. These data contain information on whether or not a measurable, auditory signal was present prior to pitches on a selection of Astros' at-bats during a subset of their Astros' 2017 season's home games. The signal, if present, was the result of banging a hard object on a metal trash can. The data were compiled by Tony Adams, a self-described Astros fan, and are publicly available on his website (Adams 2019). The data are essentially a combination of data obtained from Major League Baseball's Statcast API along with video of Houston Astros' games from Youtube. Adams matched timestamps from the MLB Statcast data to the game video, and produced a spectrogram to represent the audio before and after all pitches in his study. The spectrogram of the auditory footprint of each of these pitches was used to identify when bangs were present prior to a pitch.

Next, we looked at the prevalence of bangs for different types of pitch groups, as defined in Pitch Info, see Table 1. A χ^2-test of independence yields a p-value of $< 2.2 \times 10^{-16}$, strongly rejecting the null hypothesis of independence between the two variables. Offspeed pitches show a bang prior to the pitch at rates of 23.7%, 27.6%, and 26.8%, respectively. On the other hand, fastballs (FA) only had a bang prior to the pitch on 2.3% of the pitches. So while the Astros' method was not perfect, it is obvious that information such as an upcoming offspeed pitch was being transmitted to the batter via the trash can banging system.

Table 1: The number of pitches of Pitch Info pitch group and the incidence of bangs. The percentages correspond to the bang prevalence conditioned on pitch type category.

		Pitch Type			
		Change-up	Curveball	Fastball	Slider
Bangs	No	756 (76.3%)	707 (72.4%)	4128 (97.7%)	1470 (73.2%)
	Yes	235 (23.7%)	270 (27.6%)	97 (2.3%)	538 (26.8%)

Finally, Table 2 shows the relationship between swinging at a pitch and whether or not there was a bang preceding the pitch. Given the presence of a bang, an Astros' player swung 40.5% of the time as opposed to 46.2% when there was no bang. This translates to an odds ratio of 0.793 (95% CI: 0.69678, 0.9023; p-value 3.706×10^{-4}) , which indicates an approximate 21% reduction in the odds of swinging given a bang relative to when there was no banging before the pitch.

Table 2: The number of pitches with bangs by the number of swings. Percentages correspond to the percentage of swings given bangs/no bangs on the pitch.

		Swing	
		No	Yes
Bangs	No	3798 (53.8%)	3263 (46.2%)
	Yes	678 (59.5%)	462 (40.5%)

3 Methodology

In their most general form, we are simply fitting generalized linear mixed models (GLMM) in each case (McCulloch & Neuhaus 2014). The particular GLMM is determined by its response variable and covariates, however, they can be defined in general using Equation (1) by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}. \tag{1}$$

Here $\boldsymbol{\eta}$ is referred to as the linear predictor and is related to the response of interest through a link function $g(\cdot)$, $\mathbf{X}$ is a matrix of covariates, $\boldsymbol{\beta}$ is a vector of parameters (fixed-effects),

$\mathbf{Z}$ is the random effect design matrix, and $\mathbf{b}$ is the vector of random effect parameters. The specific GLMMs that we utilize are defined by the response variables of interest, the link $g(\cdot)$, covariates, and random effect terms. The specific terms are given explicitly in the following subsections.

Swing Model. We developed our first GLMM in order to estimate the effect of stealing a pitching sign on swinging at the subsequent pitch. Specifically, let Y_{ij} be an indicator for the j^{th} player swinging at the i^{th} pitch, $i = 1, \cdots, n_j$ and $j = 1, \cdots, N_{players}$. The link function in this situation is the well known $\text{logit}(\pi_{ij}) = \log\frac{\pi_{ij}}{1-\pi_{ij}}$, where $\pi_{ij} = P(Y_{ij} = 1)$ conditioned on an indicator variable for the presence or absence of a bang, while controlling for pitch type (*i.e.*, fastball or not), CSP, pitch count (as a factor). By pitch count, we mean the number of balls and strikes that a batter has while facing the current pitch. Additionally a random slope term for each batter is included in the model.

Contact Model. Given that a swing occurred, we now focus our attention on whether or not contact was made with the ball. In other words, did the batter actually hit the ball (not necessarily hit in play) that was pitched when he swung? In this case, we define another binary response variable to be $Y_{ij} = 1$ if the j^{th} player made contact with the ball on the i^{th} pitch, $i = 1, \cdots, n_j$ and $j = 1, \cdots, N_{players}$. Similar to the situation given in the previous subsection, the link function in this model is the $\text{logit}(\pi_{ij})$ where π_{ij} is the probability of making contact given fixed-effect covariates defined by CSP, an indicator variable for a fastball, an indicator variable for the presence of a bang, and an interaction term between fastball and bang. We include random intercepts for *both* pitcher and batter as well as a random slope for bangs for battter.

Exit Velocity Model. The hierarchical nature of our modeling process leads us to examining a hit's exit velocity (EV) based on the presence of bangs prior to the pitch. Given that the batter swung at the pitch *and* made contact with the ball, was exit velocity impacted by stealing the pitching sign? In other words, are hits better given knowledge of an upcoming pitch?

In this case, we consider Y_{ij}, the exit velocity of the ball leaving the bat, as a continuous variable. Therefore, the link function is simply the identity and, hence, we use a standard linear mixed-effects model. Covariates in this model include CSP, an indicator variable for fastball, an indicator variable for the presence of a bang, and random intercepts for both pitcher and batter.

4 Results

Swing Model. The full results of fitting the model described in Section 3.1 are not presented here; rather only the coefficients of primary interest (i.e. effect of the bangs) are presented here. The coefficient estimate for the bangs indicator variable is -0.3219 (95%

CI: -0.482, -0.163, p-value: 7.59×10^{-5}) indicating that when there were bangs prior to a pitch, the batter was significantly less likely to swing at that pitch relative to pitches with no bangs present. When all other covariates are held constant, the odds ratio for the probability of swinging comparing bangs to no bangs is 0.725 (95% CI: 0.618, 0.850). This translates to an approximate 27.5% reduction in the odds of swinging in the presence of a bang. This is conclusive and statistically significant evidence that on-field behavior was directly affected by stealing the pitcher's sign. That is, the act of banging on a drum prior to a pitch (to indicate the ensuing pitch type) provided significant information to the batter causing him to swing less often.

It is worth discussing why a player might swing more often given that he knows a fastball is being pitched. Simply put, fastballs are easier to hit. Verducci (2020) states that the MLB batting average is approximately 20% - 40% higher on fastballs and that the off-speed pitches lead to less contact.

Contact Model. Prior to fitting this model, pitches where the player did not swing were removed from the data set, along with several unrelated and rare results, e.g. batter or catcher interference. This leaves 3725 observations. Given that a swing occurs, we defined "contact" on a pitch as any result other than a swing and miss. Therefore, a ball put in play, regardless of whether or not they made an out, a foul ball, or a home run are all treated equally as "contacts".

Table 3: The fixed effect estimates for the contact model with the effect of banging on a metal can in bold.

Term	Estimate	Std. Error	Z Statistic	p-value
Intercept	-0.20	0.14	-1.46	0.15
CSP	1.90	0.12	16.37	0.00
$I_{\{\text{Fastball}\}}$	0.97	0.10	9.61	0.00
$I_{\{Bang\}}$	**0.59**	0.22	2.68	0.01
$I_{\{\text{Fastball}\}} * I_{\{Bang\}}$	**-1.19**	0.45	-2.64	0.01

A summary of the results of fitting the model described in Section 3.2 are given in Table 3. Notice that both the indicator variable for banging on the can and its interaction effect with the fastball indicator are both significant. For this reason, we interpret fastballs and off-speed pitches separately. First, the estimated effect size for bangs is 0.591 (95% CI: 0.294, 0.984) on off-speed pitches. This corresponds to an odds ratio of 1.805 (Bootstrapped 95% CI: 1.342, 2.675, see Efron & Tibshirani (1986)). In other words, given that a player swings at the pitch, the odds of making contact in the presence of a bang (i.e. the pitching sign is

known) are about *80% higher* than the odds of making contact when a bang is not present.

Next, we will consider fastballs. The coefficient for bangs when the pitch is a fastball is -0.603. This corresponds to an odds ratio of 0.547 (Bootstrapped 95% CI: 0.227, 1.774). While this odds ratio is not significantly different than one, an odds ratio less than one here would mean that a player is *less likely* to make contact in the presence of a bang.

Not all player on the Astros benefited uniformly from the information provided by a bang. In fact, of the twenty players included in our data set, ten of them do not exhibit a statistically significant increase in the odds of making contact on off-speed pitches given a swing (i.e., their confidence interval contains 1). The ten remaining players George Springer, Yulieski Gurriel, Jacob Marisnick, Evan Gattis, Josh Reddick, Max Stassi, Carlos Correa, Carlos Beltran, Norichiak Aoki, and Anthony Kemp, on the other hand, all exhibited significant increases in their respective odds of contact given the presence of bangs prior to the pitch. Nine of these ten remaining players had increases in their odds of contact on off-speed pitches ranging from 68% (Anthony Kemp) to 159% (Yulieski Gurriel). However, one player, George Springer, seems to have benefited much more than the other players with an estimated 281% increase in the odds of contact on an off-speed pitch when a pitch was preceded by a bang (OR 3.810 (95% CI: 2.042, 12.864)).

To put this in perspective, in our data set we have 390 records of George Springer swinging and 135 of those swings were at off-speed pitches. Ninety-five of these swings at off-speed pitches were not preceded by bangs and 31 of these swings resulted in no contact for a swing-and-miss rate of 32.63% (95% CI: 23.57%, 43.12%). Out of the remaining 40 swings at off-speed pitches that were preceded by bangs *only 2 resulted in no contact.* This corresponds to a miss rate of only 5% (95% CI: 0.87%, 18.21%).

Exit Velocity Model. To fit the exit velocity model, we restricted the observations used in this model to only include instances where contact (as defined above) was made and a launch speed, or exit velocity, was recorded. This leaves 2272 observations for our final analysis.

Coefficient estimates for the exit velocity model are given in Table 4. The coefficient estimate for indicator of banging on a can in this model is 2.386 (95% CI: 0.334, 4.451). Therefore, we estimate that when a batter makes contact with ball on a pitch preceded by bangs their exit velocity is 2.386 miles per hour greater on average than on pitches that were not preceded by a bang, when every other variable is held constant. To put this in perspective, a ball hit at 100 miles per hour at a launch angle of 30 degrees will travel roughly 385.3 feet before it hits the ground (Nathan 2020). A ball hit with the same launch angle, but with an exit velocity of 102.38 (i.e., 2.386 miles per hour more) will travel 397.9 feet before hitting the ground, or 12.6 additional feet. This is the difference between a long fly ball to straight away center field (likely an out) and a home run at Fenway Park, where a home run is 389' 9" inches to center field.

Table 4: The fixed-effects estimates for our model on exit velocity. The effects involving bangs in shown in bold font.

Term	Estimate	Std. Error	Z Statistic	p-value
Intercept	76.00	0.94	80.75	0.00
CSP	8.36	0.91	9.18	0.00
$I_{\{\text{Fastball}\}}$	2.20	0.70	3.15	0.00165
$I_{\{Bang\}}$	**2.39**	1.05	2.27	0.0230

5 Conclusion and Discussion

In this manuscript, we examined the effects of sign stealing by the Houston Astros during the 2017 Major League Baseball season. We first verified that the presence of banging on a trashcan prior to a pitch was indeed related to the type of pitch being thrown. That is, do the bangs indicate that an off-speed pitch is likely to be thrown. The results are presented in Table 1. Next, we showed that the presence of the banging was significantly related to the probability of an Astros batter swinging, however, we did not control for potential confounding factors.

In order to control for additional covariates, we used a series of generalized linear mixed effects models to control for known factors that potentially affect each of the outcomes considered here. Specifically, we modeled the probability that a player swings at a pitch, followed by modeling the probability of contact given a swing, and finally a model looking at exit velocity given contact. The three main findings of our paper are that the presence of bangs made it *less* likely that a player would swing at a pitch, *more* likely that a player would make contact with a off-speed pitch given that he swung, and *increased* the average exit velocity given that a player swung and made contact.

In addition, we found that there was quite a bit of variability in how much the banging aided players in making contact with the ball given a swing. A particularly notable example is that George Springer was found to make contact on swings of off-speed pitches at much higher rates when a bang was present relative to the same pitch with no bang, estimated OR 3.810 (95% CI: 2.042, 12.864).

In closing, we emphasize that these data and the results of our modeling efforts show that the effect of the Astros stealing the pitching sign significantly impacted their team's on-field performance. And while the effects were shown to differ from player to player, the overall impact on the game itself is undeniable – the Astros were beneficiaries of their sign-stealing

scheme and went on to win the 2017 World Series. Given the evidence presented here, we would argue that a cheater may indeed prosper, and occasionally even win a World Series.

Acknowledgements

The authors gratefully acknowledge Harry Pavlidis and all the members of the Baseball Prospectus Stats Slack channel, Tony Adams for collecting and disseminating the bangs data set, Tim P. Levine for the title suggestion, and Scott Sibbel for early comments and suggestions.

References

Adams, T. (2019).
URL: *http://signstealingscandal.com/*

Arthur, R. (2019), 'Moonshot: The Astros' offense took a huge leap after they started stealing signs'.
URL: *https://www.baseballprospectus.com/news/article/55450/the-astros-offense-took-a-huge-leap-after-they-started-stealing-signs/*

Axisa, M. (2019), 'Astros owner Jim Crane says sign-stealing scandal 'didn't impact the game' as team issues public apology'.
URL: *https://www.cbssports.com/mlb/news/astros-owner-jim-crane-says-sign -stealing-scandal-didnt-impact-the-game-as-team-issues-public-apology/*

Efron, B. & Tibshirani, R. (1986), 'Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy', *Statistical Science* **1**(1), 54–75.

Judge, J., P. H. & Brooks, D. (2015), 'Moving beyond WOWY: A mixed approach to measuring catcher framing'.
URL: *https://www.baseballprospectus.com/news/article/25514/moving-beyond-wowy-a-mixed-approach-to-measuring-catcher-framing/*

Lindbergh, B. (2019), 'There's no virtue in signaling. but is there any benefit?'.
URL: *https://www.theringer.com/mlb/2019/11/22/20977542/astros-sign-stealing-benefit-wins-advantage*

Major League Baseball (2020), 'About statcast'.
URL: *http://m.mlb.com/glossary/statcast*

McCulloch, C. E. & Neuhaus, J. M. (2014), 'Generalized linear mixed models', *Wiley StatsRef: Statistics Reference Online* .

Nathan, A. (2020), 'The phsyics of baseball'.
URL: *http://baseball.physics.illinois.edu/trajectory-calculator-new.html*

Pavlidis, H. & Brooks, D. (2020).
URL: *www.pitchinfo.com*

Rosenthal, K. & Drellich, E. (2019), 'The Astros stole signs electronically in 2017 — part of a much broader issue for Major League Baseball'.
URL: *https://theathletic.com/1363451/2019/11/12/the-astros-stole-signs-electronically-in-2017-part-of-a-much-broader-issue-for-major-league-baseball/?source=spotracpc=spotrac40off*

Sawchik, T. (2019), 'If the Astros stole signs, how much did it help them?'.
URL: *https://fivethirtyeight.com/features/if-the-astros-stole-signs-how-much-did-it-help-them/*

Stark, J. & Sarris, E. (2020), 'Does electronic sign stealing work? the Astros' numbers are eye-popping'.
URL: *https://theathletic.com/1573075/2020/01/31/does-electronic-sign-stealing-work-the-astros-numbers-are-eye-popping/*

Verducci, T. (2020), 'The fastball is disappearing. What does it mean for MLB's future?'.
URL: *https://www.si.com/mlb/2020/08/10/justin-verlander-fastball-usage*

Walstein, D. (2019), 'Former Astros pitcher says team electronically stole signs in 2017'.
URL: *https://www.nytimes.com/2019/11/12/sports/baseball/astros-cheating.html*