Devan Venkataraman & Nathaniel Bozzella
EC 107 Empirical Project
Sergio Turner
12/20/20

Stolen Signs to Stolen Wins?

The Trash Can Banging Scandal Heard 'Round the World

Question

To what extent, and in what ways, was the Houston Astros cheating scandal in the 2017 season effective in improving team performance?

Introduction

For the majority of the 2010's, the Houston Astros were a very middle of the pack team. From 2010-2014, the team did not finish higher than 4th in their division. For most of their history, the Houston Astros participated in the National League Central Division, up until the 2013 season. Since the 2013 season, the Astros have competed in the American League West Division, where they have seen much more success.

In 2011, the Astros, one of the worst teams in baseball with a record of 56-106, were sold to Jim Crane where he moved on from ex-GM Ed Wade, and hired Jeff Luhnow two days after the sale. While Ed Wade made some good decisions: debuting Jose Altuve in the 2011 season and drafting George Springer in the 2011 draft, his overall performance was not satisfactory for the new owner. The new GM, Jeff Luhnow, made some notable decisions as well, drafting Carlos Correa in the 2012 draft (debuting him in 2015) and drafting Alex Bregman in the 2015 draft (debuting him in the 2017 season).

After another few unsuccessful seasons with records of 55-107, 51-111, and 70-92 in the 2012-2014 seasons, Jeff Luhnow decided to fire the current manager of the team, whom he had a falling out with towards the end of the 2014 season. Jeff then hired AJ Hinch to be the Astros manager for the start of the 2015 season. Right away, with the help of Altuve, Springer, and Correa, AJ Hinch turned the Houston Astros around into a playoff contender in his first season, finishing second in the AL West with a 86-76 record where they ultimately lost in the ALDS to the Kansas City Royals (the 2015 World Series Champions). This was not a one-off performance, as AJ Hinch turned the Astros into a serious threat for the years following. In the 2016 season, the Astros won 84 games. With Bregman debuting in 2017, the Astros won over 100 games, finishing with a 101-61 record and winning their first ever World Series title. The Astros then went on to have back-to-back-to-back 100 plus win seasons, losing in the ALDS in 2018, and losing in the World Series in 2019 after winning a franchise best 107 games.

After the 2019 World Series loss, the tarnishing of the Astros long-list of accomplishments began. Ex-Astros pitcher, Mike Fiers spoke out about the Astros cheating tactics that he witnessed during his time as a Houston Astros pitcher from 2015-2017. Fiers explained that the Astros batters engaged in a sign-stealing scandal to achieve an unfair advantage over the opposing team and opposing pitchers. The Houston Astros participated in this cheating scheme at their home stadium, Minute Maid Park throughout the 2017 season, and maybe even throughout the postseason as well.

Throughout my experience in baseball, there are legal ways that as a team, you try and gain an advantage over your opponent. As a member of the Tufts baseball team, we always try and see if the opposing pitcher tips his pitches or gives us any indication as to what pitch he might throw. Maybe he pauses longer when he throws an off-speed pitch, maybe he has a lower arm slot when he throws his slider versus his curveball, maybe he is only comfortable performing a slide step when throwing a fastball. There are multiple ways that teams legally try to gain an edge. Teams and players look for patterns that the opponent does that indicates what they might do next.

The Houston Astros did not engage in actions that qualify as a legal way to gain an advantage. What they did was illegal and destroyed the integrity of the nation's greatest pastime. Specifically, the Astros positioned a camera past the center field fence that was zoomed in on the catcher's signals that he gives to the pitcher, to determine what pitch the pitcher will throw to the batter. The camera then sent the video footage to a member of the Astros down by the dugout to decode the pitch being thrown and then subsequently bang a trash can if the pitch is an off-speed pitch. What qualifies as an off speed pitch is anything that is not a fastball (two-seam, four-seam, or a cutter).

The Commissioner in his statement regarding the scandal made it clear that it is not his job to figure out whether or not the Astros benefitted from cheating. Rob Manfred, the MLB commissioner, made it clear that his job was to punish the team for cheating, not punish them for successfully doing so. In the statement he gave, he also mentioned that "some Astros players told my investigators that they did not believe the sign-stealing scheme was effective, and it was more distracting than useful to hitters. I am neither in a position to evaluate whether the scheme helped Astros hitters (who were unquestionably a very talented group), nor whether it helped the Astros win any games. There are so many factors that impact the outcome of games that addressing that issue would require rank speculation. But for purposes of my decision, regardless of whether the scheme was effective or not, it violated the rules and, at a minimum, created the appearance of unfairness, and for that, it necessitates severe discipline" (Manfred). This was important because some Astros players personally thought that the cheating negatively impacted their individual and team results due to the fact that it was distracting to them when trying to hit, which is hard enough at the MLB level as it is.

It is published that the Astros had multiple forms of communication between the center field camera footage and the batter. According to Rob Manfred, "witnesses explained that they initially experimented with communicating sign information by clapping, whistling, or yelling, but that they eventually determined that banging a trash can was the preferred method of communication" (Manfred). In this data set, we are able to only identify instances of "banging a trash can" through the video footage of the Houston Astros home games, and will be accounting only for this method as the form of cheating used by the Astros at their home stadium, in the 2017 season.

Relevant Literature

With such an interesting dataset with widespread implications in the world of sports, we are unsurprisingly not the only individuals to delve into the dataset developed by Tony Adams. A paper by Alexandre Olbrecht and Jeremy Rosen takes a similar binary approach to statistical analysis, but regresses solely on rates such as foul rates, ball-in-play rate, and home runs. Such regressions certainly hold value, as we will carry out two similar regressions on swing and whiff rates, but they do not describe the whole picture (Olbrecht and Rosen). The paper does not accurately conclude whether or not cheating was significant in *improving* performance and instead demonstrates significant *differences* in individual metrics. There is, however, a fatal flaw in the method of their analysis: Olbrecht and Rosen's rate-based analysis does not subset to an at-bat level. We will dive further into the structure of the dataset in the next section, but it is vital to understand how batting outcomes are recorded in the dataset. Imagine there is a home run on a given at-bat. Then the batting outcome is marked as "home run" for each pitch observation in the dataset. If there were 7 pitches during the at-bat, that translates as 7 distinct "home run" outcomes in the dataset. We see that in the Olbrecht and Rosen report that when performing analysis on batting outcomes, there is no subsetting to unique at-bats. This entirely distorts the results of those batting outcome regressions performed. Our thorough data organization and filtration means that we can more confidently draw conclusions from regressions that both include batting outcomes and implicate improvement in performance, as Olbrecht and Rosen were unable to do.

There are several others who have utilized summary statistics and data visualization with use of Tony Adams' dataset. None have done so better than Roderick Henderson, who was able to discern effects of cheating through summary calculations alone (Henderson). However, without statistical testing, he was not able to indicate the magnitude of or confidence behind these conclusions. Our data set up will be done in a way that allows us to perform controlled, accurate regressions from which we can draw conclusions on performance improvement.

Description of Data

Our analysis would not be possible without the work of Houston Astros superfan Tony Adams, who listened to every home game from the Astros' 2017 season and recorded when exactly these "bangs" occurred as well as recording the batter, pitcher, pitch, at-bat outcome, and more (Adams). This dataset provided the fundamentals needed to perform our analysis. The full output of our data organization and filtration is attached at the end (performed in R), but there were originally 8274 observations, or pitches, recorded in the dataset, with 29 variables included.

Given that much of the data provided was qualitative in nature, we decided that we had to segment the data further to obtain metrics necessary to facilitate regressions and statistical conclusions. But first, there were a few cases in which we needed to drop observations from the data set. First, we dropped all observations where the pitch was unable to be categorized and thus was labeled “other”. We then created a dummy binary variable called *offspeed* that is equal to 0 for every pitch that is a two-seam fastball, four-seam fastball, and a cutter-fastball and is equal to 1 otherwise. We do this because as discussed earlier, the cheating methodology was such that the team would bang the trash can for every “off-speed” pitch. Next, we dropped all pitch observations if the given at-bat had no off-speed pitches and no bangs. This is because if all pitches in an at-bat are fastballs, there is likely little way to tell if the team was cheating or not for the given at-bat.

We were then able to create a dummy variable *ischeating* where the value is 1 if the at-bat had a bang and is 0 if there were no bangs throughout the course of the at-bat. From there, we created a variable *cheatedaccurately* that only applies to observations where cheating occurred during the at-bat. This is another dummy variable that is 1 if there was a bang and it was on an off-speed pitch or if it was a fastball and there was no bang, as both situations represent the cheating system being used correctly. Alternatively, the variable is equal to 0 if there was a bang and the pitch was a fastball, or there was no bang and the pitch was off-speed, as both represent the cheating system being used inaccurately.

Before we started, we also wanted to consider the accuracy of the cheating scheme. We sensed that whether or not the pitch they banged the trash can on was correct, it would have a large impact on the variable we were looking at. After analyzing the data and the pitch type correlating to whether or not there was a bang, we were able to calculate that the Astros scandal was accurate about 79.38% of the time.

```
## [1] "Rate of cheating (successful and unsuccessful): 0.384245339747444"
print(paste0("Total number of accurate cheating attempts: ", sum(dataWithOffspeed$cheatedAccurately)))
## [1] "Total number of accurate cheating attempts: 2029"
print(paste0("Total number of inaccurate cheating attempts: ", sum(dataWithOffspeed$cheatedInaccurately
## [1] "Total number of inaccurate cheating attempts: 527"
cheatingSuccessRate <- sum(dataWithOffspeed$cheatedAccurately) / totalCheats
print(paste0("Rate of cheating successfully ", cheatingSuccessRate))
## [1] "Rate of cheating successfully 0.793818466353678"
```

As you can see above, the Astros cheated on about 38.42% of pitches during the 2017 season for a total of 2,556 times in 58 home games. They cheated correctly 2,029 times and inaccurately 527 for a total cheating accuracy rate of 79.38%.

Empirical Results

Before we delve into the results of our analysis, it is important to understand the econometric context with which we are conducting our experiment. The entirety of the dataset provided is qualitatively defined. That is, we have zero interval-based variables from which we develop our regressions. While that may be viewed as a potential limitation of statistical exploration, we argue that it can provide more meaningful, contextual results surrounding the original question we seek to answer. What this means is that our consideration of potential pitfalls in our analysis is not driven by the possibility of homoscedasticity or serial correlation in the errors. We will not utilize an (AR-1) serial correlation test or Breusch-Pagan test for homoscedasticity because we have doubts as to the significance of the magnitude in errors given a context where values are strictly 1 or 0. Instead, the main potential pitfalls to consider are in the construction of dummy variables and regressions such as overgeneralized definition of outcome, multicollinearity, dummy variable traps, and omitted variable bias. They will be discussed at lengths in the context of the regressions below.

Does cheating have an impact on the likelihood of a productive at-bat?

The first question that we wanted to look at was whether or not cheating (and cheating effectively) had an impact on the likelihood of a productive at-bat. Our prediction before running any regressions would be that the Houston Astros cheating scandal helped their batters have more productive at-bats. With this in mind, in this multiple regression analysis with qualitative information, we regressed the variable *productiveatbat* on the variables *ischeating*, *cheatedaccurately*, *on_1b*, *on_2b* and *on_3b*, where the last 3 variables represent whether or not there were runners on first, second, and third base respectively. Reasoning and implications for including these control variables are discussed further below.

But first, we had to define what a productive at-bat was. We chose that a single (1B), double (2B), triple (3B), homerun (HR), walk (BB), hit-by-pitch (HBP), sacrifice fly, and sacrifice bunt were to all be considered productive at-bats. We chose these outcomes as a productive at-bat because these events do not negatively affect the batter's batting average. Some of these: 1B, 2B, 3B, HR all increase the batting average, while BB, HBP, sacrifice bunt and sacrifice fly all do not affect the batting average at all. Our productive at-bat dependent variable is defined as a binary variable based on those outcomes above, so we can interpret its prediction as the likelihood of a productive play on a given at-bat.

```
. regress productiveatbat ischeating cheatedaccurately on_1b on_2b on_3b

      Source |       SS           df       MS      Number of obs   =     6,652
-------------+----------------------------------   F(5, 6646)      =     13.55
       Model |    15.14276         5  3.02855199   Prob > F        =    0.0000
    Residual |   1485.8843     6,646   .22357573   R-squared       =    0.0101
-------------+----------------------------------   Adj R-squared   =    0.0093
       Total |  1501.02706     6,651  .225684417   Root MSE        =    .47284

------------------------------------------------------------------------------
productiveatbat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------
       ischeating |   -.045458   .0218951    -2.08   0.038    -.0883795   -.0025366
cheatedaccurately |   .0339388   .0231977     1.46   0.144    -.0115362    .0794138
            on_1b |   .0359232   .0130578     2.75   0.006     .0103257    .0615208
            on_2b |   .0446315   .0148991     3.00   0.003     .0154244    .0738385
            on_3b |   .1143615   .0204098     5.60   0.000     .0743518    .1543712
            _cons |   .3210641   .0088019    36.48   0.000     .3038095    .3383187
------------------------------------------------------------------------------
```

As we see here in this first regression, the signage of *ischeating* is negative, with a p-value of .038 which is significant at the 5% significance level. What this initially tells us is that overall, cheating had a negative impact on the at-bat outcome productivity. However, this figure alone does not take into consideration the accuracy of the cheating attempt. It is not simply shallow to consider only the *ischeating* variable but negligent, as there is a dependence relation with the *cheatedaccurately* variable. That is, *cheatedaccurately* can only take on the value of 1 when *ischeating* is 1. We avoid a dummy variable trap by omitting the dummy variable of cheating inaccurately. We do note a low R^2 value (as is the trend throughout our analysis) that is a strong indicator that we do not have all of the necessary predictors of an at-bat outcome. For instance, we assume that the pitcher being faced by the batter to be a significant indicator of the likelihood of a positive outcome. Other factors like the team being faced, inning, and batter all likely contribute but are beyond the scope of this analysis. This also brings to consideration omitted variable bias, and the potential that some of statistical significance attributed to cheating would be better captured by other variables. For instance, given that the team did not always cheat and instead seemed to choose situations to do so, it is possible that the current game state was an important cofactor of cheating. Perhaps the Astros tended to cheat more in "big play situations," where there was a higher chance of runners scoring, and thus a higher likelihood of a productive at-bat. We are able to limit some of this with our controls (to be discussed below), but lingers as a question in our analysis.

The dependence relation does not undermine the validity of the regression via multicollinearity, either, as we have no reason to believe there is any underlying correlation between the team choosing to cheat on any given pitch and said cheating being accurate or not. If it were the case that the team was more likely to choose to cheat when, say, signs were easier to read, and such conditions meant that cheating more accurately took place, then we may potentially run into issues of correlation in the regressors. But given what is known of the scandal, we are not led to believe that is the case.

The coefficient of the *cheatedaccurately* variable is less significant, with a p-value of .144 which is just barely insignificant at the 10% significance level. However, its positive signage and near-statistical significance lets us entertain quite interesting and perhaps significant results. If we have a cheating at-bat where the team accurately cheated, our dummy variables give us an expected impact of -.045(1) + 0.034(1). This gives a -1.1% less likelihood of a productive play. When there is an at-bat and the cheating is inaccurate, the expected impact is -.045(1) + 0.034(0). This gives a -4.5% less likelihood of a productive at-bat. We conclude at just over the 10% significance level that even when cheating is accurate, it net negatively impacted the outcome of the at-bat. However when the cheating was inaccurate, there was a much more significant negative impact on the outcome productivity of the at-bat.

As expected, all three of our *on_1b*, *on_2*, and *on_3b* variables are statistically significant even at the 1% significance level. These three variables are effectively control variables because of their unique relationship with specific outcomes of an at-bat. We expected this significance given our definition of a productive at-bat. There are certain outcomes that are independent of there being runners on such as a single, double, triple, homerun, walk, and hit-by-pitch. However, there are also outcomes that are dependent on there being runners on base. For example, a sacrifice fly and a sacrifice bunt are both productive at-bats, but only qualify as such if there is a runner on base when it occurs, or else they would simply be groundout and flyouts. So it makes sense that runners being on base results in an expected increase in likelihood of a productive play. Intuitively, we also see that the coefficients increase from *on_1b* to *on_2b* to *on_3b* because of the higher likelihood that a runner scores if they are further along on the basepaths. Meaning that runners close to home are more likely to predict a higher likelihood of a productive at-bat. These effective controls help to mitigate the generalization from creating a single outcome variable that encompasses several different batting outcomes.

Does Cheating have a measurable impact on players' swing rates?

We found that cheating had a net negative impact on the likelihood of a productive at-bat. Although these findings answered the question we set out to answer most, we held hypotheses about the measurable impact of cheating on other metrics. We hypothesized that with the ability to know the incoming pitch from the opposition, players may be more confident at the plate, swinging at off-speed pitches that they may not normally swing at. So, we opted to perform a regression to put this hypothesis to the test. For this regression we developed another dummy variable based on pitch outcome that was equal to 1 if the pitch outcome was one of 10 swinging-based outcomes, and 0 otherwise.

Similar to the prior regression on at-bat productivity, we chose to add the control variables of being runners on first, second, and third base. We hypothesized that when there were runners on base, players' swing tendencies may be more aggressive, in the hopes of bringing runners home.

We included the accuracy variable given the possibility that if as the pitch came in the batter realized it was a different pitch than expected (hence cheating incorrectly), then perhaps they would be less likely to swing. The results of the regression are shown below.

```
. regress swungat ischeating cheatedaccurately on_1b on_2b on_3b

      Source |       SS           df       MS      Number of obs   =     6,652
-------------+----------------------------------   F(5, 6646)      =      2.16
       Model |  2.65834696         5  .531669392   Prob > F        =    0.0556
    Residual |  1635.80527     6,646  .246133806   R-squared       =    0.0016
-------------+----------------------------------   Adj R-squared   =    0.0009
       Total |  1638.46362     6,651  .246348462   Root MSE        =    .49612

-----------------------------------------------------------------------------------
          swungat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
       ischeating |    -.02323   .0229731    -1.01   0.312    -.0682647    .0218048
cheatedaccurately |   .0397284   .0243399     1.63   0.103    -.0079856    .0874425
            on_1b |  -.0170671   .0137008    -1.25   0.213     -.043925    .0097908
            on_2b |    .008869   .0156327     0.57   0.571    -.0217761    .0395142
            on_3b |   .0528544   .0214147     2.47   0.014     .0108747     .094834
            _cons |   .4343088   .0092353    47.03   0.000     .4162046     .452413
-----------------------------------------------------------------------------------
```

We generally conclude statistical insignificance from this regression. Both the *ischeating* and *cheatedaccurately* variables are statistically insignificant at the 10% significance level. In fact, we find that almost all independent variables in the regression are statistically insignificant. The one exception is the variable *on_3b*, which has a coefficient of about 0.053 and is significant at the 5% level. This supports the theory that batters may swing more aggressively when the chance at bringing in a run is higher. The prediction states that swing rate is expected to increase by about 5% given that there is a runner on 3rd base.

Given the original intentions of the regression, we are still hesitant to take too much away from it. It is unfortunate that we cannot make further conclusions about changes in swing patterns but provides us with the opportunity to take a look at one final metric: whiff rate.

Does Cheating have a measurable impact on players' whiffing rates?

Although the findings from our main regression on at-bat productivity conclusively pointed toward cheating have a net negative effect, we found it hard to believe that there were not metrics where we could see conclusive improvement with cheating. Plus, given the rather general definition of a productive at-bat, essentially making identical several batting outcomes that are more realistically a spectrum of varying levels of success, there was certainly more to explore. One measurable metric given the pitch outcomes is the whiff rate, defined as the rate at which a batter swings at a pitch but does not make any contact with the ball.

For this regression, we subset the data only to pitches that were swung at. Then, we created a dummy binary variable *whiffedat* which was equal to 1 if the pitch outcome was one of swinging strike, swinging strike (blocked), or missed bunt. Since we are only working with observations where the batter swung, the variable equals 0 in all other cases as contact was made.

From here, we regressed this *whiffedat* variable on *ischeating* and *cheatedaccurately*. In this case we do not include the controls of batters being on base as there is no expected relationship between the batter's ability to make contact with the ball and there being a batter on base. The results of the regression are shown below.

```
. regress whiffedat ischeating cheatedaccurately

      Source |       SS           df       MS      Number of obs   =     2,922
-------------+----------------------------------   F(2, 2919)      =     19.90
       Model |  6.79852316         2  3.39926158   Prob > F        =    0.0000
    Residual |  498.608732     2,919  .170814913   R-squared       =    0.0135
-------------+----------------------------------   Adj R-squared   =    0.0128
       Total |  505.407255     2,921  .173025421   Root MSE        =     .4133

------------------------------------------------------------------------------------
         whiffedat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
        ischeating |   .1003678   .0296512     3.38   0.001     .0422283    .1585073
 cheatedaccurately |  -.1771402   .0311388    -5.69   0.000    -.2381965   -.1160838
             _cons |   .2390817   .0097796    24.45   0.000     .2199061    .2582574
------------------------------------------------------------------------------------
```

There are most interesting findings from this regression. First off, we see extremely high statistical significance across both variables included in the regression, with p values of nearly zero. It then becomes quite interesting to look at the coefficients from this regression and discern their meaning.

We notice first that the coefficient of *ischeating* is positive and quite large in magnitude, with a value of 0.1. Once again, we cannot consider the implications of this coefficient alone - we must also take into account the value of the *cheatedaccurately* coefficient, which is -0.177. So, for a pitch in which the batter cheats accurately, we have an expected impact on the whiff rate of 0.1004(1) + -0.177(1) = -0.0766, or a 7.66% drop in the whiff rate. Alternatively, should the batter cheat inaccurately, we instead have an expected impact of 0.1004(1) + -0.177(0) = 10.04% increase in the whiff rate. For context, our summary statistics from our data set-up shows that the average whiff rate on the season was 22.2%, showing just how drastic the magnitude of these changes in percentages are.

That is to say that when the batter is able to cheat correctly and knows the pitch that is coming, there is significant improvement in the batter’s ability to make contact with the ball, a

7% increase in likelihood of contact. However, if the batter mistakes the pitch, there is a significant swing in the opposite direction, making the batter instead 10% more likely to miss.

Given both the statistical significance and the magnitude of these coefficients, these are quite significant findings. This suggests that while cheating may not have improved batters' ability to have productive at-bats overall, it could aid them in making contact with the ball. The catch? They had to get it right. Should the cheating signal predict the wrong pitch, batters were made significantly worse off than if they had not cheated at all. Intuitively, this makes sense - we expect that when the batter has some expectation of the incoming pitch but the pitch is different, it would be even more difficult to make the adjustment in swing than simply reacting to the pitch normally. But, should that prediction from cheating be correct, the batter will then have a better chance at contact.

Conclusion

As for our hypothesis that this cheating scandal actually helped the overall success of the Houston Astros in 2017, we found quite the contrary through our analysis. Rob Manfred said that some Astros players said that they did not think the sign-stealing scheme was effective because it was more distracting than useful to hitters, and it looks as though this statement was backed up by our regressions.

For our first analysis we regressed productiveatbat on the variables *ischeating*, *cheatedaccurately*, *on_1b*, *on_2b* and *on_3b*. Our conclusion was that when the Houston Astros cheated, it had a negative impact on the probability of a productive at-bat. When the Houston Astros cheated correctly, there was a -1.1% less likelihood of a productive play. But when the Astros cheated incorrectly, there was a -4.5% less likelihood of a productive at-bat. We did find a p-value that made us hesitant to conclude solely with this analysis, so we opted for further exploration.

Our second analysis looked at the impact of cheating on player swing rates. From this regression, we found that both the *ischeating* and *cheatedaccurately* variables were statistically insignificant at the 10% significance level. Our only statistically significant takeaway came from there being a runner on 3rd base, where swing rates were expected to increase by about 5%, which made sense intuitively given the higher incentive to swing aggressively.

Because swing rates were insignificant, we finally looked at whiff rates to see if there was significance, and perhaps improvement, given cheating. Right off the bat, we see that all of the variables regressed are extremely significant. When the Astros cheated accurately, there was an expected 7.66% drop in the whiff rate. However, when the Astros cheated inaccurately, we instead had an expected 10.04% increase in whiffs. Here lies the danger of cheating. When able to pull it off correctly, the Astros improved their chances at contact. But if the cheat was mistaken, the effects were fatal. Even though we saw that Astros whiffed less when cheating, most of that contact resulted in foul balls, and not productive at-bats. That is why we do not see the improvement from cheating translate to productive at-bats.

Overall, we draw the conclusion that the cheating scandal did not lead to a significant improvement in the teams ability to perform on a pitch-to-pitch, and likely game, basis but certainly was an unfair method of improving the ability of Astros batters to make contact, especially when executed correctly.

Works Cited

Adams, Tony. “Sign Stealing Scandal.” 2020, http://signstealingscandal.com. Accessed 20 December 2020.

Henderson, Roderick. “The Astros' sign-stealing: exploring a scandal using data visualization.” May 2020, https://public.tableau.com/views/ASign-StealingScandal/Story?:display_count=y&:origin=viz_share_link&:showVizHome=no. Accessed 20 December 2020.

Manfred, Rob. “Statement of the Commissioner.” 13 January 2020, https://img.mlbstatic.com/mlb-images/image/upload/mlb/cglrhmlrwwbkacty27l7.pdf. Accessed 20 December 2020.

Olbrecht, Alexandre, and Jeremy Rosen. “Do the Data Support the Major League Baseball Commissioner’s Report on the Houston Astros Sign-Stealing Scandal?” 2020, https://drive.google.com/file/d/1is3jCh851Q5iF2AT78-0brBV4a_boLCT/view. Accessed 20 December 2020.

EC107 Research Project

Devan Venkataraman and Nathaniel Bozzella

12/12/2020

Data Set-up

We begin by importing the Astros 2017 Bang Data set.

```
initialData <- read.table("astros_bangs.csv", sep=",", header=TRUE)

dim(initialData)
```

```
## [1] 8274   29
```

```
summary(initialData)
```

```
##                           game_id        game_pk           game_date
##  2017_04_05_seamlb_houmlb_1: 216   Min.   :490111   5/4/2017 : 216
##  2017_05_04_texmlb_houmlb_1: 197   1st Qu.:490528   4/5/2017 : 197
##  2017_04_09_kcamlb_houmlb_1: 192   Median :491275   9/4/2017 : 192
##  2017_06_14_texmlb_houmlb_1: 190   Mean   :491202   14/6/2017: 190
##  2017_07_14_minmlb_houmlb_1: 179   3rd Qu.:491752   14/7/2017: 179
##  2017_08_04_tormlb_houmlb_1: 175   Max.   :492424   4/8/2017 : 175
##  (Other)                   :7125                    (Other)  :7125
##     opponent    final_away_runs  final_home_runs      inning
##  ANA    :1158   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
##  SEA    :1033   1st Qu.: 2.000   1st Qu.: 3.000   1st Qu.: 3.000
##  OAK    : 944   Median : 4.000   Median : 4.000   Median : 5.000
##  TEX    : 808   Mean   : 4.169   Mean   : 4.944   Mean   : 4.803
##  TOR    : 473   3rd Qu.: 6.000   3rd Qu.: 7.000   3rd Qu.: 7.000
##  MIN    : 472   Max.   :13.000   Max.   :16.000   Max.   :13.000
##  (Other):3386
##   top_bottom               batter        at_bat_event  pitch_type_code
##  bottom:8274   George Springer: 933   Strikeout:1851   FF     :2326
##                Jose Altuve    : 866   Groundout:1418   SL     :1500
##                Alex Bregman   : 800   Single   :1138   FT     :1474
##                Marwin Gonzalez: 776   Walk     : 986   CH     : 767
##                Carlos Beltran : 762   Flyout   : 782   CU     : 704
##                Josh Reddick   : 725   Lineout  : 452   FC     : 488
##                (Other)        :3412   (Other)  :1647   (Other):1015
##  pitch_category has_bangs bangs       call_code            description
##  BR:2518        n:7132       :7132   B       :2842   Ball            :2842
##  CH: 772        y:1142    1B: 790    C       :1397   Called Strike   :1397
##  FB:4959                  2B: 330    F       :1374   Foul            :1357
```

```
##  OT:  25                  3B:  16   X      :1018   In play, out(s):1018
##                           4B:   5   S      : 668   Swinging Strike: 668
##                           5B:   1   D      : 346   In play, no out: 346
##                                     (Other): 629   (Other)        : 646
##  on_1b    on_2b    on_3b         youtube_id   pitch_youtube_seconds
##  f:5878   f:6607   f:7503   WTsHakP381M: 216   Min.   :  874
##  t:2396   t:1667   t: 771   ZIYTuQ0OckU: 197   1st Qu.: 3416
##                             lnQFgqF5ZQs: 192   Median : 5982
##                             MYrjL7KG1fc: 190   Mean   : 6319
##                             7FYj9_5sQk4: 179   3rd Qu.: 9046
##                             pkwjm2oglGw: 175   Max.   :16594
##                             (Other)    :7125
##                                                youtube_url
##  https://www.youtube.com/watch?v=2viSQ9q3SoQ&t=7343 :   2
##  https://www.youtube.com/watch?v=af5e55Cc8ZA&t=7997 :   2
##  https://www.youtube.com/watch?v=iP2zDfaBz9w&t=5801 :   2
##  https://www.youtube.com/watch?v=iP2zDfaBz9w&t=9365 :   2
##  https://www.youtube.com/watch?v=MUMI3u7ftZw&t=10836:   2
##  https://www.youtube.com/watch?v=_h8iCp005rg&t=10480:   1
##  (Other)                                            :8263
##                   pitch_datetime      game_pitch_id   event_number
##  17/4/2017 20:36:49-05:05:   2   170715035707:  10   Min.   : 17.0
##  17/4/2017 21:36:13-05:05:   2               :   8   1st Qu.:157.0
##  3/4/2017 21:15:00-05:05 :   2   170823023949:   7   Median :297.0
##  6/4/2017 22:00:44-05:05 :   2   170801032800:   6   Mean   :313.5
##  1/7/2017 18:24:37-05:05 :   1   170524032323:   5   3rd Qu.:458.0
##  1/7/2017 18:24:54-05:05 :   1   170629203634:   5   Max.   :908.0
##  (Other)                 :8264   (Other)     :8233
##                                pitch_playid
##  00041f48-3fd9-4c9f-a10c-6155e30b2dde:   1
##  00277525-b8d2-4f40-b067-8885e3b701e1:   1
##  002ab87f-bf4f-443b-83a9-d60e73700825:   1
##  002ad03b-8daa-4462-9c5e-5469dcff133f:   1
##  002b9748-b10c-4d77-a174-485212fef4db:   1
##  00353bc9-ae29-4e18-b52b-864b9f94b38e:   1
##  (Other)                             :8268
##                                atbat_playid   away_team_id    home_team_id
##  64e881a0-adc8-482b-b8d8-36c582c57f0a:  12   Min.   :108.0   Min.   :117
##  98597c10-2545-48db-b217-204d4f092c22:  12   1st Qu.:116.0   1st Qu.:117
##  ce1ec57d-fdc8-474a-ba43-241420626701:  12   Median :136.0   Median :117
##  de0fe58a-d083-4d75-91ac-4b47e03ec89b:  12   Mean   :129.2   Mean   :117
##  3eb1dc5a-95eb-4b80-9d2b-ca62526ecce2:  11   3rd Qu.:140.0   3rd Qu.:117
##  6b4ad3c4-2345-45a3-9508-f5bd1298db87:  11   Max.   :147.0   Max.   :117
##  (Other)                             :8204
```

```r
cleanedData <- subset(initialData, select = -c(youtube_id, pitch_youtube_seconds, youtube_url, pitch_da
```

Data Clean-up + At-Bat Setup

The data set is composed of observations that represent individual pitches from pitcher to catcher with a given batter. In addition to the testing we will do on these individual pitches, there is addititional testing that we hope to perform using at-bats as the “universe” in which we are investigating. With this in mind, we will create several variables that are "at-bat based," which will allow us to subset a separate data set that includes only unique at bats, letting these new variables drive inference.

```r
# An off speed pitch is any pitch that is not a fastball
isOffspeed <- ifelse(cleanedData$pitch_category == "FB", 0, 1)
dataWithOffspeed <- cleanedData

# Create the binary variable
dataWithOffspeed["isOffspeed"] <- isOffspeed

# We cannot infer when the pitch isn't categorized - drop "other" pitch observations
dataWithOffspeed <- dataWithOffspeed[!(dataWithOffspeed$pitch_category == "OT"),]

# WQe also cannot make inferences about at-bat when it ends with fan or catcher interference.
# Drop those observations as well
dataWithOffspeed <- dataWithOffspeed[!(dataWithOffspeed$at_bat_event == "Catcher Interference"),]
dataWithOffspeed <- dataWithOffspeed[!(dataWithOffspeed$at_bat_event == "Fan Interference"),]

# We're going to want to use the on-base variables as controls in our regressions, so convert
# our on-base factor variables to be a 0/1 integer binary
dataWithOffspeed$on_1b <- ifelse(dataWithOffspeed$on_1b == "t", 1, 0)
dataWithOffspeed$on_2b <- ifelse(dataWithOffspeed$on_2b == "t", 1, 0)
dataWithOffspeed$on_3b <- ifelse(dataWithOffspeed$on_3b == "t", 1, 0)

# Create columns for new, at-bat-based binary variables
dataWithOffspeed["abHadAnOffspeed"] <- 0
dataWithOffspeed["abHadABang"] <- 0

# Create binary values that we'll conditionally set
abWithOffspeed <- 0
abWithBang <- 0

# For each unique at bat...
for (i in unique(dataWithOffspeed$atbat_playid)) {
  thisPitch <- dataWithOffspeed[i,]
  # Create a data frame that represents the set of pitches in a unique at bat
  thisAB <- dataWithOffspeed[dataWithOffspeed$atbat_playid == i, ]
  thisAB <- thisAB[complete.cases(thisAB),]

  # Do we have an offspeed pitch anywhere in this at bat?
  if(any(thisAB$isOffspeed == 1)) {
    abWithOffspeed <- 1
  } else {
    abWithOffspeed <- 0
  }

  # Do we have a bang anywhere in this at bat?
  if(any(thisAB$has_bangs == "y")) {
    abWithBang <- 1
  } else {
    abWithBang <- 0
  }

  # Given our findings from the at bat, set the respective variable
```

```
  # value for each observation in the at bat
  dataWithOffspeed[dataWithOffspeed$atbat_playid == i, ]$abHadAnOffspeed <- abWithOffspeed
  dataWithOffspeed[dataWithOffspeed$atbat_playid == i, ]$abHadABang <- abWithBang
}
```

Cheating Setup

Now we have information about whether there were fastballs / off-speed pitches in the at bat and whether there were bangs during the at-bat. We can now apply our understanding of how the Astros cheating scandal worked to create variables that recognize and signify both if the team was cheating or not and if they cheated accurately or not. These two binary variables will drive our statistical testing.

```
# If the at bat doesn't have an offspeed pitch we cannot tell if they were cheating or not
# because they wouldn't bang anyway. So drop those observations
dataWithOffspeed <- dataWithOffspeed[!(dataWithOffspeed$abHadAnOffspeed == 0),]

# Our "at bat has a bang" variable is now synonymous with cheating because we know that
# each of these at bats should have a bang... if they're cheating
names(dataWithOffspeed)[names(dataWithOffspeed) == "abHadABang"] <- "isCheating"

# Now, create variables to see if they were cheating accurately - this will
# only going to apply to observations where cheating = 1
dataWithOffspeed["cheatedAccurately"] <- 0
dataWithOffspeed["cheatedInaccurately"] <- 0

# First case of cheating accurately - pitch was a fastball and there was no bang
dataWithOffspeed$cheatedAccurately <- ifelse((dataWithOffspeed$isCheating == 1 & dataWithOffspeed$isOff

# Alternative case of cheating accurately - pitch was offspeed and there was a bang
dataWithOffspeed$cheatedAccurately <- ifelse((dataWithOffspeed$isCheating == 1 & dataWithOffspeed$isOff

# First case of cheating inaccurately - pitch was a fastball and there WAS a bang
dataWithOffspeed$cheatedInaccurately <- ifelse((dataWithOffspeed$isCheating == 1 & dataWithOffspeed$isO

# Alternative case of cheating inaccurately - pitch was offspeed and there WAS NOT a bang
dataWithOffspeed$cheatedInaccurately <- ifelse((dataWithOffspeed$isCheating == 1 & dataWithOffspeed$isO

# Let's calculate some basic summary metrics
totalCheats <- sum(dataWithOffspeed$cheatedAccurately) + sum(dataWithOffspeed$cheatedInaccurately)
cheatingRate <- totalCheats / nrow(dataWithOffspeed)
cheatingSuccessRate <- sum(dataWithOffspeed$cheatedAccurately) / totalCheats

print(paste0("Rate of cheating (successful and unsuccessful) to all at-bats: ", cheatingRate))
```

```
## [1] "Rate of cheating (successful and unsuccessful) to all at-bats: 0.384245339747444"
```

```
print(paste0("Total number of accurate cheating attempts: ", sum(dataWithOffspeed$cheatedAccurately)))
```

```
## [1] "Total number of accurate cheating attempts: 2029"
```

```r
print(paste0("Total number of inaccurate cheating attempts: ", sum(dataWithOffspeed$cheatedInaccurately
```

```
## [1] "Total number of inaccurate cheating attempts: 527"
```

```r
print(paste0("Rate of cheating successfully to all instances of cheating", cheatingSuccessRate))
```

```
## [1] "Rate of cheating successfully to all instances of cheating0.793818466353678"
```

Result-based Variable Setup

Now we have information about the team's cheating on an at-bat and pitch-based basis. However, the data set, as provided, does not provide us quantified results of pitches and at-bats. While we could theoretically perform statistical analysis using the batting results as a factor for our dependent variable, there are 25 batting outcomes in the data, with varying totals from 5 to over 700. Regression analysis on these individual results would lead to highly fractured results with wide confidence intervals. Instead, we will take an approach by determining whether batting outcomes are considered what we will call "productive at-bats." These are said to positively contribute to a team's offense. We will then create additional binary variables that will support inference on players' swinging and contact rates.

```r
# create the variable
dataWithOffspeed["productiveAtBat"] <- 0

# a productive at bat is one where the batting outcome is one of several defined outcomes.
# If it's not one of these outcomes, it is not productive.
dataWithOffspeed$productiveAtBat <- ifelse((dataWithOffspeed$at_bat_event == "Single" | dataWithOffspee

# create the variable
dataWithOffspeed["swungAt"] <- 0

# a pitch is swung at if the "description" is any of the below outcomes. Otherwise it is not swung at
dataWithOffspeed$swungAt <- ifelse((dataWithOffspeed$description == "Foul" | dataWithOffspeed$descripti

# create the variable
dataWithOffspeed["whiffedAt"] <- 0

# a pitch is whiffed at if they swing and do not make contact with the pitch, defined as the outcomes b
dataWithOffspeed$whiffedAt <- ifelse((dataWithOffspeed$description == "Swinging Strike" | dataWithOffspe

whiffrate <- sum(dataWithOffspeed$whiffedAt) / sum(dataWithOffspeed$swungAt)
print(paste0("Rate of whiffing: ", whiffrate))
```

```
## [1] "Rate of whiffing: 0.222450376454483"
```

Data Exports (for Stata)

We export 3 unique datasets from our master set we have created with the above. The first data set exported is the original data set, which will support inference performed on all pitches. The second is a dataset of unique at-bats, which will support inference performed on at-bats. Finally, we export a third data set that is a subset of pitches, those that the batter swung at. This will support inference performed on the pitch.

```r
# rename
dataMaster <- dataWithOffspeed

# subset original data set to create needed data sets
atBatData <- dataMaster[!duplicated(dataMaster$atbat_playid),]
swingData <- dataMaster[dataMaster$swungAt == 1, ]

write.csv(dataMaster, "dataMaster.csv")
write.csv(atBatData, "atBatData.csv")
write.csv(swingData, "swingData.csv")
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.