Behind the batter: underlining statistics

Exploring year to year and performance correlations

Prospective and explorative considerations

*Written by Zilio Alessandro*

What’s more important: the way you are doing things or the endgame, the results of all your peripeties? Are you a fan of the Process™ or a “all meets as long as they lead to the end” kinda guy?

In today’s baseball the choice comes yet again knocking at the door thanks to technology’s spread: nowadays pretty much everything that happens on a baseball field can be measured, not just something trivial as the speed of a fastball but also its rpms, spin direction and efficiency, horizontal and vertical movement and another good bazillion of intricacies revolving around a guy throwing a ball 60.6 feet from home plate.

The same concept applies to batters: while our eyes could already tell us whether a ball was hit in the ground, on a line or in the air, now we have at our disposal the velocity at which the ball exited after its impact (if any) with the bat, the angle at which the same ball starts its journey and the amount of top/backspin imparted.

That’s not to say technology is the only hero in this story, we as humans, sabermetricians at least, became smarter: we always had the data, we knew the rules, but we lacked the intuition to put them together.

Take the strike zone: the rulebook gives somewhat of an idea of its dimensions and since the good old days we are used to see a (sponsored) 9 cells strike zone on our screens, yet plate discipline statistics, the ones related to the existence of the strike zone itself, have become increasingly popular only in the last few years. That’s because writers and sabermetrics found out how those stats relate to the core performance of a player, therefore his value and his worth in juicy dollars.

What I’m going to do is something that will sound incredibly boring and yet, as almost all of the basics we tend to learn and forget with time, it’s going to provide a bounty of fodder for thoughts: correlations, and a lot of them!

First of all, what’s in store for stats? I chose a succulent selection of some well-known statistics, from the ones describing a player’s (in)competence to the so called *underlying* traits, numbers and percentages that tell the story of a player’s way to go about his craft behind the scenes, his swing decisions and propensions and what they mean when applied to a seamed spherical object going his direction. For your enjoyment, here is the main course:

* **WAR**: you know what it is! Our overlord since the early 2000s, the “tell all” we like to dislike and yet keep using when it’s about recommending a player, criticizing a trade or projecting a prospect’s future value;
* **wOBA**: weighted On Base Average, one of the steps before getting to WAR that takes the old AVG further by pondering each hit outcome for its value in bases depending on a system of linear weights;
* **wRC+:** weighted Runs Created +, a derivate of wOBA that is also park and league adjusted, where the plus sign indicates a normalization so that 100 is baseline and each point more/less than that is a percentage point more/less in terms of performance with respect to league average;
* **ISO**: Isolated Power, as the number of extra bases per at bat (or SLG – AVG for the nostalgic crowd);
* **BABIP**: Batting Average on Balls In Play, a retooled AVG that considers only balls hit into the confines of the stadiums, defined as the ratio between all hits without home runs and the number of ABs in play, that excludes the aforementioned bombs, sacrifice hits and strikeouts. This little guy will need some further explanation as to why it is also known as the “luck” factor;
* **GB/FB** ratio: simply the number of balls hit in the ground divided by those hit in the air, this gimmick helps to identify a player as either a GB hitter, ratio >1, or a FB hitter, ratio <1, and to reason on the value of where a ball is put in play;
* **Swing%**, **O-Swing%** and **Z-Swing%:** in order, the percentage of swings at pitches by a hitter, the percentage of swings at pitches outside the strike zone and that at pitches inside the zone. Note that the zone can be defined in several ways (by the book, by sites s.a. Baseball Info Solutions and more recently by Statcast) so small discrepancies are possible. These are a part of the so-called *plate discipline* statistics, telling us about the eye of a player at the batter’s box and if he’s either aggressive or patient;
* **Contact%**, **O-Contact%** and **Z-Contact%:** respectively, percentage of pitches resulting in a bat to ball contact on the total number of swings and percentage of pitches contacted outside/inside the strike zone over the number of swings on pitches outside/inside the said zone. Intuitively these are also plate discipline stats, but rather than telling about the approach of a player’s at bats they recount his success in following his plan, although a contact is different than a hit, so onlookers beware;
* **Swinging Strike%:** the percentage of swings and misses by a batter on the pitches he sees, can be attached to the contact percentages as a Non-Contact%;
* **Soft%**, **Med%** and **Hard%**: also called *quality of contact* stats, they consider the percentages of balls hit by a player on a pure strength profile as per a B.I.S. algorithm. Note that these stats have lost some traction since the entry of Statcast and its more precise definitions of contact quality;
* **Pull%**, **Cent%** and **Oppo%**: also called *batted ball direction* stats, they profile a player according to the zones of the field where he hits balls. Pulling a ball means that a RH hitter hits the ball to left field and the opposite for LH hitters, while going oppo means hitting to the other part of the field with respect to the batter’s box. A pull hitter is one that pulls more than 40% of his batted balls, while oppo hitters are rare and a spray hitter is one without a clear pull/oppo preference;
* **Exit Velocity**: the average velocity at which balls hit by a certain batter exit the bat and start their trajectory. A feat only possible with the introduction of Statcast, it is quite telling about the raw power of a player: someone able to hit balls at 110+ mph EV is a scary dude to pitch to for sure;
* **Launch Angle**: the average angle in grades at which balls hit by a certain player exit the bat and start their trajectory. Another gift of Statcast, it is not as straightforward as EV, where the more the better, as hitting balls too down (<10°) or too much under (>35°) at the speed of light will maybe give you a hit but not a trip around the bases. A good launch angle can be considered into the 15°-30° range, although it’s also possible to sneak a liner or a towering popup out in some stadiums (the Pesky Pole in Boston, the Crawford Boxes in Houston and….well the whole Yankee Stadium);
* **Hard Hit%** and **Barrel%**: the new *quality of contact* stats made possible by Statcast, they identify respectively the percentage of balls hit at 95+ mph and the percentage of balls struck at a combination of EV and LA returning a minimum expected .500 AVG and 1.500 SLG (those known before as “well-struck” balls by the eye test). These two metrics are increasingly important as they define not only the propension of a player on hitting the ball with some thump but also with the right bat path leading to an optimal result.

Ideally, I divided all these stats into three major categories:

* The first five, from WAR to BABIP, are marked as *resume stats,* sort of plug and play figures that tell us the value of a player in a general way (WAR), with his bat (wRC+,wOBA), his power output (ISO) and whether he had a rough or blessed season (BABIP).
* The GB/FB ratio and all the %stats until EV comprise the *profile/tendency stats,* recounting a player’s whole existence at the plate as a batter: where he hits, how much he swings in different pitch locations and how many of those swings get to the ball or miss. Note that I consider a *profile* and a *tendency* to be different: while for the former the batter has some kind of control over, a set plan or preference, the latter is more of a resultant of his reactions, instincts and pure skills with the bat. Therefore GB/FB ratio, Swing% stats and direction stats, mostly Pull%, are *profile stats*: a batter can decide to just be a free swinger (hello, Javy Baez!) or a patient stoic (I see you, Mike Trout), an extreme flyball pull hitter a la Joey Gallo or a spray line driver in the DJ LeMahieu mold. On the other hand, the Contact% stats, Swinging Strike% and quality of contact stats are *tendency stats:* you can’t simply decide to hit every single ball thrown into the zone as a Willians Astudillo does, but you also can’t wake up in the morning and become a bad ball hitter at will (think at Jose Altuve or Eddie Rosario for pitches in) or a hard hitting, hard whiffing Miguel Sano-like player.
* The last four, from EV to Barrel%, are the *Statcastats*, because…well you get the gist! EV and LA are side dishes, the first to the tendency stats and the second to profile stats, kicking in some more flavor when needed, either reinforcing a player’s swing choices (high avg LA signals a FB hitter) or his hitting talents (high avg EV means hitting the ball harder). HardHit% and Barrel% are instead closer to the resume stats as they point to characteristic of a player as a value producer: hitting the ball extremely hard (95+ mph EV) and dealing damage (at least 1.500 xSLG). Barrels can be considered as an EV+LA combo, although more of a subset in this dataset as both components are expressed as averages.

What about those whom the stats refer to, the cast of this short form movie, the players?

Usually I would have taken full season stats and call it a day but the 2020 season, in its weirdness given length, circumstances and rule changes, made it a less obvious choice: for the scope of this analysis I took all batters that had at least 80 PA in 2020 and at least 200 PA in 2017,2018 and 2019, but not necessarily only those who met both PA thresholds. By the way, the thresholds I’ve set are a third of full season PAs, a chunk of events that should stabilize at least some of the stats in the platter.

What I ended up are 4 datasets of 300+ rows (batters) per 26 columns (23 stats + team and batter name and correspondent player id), with values in “pure numbers”, wRC+ in hundreds or ISO in tenths, but also percentages.

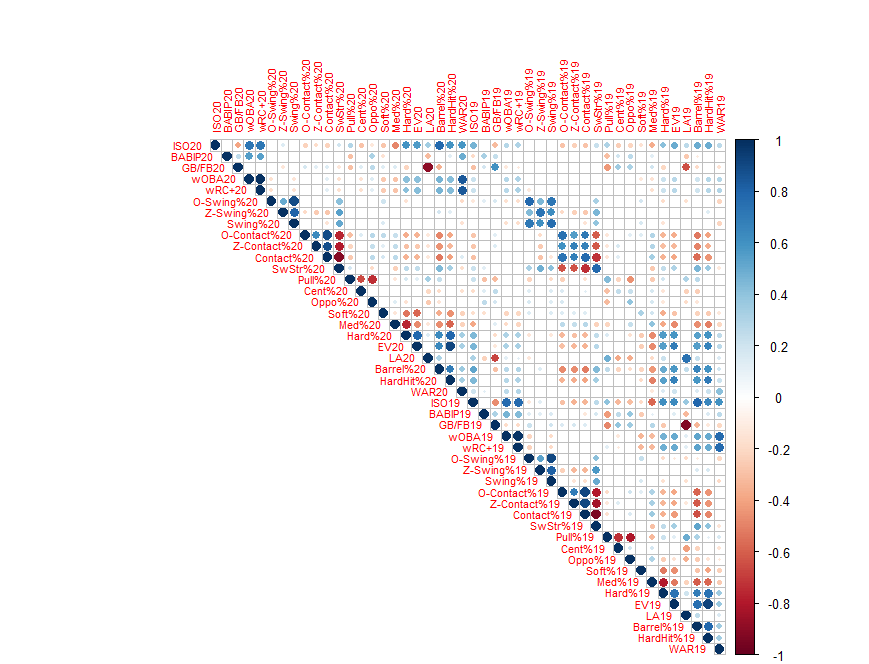
Before going into more detail, you could ask why I chose to go back all the way to 2017 and that is closely linked to what was the original scope of this analysis: finding the steadiest underlining stats for batters and then consider their value in a predictive way on future performances.

By studying the correlation between each 2020 stat on our bucket with itself but on previous years I wanted to check which, among resume, profile, tendency and Statcast stats are the toughest to wane with time, such that they are a good look at a batter’s characteristics also after a couple of years. My measure of toughness, or steadiness if you like, is the canonical correlation (R) and a positive one as you would expect from the same stat seen on a yearly basis: the higher the correlation the closer the relationship among year(s) to year statistic.

Then I would have proceeded by taking only those stats after a R threshold (0.5 in my case) and performed a linear regression of the year n WAR on those stats but at year n-1, n-2 and n-3: in this article it would have been regressing 2020 WAR on 2019, 2018 and 2017 “steady” stats to then have a look at their explanatory value on performance variability (as in R²) and whether that value goes down the more we go back in time.

Therefore, I started by computing the correlation matrix for all batters who were both into the 2020 and 2019 threshold for PAs: to do that I first inner-joined the two datasets for 2020 and 2019 batters that met the PA quota for name and player id. That left a dataset of 276 batters, a decent enough sample size, for 23 statistics (team column was taken away). With the help of the *ggpubr* R package it is easy to build a correlation plot for the entire combined dataset on the form of a heatmap:

*Figure 1: Correlation Matrix for qualified batters, 2020 to 2019.*



I’d like to apologize for my laziness but I kept the usual color palette as I found it catchy although you may be used to the opposite: good stuff in red and bad stuff in blue, at least on a batter perspective, to indicate, instead of hot and cold zones, negative and positive correlations respectively.

Now it’s time for a little search and exploit: all I wanted to know resides on a single diagonal starting from the ISO20-ISO19 cell to the down-right until the WAR20-WAR19 dot. That long line of dots tells us how the 2020-2019 correlations between stats are, and as you can see there’s a lot of blue to be looked at! That is good news at a first glance: the same stats, for “qualified” batters, have a close and positive relationship on a single year scope. More into our previous distinction:

* Resume stats are the one with less y2y correlation among themselves: ISO stands atop, then there’s WAR while just weak positive correlation can be found for both wRC+ and wOBA, a surprising result at least to me, and for BABIP, much easier to expect given its nature;
* Profile stats, particularly Swing% and GB/FB ratio, present strong positive correlations, and so do some of the tendency stats, Contact% and SwStr% foremost, while batted ball direction and quality of contact ones show fewer satisfying results apart from Hard%;
* Statcast stats are solid across the board, with high positive correlations for all 4 of the stats with LA as the “steadiest” of the lotto.

That also means we can draw some first glance conclusions: to my utmost surprise batters are hard to change when it comes to swing choices and results. While I get that hitting the ball hard can’t be taught, I figured that modifying a swing’s path to create more lift, LA and flyballs was something doable nowadays and instead it seems that not only LA, but also EV, GB/FB ratio and HardHit/Barrels% are the toughest nuts to crack, showing some year to year strong relationships.

This means that for every Justin Turner or JD Martinez, who thrived after a complete retooling of their swing, there are hundreds of Yandi Diaz, guys who hit the ball hard but can’t lift it for the life of them (fun fact, Diaz had a 5+ GB/FB ratio in 2020 but positive resume stats because he literally scorches balls to the ground).

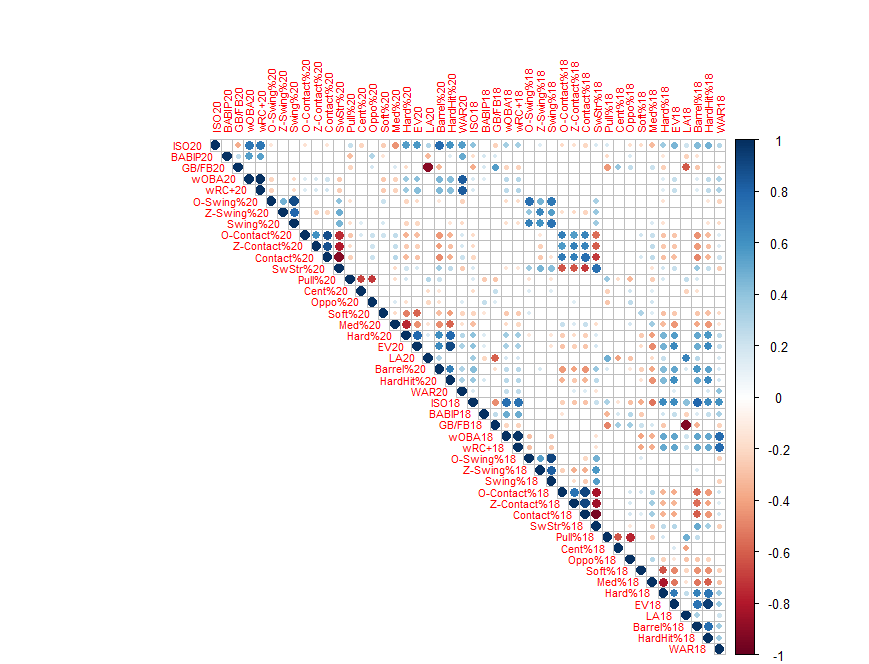
Moreover, swing decisions are also hard to modify it seems: they have some of the highest positive y2y correlations on the board, and so do the respective contact percentages, which tells us that it is not that easy to stop swinging at pitches outside the zone because you want to, and at the same time reassures us that hitting those pitches can’t be done out of sheer will, so yes, you can’t just become Vladimir Guerrero or Corey Dickerson on bouncing balls out of the blue.

On the other side, where and how hard you hit the ball seems to be quite fluctuating, unless you are a dead pull hard hitter, for that Hard% and Pull% pertain the highest positive correlation among quality of contact and batted ball direction respectively. That is both bad and good news: while these stats aren’t that solid, it also means that a batter that had a bad round could improve the following season only by getting some more punch into some Abs.

A preoccupying result is the scarce positive correlation between WAR in 2020 and 2019, although the caveat here is in the nature itself of WAR as pure counting stat and how it relates to a truncated season as in 2020: in a normal 162 games dance we probably would have seen some batters decline but also a lot of them go up to their standards (Christian Yelich, was that really you?).

All things considered we are left with a lot of possible *steady stats* in the sole year scope, but do they hold up if we go back in time? To do that I turned back the clock and performed the same process to get the correlation matrix in a 2-years scope (2020 to 2018) and 3-years scope (2020 to 2017) and checked whether there were any clear changes. To your eyes only:

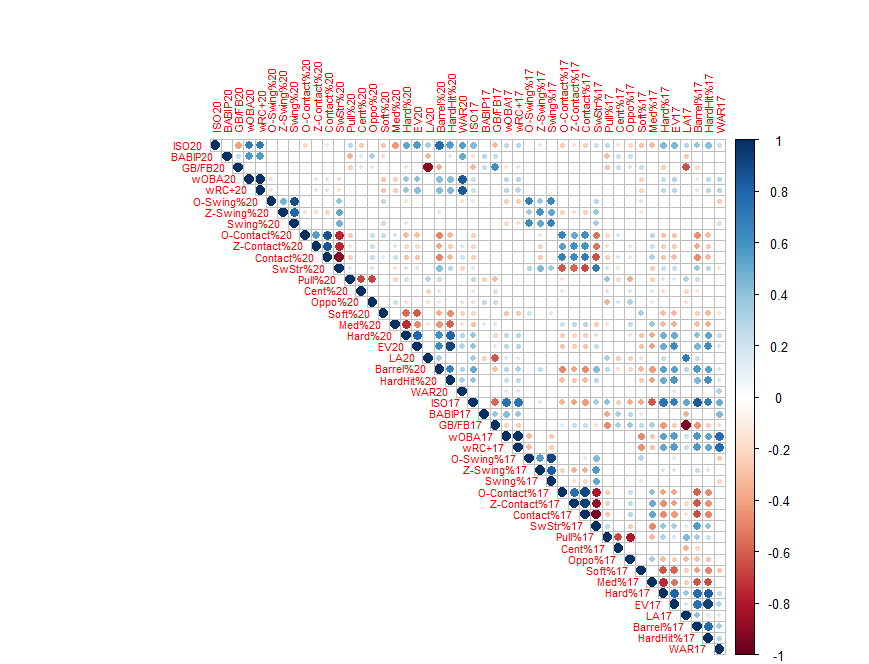
*Figure 2: Correlation Matrix for qualified batters, 2020 to 2018.*

**

As you can see there’s a subtle loss of positive correlation in some areas, particularly ISO, Pull%, Hard% and WAR, but the nature and strength of relationships seems to be in synch with that of the single season gap. Note that some of the change could be due to a loss of sample size to little more than 220 batters qualified according to set thresholds in both seasons.

Now yet another step back into baseball history:

*Figure 3: Correlation Matrix for qualified batters, 2020 to 2017.*

**

You need a proper eagle eye to see some feasible changes from the previous corrplot, and that is once more a confirmation that some among the chosen stats are positively correlated also when time gets in between.

To get a panoramic view of our stats in terms of gap years and correlations it’s easier to refer to something like this:

*Figure 4: Correlation table, 2020 to past seasons.*



As you can see, I’ve highlighted in yellow the “steady” correlation according to my personal 0.5 threshold, moreover the red signals that those correlations for Cent% are not statistically significant for the usual level of confidence with α = 0.05.

The general pattern is a loss of relationship strength as we go back in time, particularly for WAR and Swing% stats, while other stats retain solid and almost identical correlations even on a three-year gap.

Now it’s the hour of the truth: once the *steady stats* are in place, can they be important in a prospective analysis? Rather, can they tell us something about a player’s future performances? That would mean a lot, given that we are not considering either basic player statlines (H, HR, BB….) nor resume stats, as all of them don’t make it past the threshold. Roll the drums!

*Figure 5:* *linear regression model, 2020 WAR to 2019 steady stats*

Call:

lm(formula = war20i ~ gbfb19i + osw19i + zsw19i + sw19i + ocnt19i +

zcnt19i + cnt19i + swstr19i + hard19i + ev19i + barr19i +

hhit19i)

Residuals:

Min 1Q Median 3Q Max

-2.0576 -0.5985 -0.1039 0.5808 2.5083

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.369e+01 8.276e+00 -1.654 0.0992 .

gbfb19i -1.121e-01 1.365e-01 -0.821 0.4123

osw19i 1.978e+00 7.914e+00 0.250 0.8028

zsw19i 3.810e-01 4.031e+00 0.095 0.9248

sw19i -6.110e+00 1.192e+01 -0.512 0.6087

ocnt19i -5.685e+00 5.381e+00 -1.056 0.2918

zcnt19i -6.547e+00 8.636e+00 -0.758 0.4491

cnt19i 2.844e+01 1.744e+01 1.631 0.1042

swstr19i 2.460e+01 1.598e+01 1.539 0.1250

hard19i -7.568e-01 1.487e+00 -0.509 0.6112

ev19i -2.812e-04 6.924e-02 -0.004 0.9968

barr19i 3.451e+00 2.786e+00 1.239 0.2166

hhit19i 3.158e+00 2.034e+00 1.553 0.1217

---

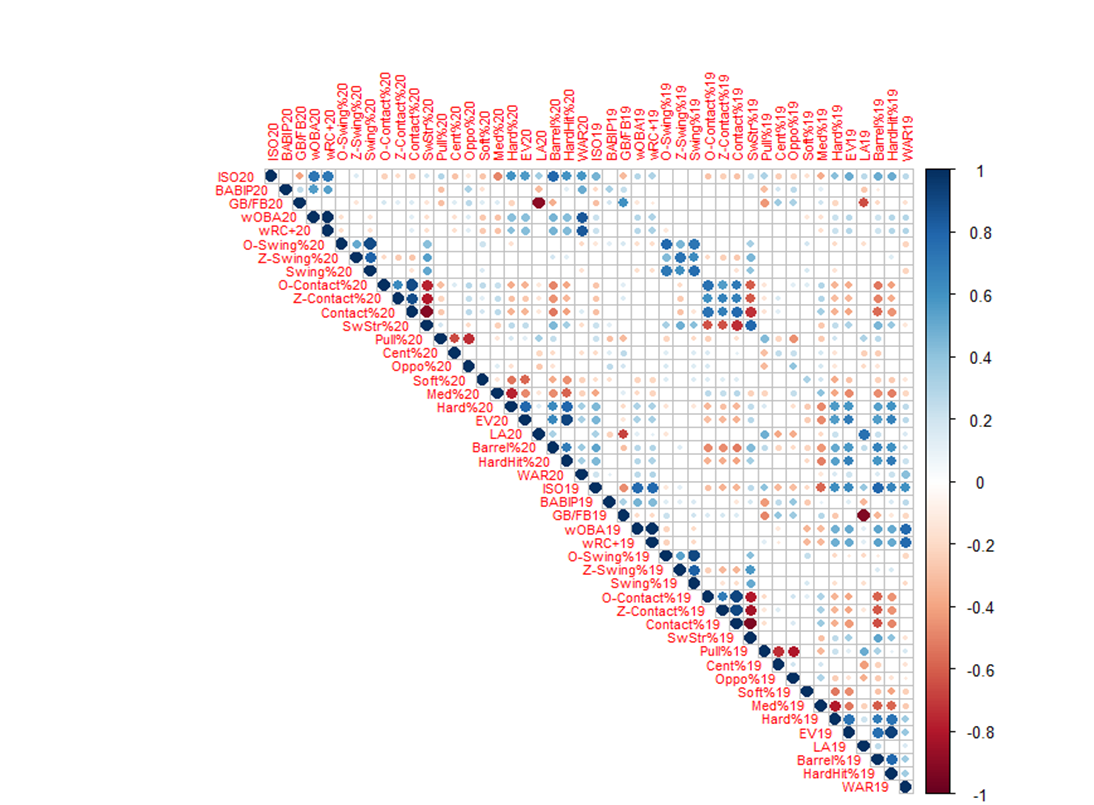
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8416 on 263 degrees of freedom

Multiple R-squared: 0.1043, Adjusted R-squared: 0.06339

F-statistic: 2.551 on 12 and 263 DF, p-value: 0.003318

And that is a big whiff! All our 2019 underlining stats are able to give us is just 10% of the overall 2020 WAR variability: without beating much around bushes I’ll call it a failure, as none of the indicators is statistically significant for any level of confidence, moreover explaining a tenth of something isn’t really something to be proud of. I’m making *mea culpa* for dragging this one out as someone could have got this result by simply relooking at the first corrplot:



If you start from the left-side WAR20 and go right till the end you can see the disaster coming: almost every correlation with 2019 stats is either pale white (≈ 0) or not that positively strong, with the sole WAR19 showing up but still less than my 0.5 target.

I’ll spend some words here on a crucial distinction that a lot of people seem to get mixed up: correlation and variability explanation are two different things altogether! The former expresses the strength and direction on the relationship between two variables on a [-1;1] range with the sign leaning towards inverse (-) or positive (+) proportionality expressed by a R figure; the latter instead shows how one or more independent variables are able to determine the variance of a dependent variable, on a 0-1 positive range defined by a R² measure that tells how much of the variability of the dependent is covered by a model.

Now you may think that there should be a link between them and you are spot on: if we take a 1-1 linear model where a single dependent is explained by a sole independent then the R² is simply the square of the R correlation measure for the two variables in question. Take WAR20 and WAR19: according to the table their correlation (R) is 0.42, and by squaring it up we get that WAR19 has a ≈ 0.18 R² for WAR 2020, that is to say that the previous year WAR explains 18% of the 2020 WAR variability.

This is telling us something both important and discouraging: looking at what a player has done in years past is not enough to assess what he’s going to provide in performance and value to the team, and that shouldn’t be surprised as there are a million reasons someone could underperform his past (age and injuries, bad luck coming or good luck fading away) or overcome his worst memories (the god of baseball looking down, a swing change or newfound health).

That’s also why projecting a player’s future career and worth is actual hell on earth: working on something you know has a serious chance of being dead wrong is a job for few brilliant minds, the ones able to develop models and algorithms that minimize a shot at embarrassment. I’ll pour one to you, Dan Szymborski, you ZiPS genius!

It’s not even worth the effort to employ the model on 2018 and 2017 indicators and 2020 WAR dependent as the results are bound to be even worse than the one closer in time, so is it over just like that? Yes, in the predictive sense, but not really if we take another avenue: what about the explanatory power of those *steady stats*? In layman’s terms, how do those 2020 stats contribute into the same-year WAR variability?

Go back to the aforementioned left-side WAR20 dot but this time instead of going right, aim to the top and follow the line: those are all the correlations between WAR20 and each other stat in our bucket from the same year (but only on hitter that qualified both in 2020 and 2019, so we are even missing some players).

You can easily see that the *Statcastats* are the ones with a clearer and stronger positive correlation with WAR among underlining stats, while good results were to be expected from both wOBA and wRC+, that are directly into the WAR formula or a product of one of its parts. Lastly, ISO shows some good correlation with WAR, a sign that power never gets old, and it’s the most important mean of batting production in this era of baseball.

First let’s redo the model as before but on 2020 stats explaining 2020 WAR:

*Figure 6: linear regression model, 2020 WAR to 2020 steady stats*

Call:

lm(formula = war20 ~ gbfb20 + osw20 + zsw20 + sw20 + swstr20 +

ocnt20 + zcnt20 + cnt20 + hard20 + ev20 + la20 + barr20 +

hhit20)

Residuals:

Min 1Q Median 3Q Max

-1.9436 -0.4644 -0.0480 0.4443 1.8816

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -9.31208 4.58241 -2.032 0.04292 \*

gbfb20 -0.10172 0.14821 -0.686 0.49297

osw20 7.84457 5.33452 1.471 0.14235

zsw20 6.69658 2.50417 2.674 0.00786 \*\*

sw20 -18.60740 7.95107 -2.340 0.01985 \*

swstr20 18.64918 10.82261 1.723 0.08577 .

ocnt20 -4.46721 3.52691 -1.267 0.20617

zcnt20 -7.73757 5.69953 -1.358 0.17550

cnt20 26.74133 11.52360 2.321 0.02090 \*

hard20 0.95871 0.91723 1.045 0.29667

ev20 -0.03146 0.03508 -0.897 0.37052

la20 -0.01727 0.01606 -1.075 0.28296

barr20 8.16086 1.37264 5.945 6.88e-09 \*\*\*

hhit20 1.89261 1.12033 1.689 0.09208 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6965 on 338 degrees of freedom

Multiple R-squared: 0.342, Adjusted R-squared: 0.3167

F-statistic: 13.51 on 13 and 338 DF, p-value: < 2.2e-16

Nothing spectacular, but it’s much better than the predictive case! The R² sits at a decent 0.34, as to say that the “pure underlining” model explains 34% of the 2020 WAR variability, not a bad result if we consider that none of the usual counting stats, from hits to homers, has been taken into the equation.

The real problem is the absence of statistical significance for almost all the variables, a lot due to correlation among themselves I feel like, yet there’s one of them that seems most solid and shines on a minuscule p-value: Barrel%.

As I’ve already written, Barrel% is the percentage of balls squared up such that results are at their peak in terms of average and power, and per definition it encompasses a certain number of EV + LA combinations. But if barreling a ball is the best-case scenario for a batter, that must be somehow connected to his worth, right??

That made me look at the 2020-2019 correlation table once again: Barrel% is obviously positively and strongly correlated to both HardHit% and EV, but it also goes hand in hand with almost all the resume stats, showing a bond with wRC+, wOBA but mostly ISO. This should surprise no one: barreling a ball means sending one to oblivion and, if you are a man of barrels, you’re bound to have a high SLG, which helps given ISO = SLG – AVG! But this is big news: ISO is the steadiest stat on the resume group, it connects with wRC+ and wOBA strongly and therefore to WAR, our player’s value proxy.

I then built a sort of chain, from solid underlining stats to the top of the food chain and looked how it holds up when compared to the full model. Note that decisions are taken given the correlation matrix table (look for the blue!) but always considering the logic behind each step and relationship. The dataset in this case is the original one for all hitters that qualified in the sole scope of the 2020 season given the 80 PA self-made threshold, different from the one correlations are calculated on.

The first element is not Barrel% itself, rather HardHit%: almost all the EV+LA combinations needed to produce a barrel require an EV over 95 mph, such that it enters the scope of the HardHit definition. On a down to ground 1-1 model:

*Figure 7: linear regression model, 2020 HardHit% to 2020 EV*

Call:

lm(formula = hhit20 ~ ev20)

Residuals:

Min 1Q Median 3Q Max

-0.13400 -0.02440 0.00008 0.02408 0.11779

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.3713695 0.0710199 -33.39 <2e-16 \*\*\*

ev20 0.0310532 0.0008025 38.70 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03885 on 350 degrees of freedom

Multiple R-squared: 0.8106, Adjusted R-squared: 0.81

F-statistic: 1498 on 1 and 350 DF, p-value: < 2.2e-16

Average exit velocity explains 81% of HardHit%’s variability: the harder you tend to hit the ball the higher your EV is and the chances to get more Hard Hits™, this is child’s play!

Then how about our Barrels now?

*Figure 8: linear regression model, 2020 Barrel% to 2020 HardHit%, LA and SwStr%*

Call:

lm(formula = barr20 ~ hhit20 + la20 + swstr20)

Residuals:

Min 1Q Median 3Q Max

-0.090779 -0.018796 -0.001334 0.019705 0.091797

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.1008875 0.0078836 -12.797 < 2e-16 \*\*\*

hhit20 0.3157586 0.0179481 17.593 < 2e-16 \*\*\*

la20 0.0023160 0.0002817 8.222 4.01e-15 \*\*\*

swstr20 0.2650712 0.0454116 5.837 1.22e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02851 on 348 degrees of freedom

Multiple R-squared: 0.6082, Adjusted R-squared: 0.6048

F-statistic: 180.1 on 3 and 348 DF, p-value: < 2.2e-16

Before you get all angry and flustered, I’m not saying that swinging and missing a lot is good! This is more likely a consequence of power-oriented approaches at the plate: if you want to murder the ball at every swing, you’re bound to miss more balls, that is the old discipline vs damage payoff. As previewed, both HardHit% and LA, and that strange SwStr%, are statistically significant and convey 60% of the overall Barrel% variability.

Watch out to how much explanatory power we are losing at each step: that is why simply relying on underlining stats is a dollar short and a second late, as on our way to the WAR pinnacle a lot is lost in terms of variability we can assert!

Now the key step on our ascent to WAR: from Barrel%, an underlining stat, to ISO, a resume one, how much is going to be lost R²-wise and is there another stats that could help reinforce the validity and strength of our chain?

*Figure 9: linear regression model, 2020 ISO to 2020 Barrel% and Pull%*

Call:

lm(formula = iso20 ~ barr20 + pull20)

Residuals:

Min 1Q Median 3Q Max

-0.138750 -0.027609 -0.002957 0.026450 0.146133

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.03761 0.01322 2.845 0.00470 \*\*

barr20 1.21008 0.05213 23.213 < 2e-16 \*\*\*

pull20 0.10194 0.03270 3.117 0.00198 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04292 on 349 degrees of freedom

Multiple R-squared: 0.6406, Adjusted R-squared: 0.6386

F-statistic: 311.1 on 2 and 349 DF, p-value: < 2.2e-16

An almost 65% explanatory power is nothing to sneeze at, and to reinforce our lines came a somewhat surprising ally: Pull%. While its estimated coefficient is minuscule, the rationale behind a statistical significance stems on the current trend in baseball of “elevating and celebrating”, focused on hitting the ball hard in the air in order to maximize the damage.

In this aspect pulling the ball is much more favorable than going oppo, unless you have prodigious power the other way as a Juan Soto or else you play in a favorable park to go oppo taco, like Yankee Stadium.

That said we finally got into the well-known and hard-boiled resume stats, knocking at WAR’s doors, but before that we need to pass the wRC+ hurdle:

*Figure 10: linear regression model, 2020 wRC+ to 2020 ISO and BABIP*

Call:

lm(formula = wRC20 ~ iso20 + babip20)

Residuals:

Min 1Q Median 3Q Max

-44.560 -9.463 1.119 9.372 51.418

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -55.574 4.965 -11.19 <2e-16 \*\*\*

iso20 339.439 11.828 28.70 <2e-16 \*\*\*

babip20 332.391 15.415 21.56 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.81 on 349 degrees of freedom

Multiple R-squared: 0.7919, Adjusted R-squared: 0.7907

F-statistic: 664.1 on 2 and 349 DF, p-value: < 2.2e-16

This is the quirkiest step in our journey, as it includes dealing with the baseball edition of fate: BABIP. Although we introduced its definition, Batting Average for Balls In Play, and formula, this strange construct is mostly known as the luck factor, a Final Fantasy-like attribute that tells whether a batter is hitting it where they are or ain’t, but is it?

Yes, and no! Firstly, there’s no set standard for luck, although a .300 BABIP sometimes is taken as benchmark, so simply comparing a player’s BABIP with a stone-cold threshold is not useful. What BABIP tells us is twofold: how a batter is running in terms of finding holes with his hits, the dumb luck, but also what kind of balls he’s hitting. A flyball hitter or a TTO, three true outcomes (HR,BB,K), in the Joey Gallo mold will always have a BABIP hovering even under the .250 zone, but that doesn’t mean he’s being struck by a curse!

The key factor is that each kind of hit has its own expected results: as we know hitting a barrel is *chef’s kiss* for output purposesbut also harder than almost anything else, after that a player has a greater shot at hits and damage with line drives, groundballs and flyballs in this order. Note that hitting a ball in the air is the worst among all possible kinds of hit in this macro-groups: that is because you can’t use your speed to get on base, just gain some, while burners can pile up infield hits, but also because well…that’s how stadiums are built! Outfields in general are cavernous with a lot of ground that defenders can cover with the extra time that the flight of the ball gives them to run their routes.

So, while hitting the ball high is appreciated, it also comes with a strong requirement: power. There’s no prize for warning track flyouts! Therefore it’s “easier” to sneak a measly grounder through the hole, beat a shift with a bunt (????) or outrun a throw than find real estate 200+ feet from home plate, and that explains BABIP differentials among batters.

But then how can BABIP help us? We need to consider it into a bigger range, sample size and number of ABs : this stat needs a lot of stabilizing as it is not strongly correlated y2y so a good catch is to compare real-time BABIP for a player with his career or three year average to only then call an eventual deviation from that average a bad roll or hot streak.

There’s also another consideration: what is the weight of BABIP alone in explaining wRC+ variability? Is luck really a big factor? As it turns out it amounted for 30% of the total in 2020, while the other 50% was provided by ISO: almost half of the total explanatory power of the model comes from a pseudo-luck measure?! Something is fishy, and, spoiler alert, it has to do with this particular (and truncated…hint!) season.

Last but not least our final destination, WAR:

*Figure 11: linear regression model, 2020 WAR to 2020 wRC+*

Call:

lm(formula = war20 ~ wRC20)

Residuals:

Min 1Q Median 3Q Max

-1.16638 -0.32606 -0.00625 0.29507 1.53471

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.371526 0.079317 -17.29 <2e-16 \*\*\*

wRC20 0.019973 0.000747 26.74 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4837 on 350 degrees of freedom

Multiple R-squared: 0.6713, Adjusted R-squared: 0.6704

F-statistic: 714.9 on 1 and 350 DF, p-value: < 2.2e-16

Almost 70% of the WAR variability is explained by wRC+: it makes sense as the latter derives from wOBA, the main component of the Batting Runs part of WAR, so it is a good result coming with the limitation that it is solely related to batting and not the other aspects of the game that influence the WAR figure, such as baserunning, defense and other adjustments.

If we employ some BadMath and consider the percentage of WAR variability explained by starting from the first link of our chain by simply computing the product of each R² start to finish, ignoring all possible correlations for variables at different steps, confounding and not considered ones also (but hey, BadMath rules!) we get that ≈ 17% of the WAR variability for qualified (>= 80 PAs) batters in 2020 is explained by a [EV, LA, SwStr, Pull%, BABIP] package and its combinations, but what about if we start straight from Barrel%?

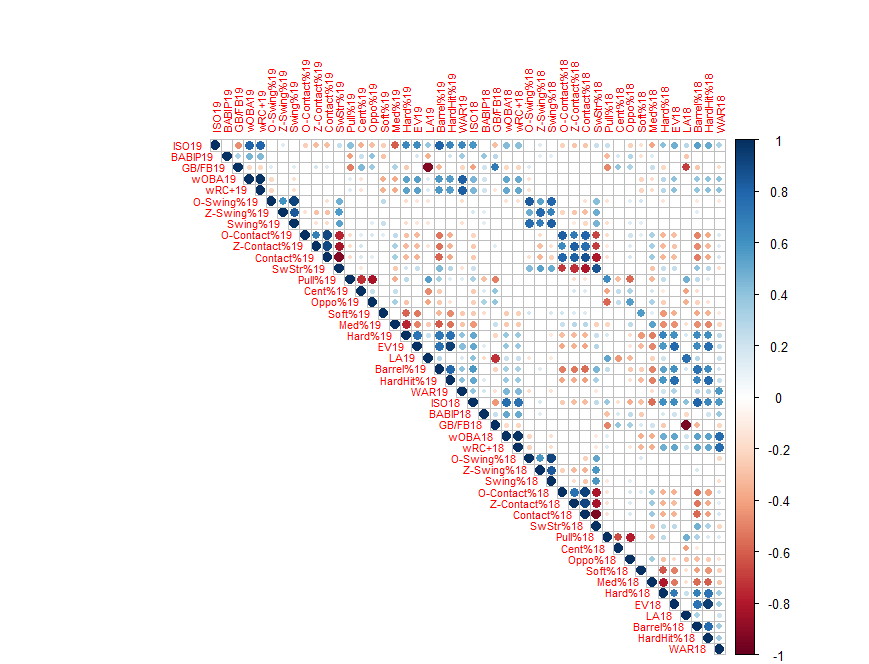
The advantage of this measure is that it stabilizes, along with others, in a relatively low number of events such that after just a slice of a season we can have a decent idea of a WAR range for batters given their actual Barrel%. By skipping stones we find out that, if we perfectly knew each batter’s Barrel% for a year, and some other stats, we could explain ≈34% of the WAR variability…exactly what the starting explanatory model told us, so to get the same R² power we need “only” 100% certified Barrel%, Pull rates and BABIP instead of 20+ stats, quite the family discount!

One final question before wrapping it up: did the distinctive nature of the 2020 season, with less than half of the standard 162 games and all other strange settings pandemic-related, impacted these results?

To check for that I considered the last two complete seasons, 2019 and 2018, and started with the correlation matrix on the inner-joined combined dataset for qualified batters (>= 200 PAs in both seasons) to see if there’s anything different, then proceeded to evaluate both predictive and explanatory cases for the 2019 season and finally compared its findings with those of the 2020 crown season.

Step 1 is the corrplot in all its shades of blue and red:

*Figure 12: Correlation Matrix for qualified batters, 2019 to 2018*

**

You’ll be happy to see that there’s not much that changed with respect to the 2020 season, so we weren’t missing the point entirely. The eye test tells us that relationships between variables are all in the same directions with some different strengths, but nothing stands out.

If we look at the diagonal starting from the dot ISO19-ISO18 down to right until the WAR cell we can clearly see that the *steady stats* in 2020 are still there while some of the others joined the party, among them almost all the resume stats, the quality of contact stats and the Oppo%. This upgrade on steadiness is surely due to more than a single factor, although the fact that two full 162 games season are compared with the same PA threshold for batters instead of a mixed bag as with 2020 could be the main reason things are much more blue than before.

To get a better grip of these improvements let’s add a column to the correlation table:

*Figure 13: Correlation table, 2020 to past seasons with 2019-18 correlations*



A yellow brick road! Every stat in our bucket, apart from Cent% which is not that useful and BABIP as “luck” is somewhat fleeting, is steady if taken the 0.5 correlation limit: this is good news because it also means that a better predictive model can be constructed!

*Figure 14: linear regression model, 2019 WAR to 2018 steady stats*

Call:

lm(formula = war19c ~ wrc18c + iso18c + gbfb18c + osw18c + zsw18c +

sw18c + swstr18c + ocnt18c + zcnt18c + cnt18c + pull18c +

oppo18c + soft18c + med18c + hard18c + ev18c + la18c + barr18c +

hhit18c)

Residuals:

Min 1Q Median 3Q Max

-4.7207 -1.2165 -0.1517 1.1007 5.7554

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.918e+02 2.295e+02 1.707 0.0891 .

wrc18c 1.854e-02 8.464e-03 2.191 0.0294 \*

iso18c 7.392e+00 5.245e+00 1.409 0.1600

gbfb18c -1.276e+00 8.735e-01 -1.461 0.1452

osw18c -8.487e+00 2.085e+01 -0.407 0.6844

zsw18c -7.023e+00 1.125e+01 -0.624 0.5330

sw18c 1.579e+01 3.274e+01 0.482 0.6301

swstr18c -3.124e+00 3.595e+01 -0.087 0.9308

ocnt18c -4.212e+00 1.310e+01 -0.322 0.7481

zcnt18c -1.273e+01 2.168e+01 -0.587 0.5577

cnt18c 1.875e+01 4.278e+01 0.438 0.6615

pull18c -1.580e+00 3.741e+00 -0.422 0.6732

oppo18c 3.176e-01 4.550e+00 0.070 0.9444

soft18c -3.980e+02 2.290e+02 -1.738 0.0835 .

med18c -4.034e+02 2.288e+02 -1.763 0.0791 .

hard18c -4.054e+02 2.286e+02 -1.773 0.0775 .

ev18c 1.400e-01 1.768e-01 0.792 0.4290

la18c -9.259e-02 8.775e-02 -1.055 0.2924

barr18c 1.868e+00 7.453e+00 0.251 0.8023

hhit18c 8.402e-01 5.203e+00 0.161 0.8719

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.764 on 242 degrees of freedom

Multiple R-squared: 0.2695, Adjusted R-squared: 0.2122

F-statistic: 4.7 on 19 and 242 DF, p-value: 3.251e-09

I mean, it’s “better” but still underwhelming: the predictive ability of the model soars to 27% of the overall WAR variability although almost all the stats end up being statistically not significant, due to correlations between them and whatnot. Still a 3x improvement is something to consider, so dealing with complete seasons and comparable dataset per thresholds gives more consistency to a rather fleeting model in terms of results and strength.

What about the explanatory counterpart? Will there be better results with respect to the 2020 model?

*Figure 15: linear regression model, 2019 WAR to 2019 steady stats*

Call:

lm(formula = war19 ~ wRC19 + iso19 + gbfb19 + osw19 + zsw19 +

sw19 + swstr19 + ocnt19 + zcnt19 + cnt19 + soft19 + med19 +

hard19 + pull19 + oppo19 + ev19 + la19 + barr19 + hhit19)

Residuals:

Min 1Q Median 3Q Max

-2.2114 -0.7207 -0.1272 0.5989 3.5882

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -79.221544 117.779016 -0.673 0.5016

wRC19 0.058567 0.004771 12.275 <2e-16 \*\*\*

iso19 1.206863 2.758874 0.437 0.6621

gbfb19 0.060995 0.377794 0.161 0.8718

osw19 -11.655712 8.926525 -1.306 0.1925

zsw19 -3.688638 4.665350 -0.791 0.4297

sw19 9.768103 13.541901 0.721 0.4712

swstr19 16.871400 17.073148 0.988 0.3238

ocnt19 6.473940 5.844867 1.108 0.2688

zcnt19 12.371178 9.268374 1.335 0.1828

cnt19 -8.924219 18.581416 -0.480 0.6313

soft19 55.799548 117.317274 0.476 0.6346

med19 55.199066 117.300231 0.471 0.6382

hard19 55.543840 117.390964 0.473 0.6364

pull19 0.834836 1.921270 0.435 0.6642

oppo19 -1.067716 2.551436 -0.418 0.6759

ev19 0.139612 0.075194 1.857 0.0642 .

la19 0.026314 0.037473 0.702 0.4830

barr19 -4.488659 3.784482 -1.186 0.2364

hhit19 -3.500955 2.232135 -1.568 0.1177

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.09 on 340 degrees of freedom

Multiple R-squared: 0.6805, Adjusted R-squared: 0.6626

F-statistic: 38.11 on 19 and 340 DF, p-value: < 2.2e-16

Having either one between wOBA and wRC+ as steady is a boon as both are close to WAR as a construct, so the 2x improvement in WAR variability’s explanatory power of the model has a lot to do with that. Note that while in the predictive model we considered the inner-joined dataset for batters meeting the 200 PA threshold in both seasons, here we just take the 2019 dataset for “qualified batters” no matter if they were also in the 2018 one.

While wRC+ seems to be the only solid stat, there’s a little dot of significance for EV, which means almost nothing but if you remember EV itself was the starting point for the chain-like model that went all the way from the underlining stats to WAR. So, it’s BadMath time once again!

*Figure 16: linear regression model, 2019 HardHit% to 2019 EV*

Call:

lm(formula = hhit19 ~ ev19)

Residuals:

Min 1Q Median 3Q Max

-0.134457 -0.016998 0.001084 0.016704 0.115744

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.4923846 0.0636671 -39.15 <2e-16 \*\*\*

ev19 0.0322487 0.0007166 45.00 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02871 on 358 degrees of freedom

Multiple R-squared: 0.8498, Adjusted R-squared: 0.8493

F-statistic: 2025 on 1 and 358 DF, p-value: < 2.2e-16

Slightly better, a +5% in R² from the 2020 edition. Now onto the Barrel%:

*Figure 17: linear regression model, 2019 Barrel% to 2019 HardHit%, LA and SwStr%*

Call:

lm(formula = barr19 ~ hhit19 + la19 + swstr19)

Residuals:

Min 1Q Median 3Q Max

-0.047518 -0.013015 -0.001352 0.012206 0.092803

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.1247994 0.0067918 -18.375 < 2e-16 \*\*\*

hhit19 0.3701905 0.0159094 23.269 < 2e-16 \*\*\*

la19 0.0020631 0.0002643 7.805 6.69e-14 \*\*\*

swstr19 0.2847086 0.0360163 7.905 3.39e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02119 on 356 degrees of freedom

Multiple R-squared: 0.7144, Adjusted R-squared: 0.712

F-statistic: 296.9 on 3 and 356 DF, p-value: < 2.2e-16

Now we are talking business! A +10% R² is quite the gain, so that we are in a good position to move forward.

*Figure 18: linear regression model, 2019 ISO to 2019 Barrel% and Pull%*

Call:

lm(formula = iso19 ~ barr19 + pull19)

Residuals:

Min 1Q Median 3Q Max

-0.129874 -0.022800 -0.001063 0.022256 0.128548

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.03805 0.01277 2.980 0.00308 \*\*

barr19 1.11846 0.04873 22.952 < 2e-16 \*\*\*

pull19 0.17207 0.03239 5.313 1.9e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03464 on 357 degrees of freedom

Multiple R-squared: 0.6619, Adjusted R-squared: 0.66

F-statistic: 349.4 on 2 and 357 DF, p-value: < 2.2e-16

One step forward, two steps back! We are now at almost the same situation as in 2020, with a small 2% improvement. Time to bring luck into the equation:

*Figure 19: linear regression model, 2019 wRC+ to 2019 ISO and BABIP*

Call:

lm(formula = wRC19 ~ iso19 + babip19)

Residuals:

Min 1Q Median 3Q Max

-43.246 -7.624 -0.722 8.216 37.098

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -42.929 5.408 -7.938 2.68e-14 \*\*\*

iso19 327.144 10.991 29.764 < 2e-16 \*\*\*

babip19 274.009 17.115 16.010 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.32 on 357 degrees of freedom

Multiple R-squared: 0.776, Adjusted R-squared: 0.7747

F-statistic: 618.3 on 2 and 357 DF, p-value: < 2.2e-16

Well, that was unexpected! On what should be a stronger dataset the variability of wRC+ explained by ISO and BABIP goes a smidge down! Maybe this is the real amount that the model can predict while the 2020 scenario was affected by its shortened nature.

What is clear as the skies in a beautiful summer afternoon is the role of BABIP, the Luck factor, this time around: in itself the stat explains ≈ 22% of the R², while the remaining 55% is due to sheer ISO (and a small amount of BABIP-ISO correlation that we do not consider). This is a big drop from the 30% that BABIP was worth in 2020 R²-wise and for a good reason: in dealing with resume stats we are considering measures that stabilize much later than the underlining ones, so that with a bigger PA threshold, and a complete season, ISO became stronger while the effect of BABIP went down accordingly.

That doesn’t mean that luck has no effect in the scope of a full season and that BABIP is just a proxy devoid of meaning: while for a lot of batters a high BABIP could mean catching breaks left and right, others are routinely able to post .340+ BABIP seasons and if probabilities are a thing, it’s extremely tough to be lucky year in and year out! That’s why single season BABIP is a good measure only if compared with a career or span average: sometimes a man brings his own luck!

To end all things, we meet WAR again:

*Figure 20: linear regression model, 2019 WAR to 2019 wRC+*

Call:

lm(formula = war19 ~ wRC19)

Residuals:

Min 1Q Median 3Q Max

-2.3841 -0.7710 -0.1094 0.5742 3.8235

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.268302 0.231997 -18.40 <2e-16 \*\*\*

wRC19 0.058721 0.002227 26.37 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.095 on 358 degrees of freedom

Multiple R-squared: 0.6602, Adjusted R-squared: 0.6592

F-statistic: 695.5 on 1 and 358 DF, p-value: < 2.2e-16

There’s almost no difference with 2020, and that is a little surprising: while wRC+ is a pure batting stat, WAR also consider the other facets of the game, defense and value on the bases alike.

On a whim, there could be at least a pair of explanations: one is that other means of value have lower variance in the entire season’s scope than batting, so that for example there are not a lot of great or awful defenders and a bunch of average ones, causing almost no explanatory power loss except for the extremes; the other reason could be that defensive and baserunning indicators are fast to stabilize so that a low number of events, those happening in the span of 80 PAs, is enough to have a clear view of their progression and project accordingly with some confidence.

A last BadMath consideration: the R² of WAR explained in 2019 starting from EV, on the [EV, LA, SwStr, Pull%, BABIP] package and its combinations, is almost 21%, a small raise from 2020, while by starting with perfect Barrel knowledge on the sole [Barrel%, Pull%, BABIP] discount combo we arrive at the same 34% we had in 2020, so no major changes on this funny and flawed escape.

What have we learned then?

1. Almost all the underlining stats, from Swing and Contact percentages to quality of contact and batted ball direction metrics, are steady in the sense they are positively and strongly (R > 0.5) correlated y2y but…;
2. The same can’t be said for resume stats, that are steady only over a greater number of PAs or not even then in the BABIP case. That said…;
3. Underlining stats are not enough to have confident and significant predictive value on a future player’s performances, stopping at a maximum 10% of WAR R² in a pure underlining model….
4. But let’s not throw everything away! Underlining stats are much better at explaining rather than predicting, for both resume stats as wRC+ and ISO, and attain more than 30% of the WAR total variability in 2020 and 2019 which is relevant as…
5. The truncated nature of the 2020 season was crucial for the solidity of resume stats and WAR itself, not to mention the different impact of BABIP on shorter samples, although in the end…;
6. Predicting or explaining a player’s performance with a certain degree of trust in your results is something that requires models much more complicated that a linear regression, where cold results are considered in an age/career perspective in their development or decline and unforeseen events are also in the mix, such as injuries or swing revolutions.

What is then the value of all this analysis? While not being able to assert about a player’s worth in a predictive sense sucks, having an idea of the correlations existing between underlining and resume stats, such as the one between WAR and Barrel% on a same-season consideration, can help us to identify players who have been somewhat strange and that have some value to be found.

Think about a player with a sky-high EV, good HardHit% but low Barrel%: that is a candidate for an infusion of LA! Or what about someone with good EV, a pull heavy approach and a decent LA but low ISO: maybe he was unlucky or bound by his stadium!

In this mold I’ll try to identify a player or two that have some weird shenanigans going on between underlining stats and results, why they are there and if there’s a shot at something better than it seems.

Until then, barrel as much as you can!