Expected Hits vs. Actual Hits 2022 Season

Harry Svendsgaard

January 14, 2023

1 Introduction

With the 2022 season concluded, now is the best time to take a look at players from the previous year that may have over or under performed. It is easy to look at how a player has performed over the course of their career and highlight outlier seasons where they failed to meet high expectations, or exceeded a low bar that was set for them. Something like this can be impacted by many factors, such as progression/regression, injuries, etc. In order to mitigate some of those factors, I chose to look at performance in the 2022 season in relation to expected performance. In this article, I will explain the process for building a model that finds expected hits in 2022 for each player, and analyze the results of that model. By the end of the article, it will be clear which players under or over performed at the plate this past season. I will also highlight numbers for the young players in the league and define what this means early on in their career.

The data was downloaded from Baseball Savant's website, and includes all batters with 100 or more plate appearances from the previous season. Using 100 as the baseline is probably the lowest to go when working with MLB data. I wanted to include players that may have dealt with injuries this season, or players that could have bounced between the Majors and Minors.

2 Building a Model

The first thing to do was build a model that predicts the amount of hits for all players from a variety of variables using Statcast data. I knew the number of plate appearances was the one variable that needed to be included, but beyond that, the remaining impactful variables were found through stepwise regression in R. The resulting model predicts hits from the number of plate appearances, average launch angle, and the percentage of balls hit in the sweet spot of the bat. This linear model resulted in an R-squared value of 0.9541. When building a model, the goal is to try to achieve an R-squared value that is as close to 1 as possible. That being said, it is pretty evident that the stepwise regression resulted in a very strong model.

The next step was to use this model to generate expected hit predictions for every player in the data set. This was done by utilizing the *predict()* function in R, where mod is the model created, and stats is the data set being used.

```
xHits <- predict(mod, stats)
```

I will plot the generated values against the true base hit value for each player to visualize the correlation between the two.
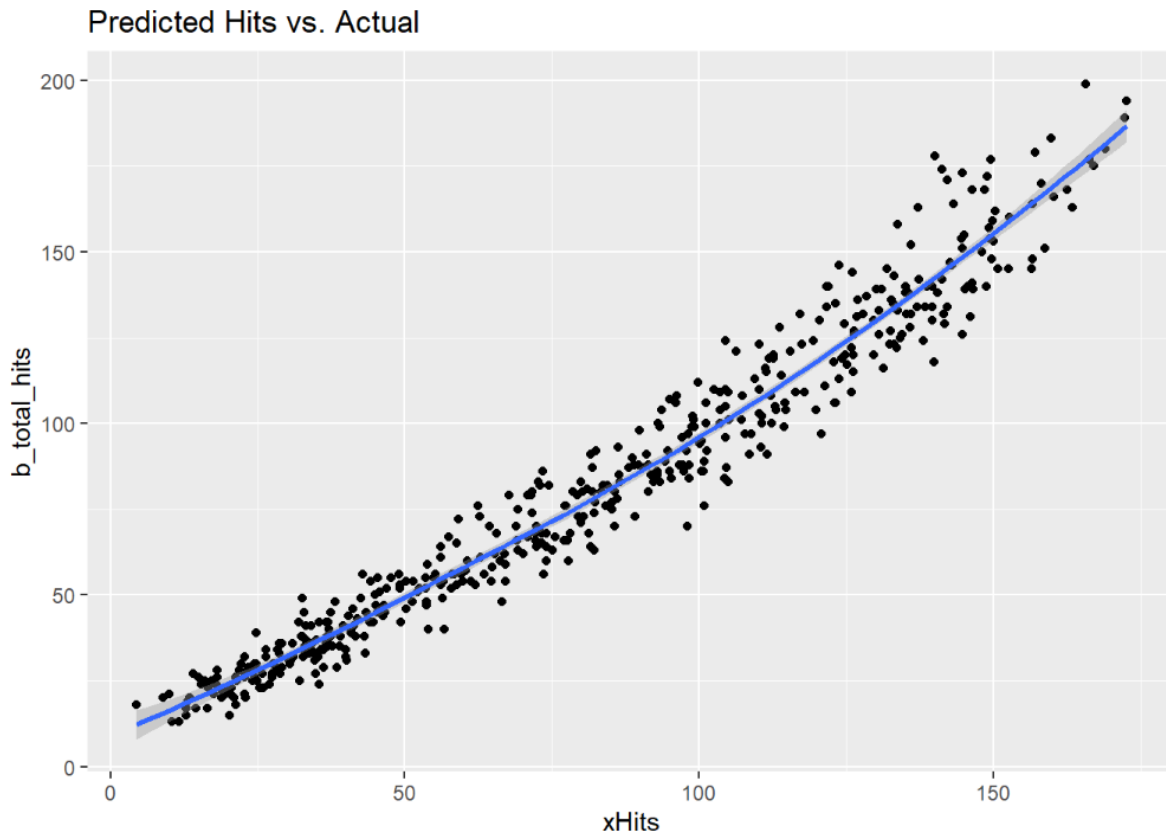
Figure 1: Predicted Hits vs. Actual Hits 2022

We see that there are no true outliers, and the majority of the points are fairly close to the line from the model. That being said, we will take a look at the more extreme cases, where expected hits and actual hits varies the most.

Before doing this, it is necessary to accommodate for the fact that the number of at-bats for each player can vary drastically, and this can impact the data. For example, a player with 500 at-bats has a lot higher chance of having a larger number than someone with 100 at-bats when it comes to the absolute value of $xHits - actual hits$. In order to get rid of this problem, I will create a new variable, which divides xHits – actual hits by total number of at-bats. This leaves a difference of expected hits versus actual hits, per at-bat.

This results in values that range from approximately -0.1 to 0.1, which can be a little painful to look at and manage. That being said, one final step will be taken in order to optimize this data, and it will be to normalize these values. All values will be normalized on a 0 to 1 scale. The player with the highest value will be assigned the new value *1*, and the player with the lowest value will be assigned *0*.

3 Analysis

Let's take a look at the top ten players from the league in this statistic. These players, according to the model, deserved the highest percentage of hits per at-bat, as compared to their actual statistics.

Top 10 Players in (Expected Hits - Hits) per AB

Last	First	Age	AB	PA	H	BA	xHits	xHits - H	(xHits - H)/AB	Normalized	xBA	BAdiff
Piscotty	Stephen	31	126	139	24	0.190	35.40176	11.401762	0.0904902	1.0000000	0.2809664	0.0909664
Lowe	Josh	24	181	198	40	0.221	53.98333	13.983334	0.0772560	0.9392974	0.2982505	0.0772505
Pache	Cristian	23	241	260	40	0.166	56.76864	16.768643	0.0695794	0.9040865	0.2355545	0.0695545
Walls	Taylor	25	407	466	70	0.172	98.01319	28.013191	0.0688285	0.9006420	0.2408187	0.0688187
Calhoun	Kole	34	388	424	76	0.196	100.83963	24.839628	0.0640197	0.8785849	0.2598959	0.0638959
Hedges	Austin	29	294	338	48	0.163	66.47674	18.476739	0.0628461	0.8732018	0.2261114	0.0631114
McKinstry	Zach	27	166	185	33	0.199	43.25282	10.252821	0.0617640	0.8682385	0.2605592	0.0615592
Velazquez	Andrew	27	322	349	63	0.196	82.14452	19.144522	0.0594550	0.8576479	0.2551072	0.0591072
Tsutsugo	Yoshi	30	170	193	29	0.171	38.47872	9.478725	0.0557572	0.8406866	0.2263454	0.0553454
Stassi	Max	31	333	375	60	0.180	77.77254	17.772536	0.0533710	0.8297415	0.2335512	0.0535512

Figure 2: Top 10 Players in $|xHits - truehits|$

Stephen Piscotty, a veteran in the league, sits atop this list. Piscotty only recorded 24 hits in his limited plate appearances in 2022, but was expected about 35. This is nearly a 50% increase from his actual statistics. On paper, Piscotty had a shameful year, batting only .190. However, if he had reached his expected number of hits, he would have hit about .281. Josh Lowe is another name that pops out on this list. He is the only player in this top ten list that batted above the .200 mark. Taking into account his expected hits, Lowe would have hit nearly .300 this past season. Lowe did not have a lot of at-bats, only appearing in 52 games, but he was clearly hitting the ball well in these contests.

Now, let's take a look at the bottom ten players in this statistic. These players got more hits this year than they were expected to get, based on the model.

Bottom 10 Players in (Expected Hits - Hits) per AB

Last	First	Age	AB	PA	H	BA	xHits	xHits - H	(xHits - H)/AB	Normalized	xBA	BAdiff
Moniak	Mickey	24	106	112	18	0.170	4.482195	-13.517805	-0.1275265	0.0000000	0.0422849	-0.1277151
Call	Alex	27	114	131	27	0.237	14.079636	-12.920364	-0.1133365	0.0650865	0.1235056	-0.1134944
Carpenter	Matt	36	128	154	39	0.305	24.748451	-14.251549	-0.1113402	0.0742431	0.1933473	-0.1116527
Trammell	Taylor	24	102	117	20	0.196	8.886956	-11.113044	-0.1089514	0.0852002	0.0871270	-0.1088730
White	Eli	28	105	117	21	0.200	9.956256	-11.043744	-0.1051785	0.1025057	0.0948215	-0.1051785
Bryant	Kris	30	160	181	49	0.306	32.574922	-16.425078	-0.1026567	0.1140726	0.2035933	-0.1024067
Palacios	Richie	25	112	123	26	0.232	14.984555	-11.015445	-0.0983522	0.1338167	0.1337907	-0.0982093
Stubbs	Garrett	29	106	121	28	0.264	18.145568	-9.854432	-0.0929663	0.1585206	0.1711846	-0.0928154
Fairchild	Stuart	26	97	110	24	0.247	15.375618	-8.624382	-0.0889112	0.1771209	0.1585115	-0.0884885
Mercado	Oscar	27	121	128	25	0.207	15.989390	-9.010610	-0.0744679	0.2433695	0.1321437	-0.0748563

Figure 3: Bottom 10 Players in $|xHits - truehits|$

Matt Carpenter and Kris Bryant are the names that pop out the most from this list, as they both batted over .300. Although there could be a vast array of reasons for this, I have come up with a few theories as to why these two names in particular are on the list. For Carpenter, I would assume that he generated a few hits against the shift, and potentially a few home run balls in Yankee Stadium that would be fly outs in other ballparks. Yankee Stadium is known for it's short left and right field, so as a major pull hitter, Carpenter is able to benefit from this. In regards to Bryant, there are also a few factors to consider. Firstly, being a Rockie typically leads to slightly inflated batting numbers due to the elevation. It is easier to hit home runs in Coor's Park than in any other stadium. Also, we can take a look at Bryant's spray chart of hits using Baseball Savant's website.
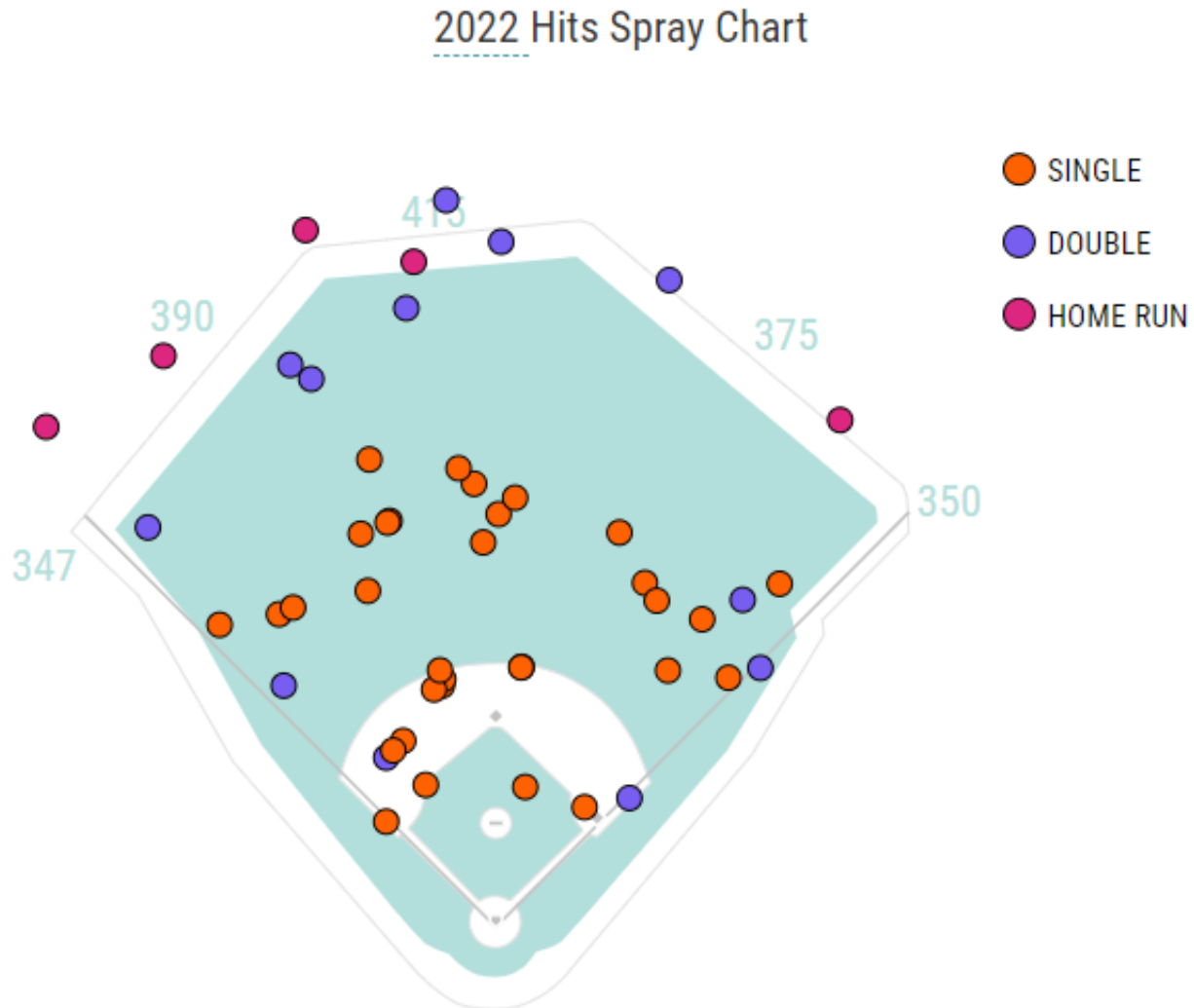

Figure 4:   Kris Bryant Hit Chart 2022

This reveals that a lot of Kris Bryant's hits this year came from the infield or approximately 250 feet or less. Compared to previous years, it appears that Bryant was hitting the ball with a lot less power this year. This can be backed up with the fact that he only hit five home runs in 160 at-bats; the lowest rate of his career,

with exemption to the 2020 season, where he only appeared in 34 games.

Having looked at the top ten and bottom ten players in the league, I decided it would be interesting to look at the same thing for players under 25 years old. It is currently the offseason, and teams are looking to sign players that can become impactful for their organization, or simply give more playing time to guys that are currently on their rosters. Using the following tables will help reveal which players to look out for in the years to come.

First, we'll look at the top ten young players from 2022.

Top 10 Players Under 25 in (Expected Hits - Hits) per AB

Last	First	Age	AB	PA	H	BA	xHits	xHits - H	(xHits - H)/AB	Normalized	xBA	BAdiff
Lowe	Josh	24	181	198	40	0.221	53.98333	13.983334	0.0772560	0.9392974	0.2982505	0.0772505
Pache	Cristian	23	241	260	40	0.166	56.76864	16.768643	0.0695794	0.9040865	0.2355545	0.0695545
Walls	Taylor	25	407	466	70	0.172	98.01319	28.013191	0.0688285	0.9006420	0.2408187	0.0688187
Grisham	Trent	25	451	524	83	0.184	104.87225	21.872246	0.0484972	0.8073865	0.2325327	0.0485327
Perdomo	Geraldo	22	431	500	84	0.195	104.24001	20.240015	0.0469606	0.8003383	0.2418562	0.0468562
Torkelson	Spencer	22	360	404	73	0.203	89.03617	16.036168	0.0445449	0.7892580	0.2473227	0.0443227
Toro	Abraham	25	324	352	60	0.185	74.03054	14.030539	0.0433041	0.7835668	0.2284893	0.0434893
Barrero	Jose	24	165	174	25	0.152	32.07882	7.078825	0.0429020	0.7817221	0.1944171	0.0424171
Castillo	Diego	24	262	283	54	0.206	64.70374	10.703739	0.0408540	0.7723283	0.2469608	0.0409608
Gorman	Nolan	22	283	313	64	0.226	74.05377	10.053769	0.0355257	0.7478886	0.2616741	0.0356741

Figure 5: Top 10 Players Under 25 Years Old in $|xHits - truehits|$

The biggest name to look out for on this list is Nolan Gorman. Gorman was a rookie in 2022, and only appeared in 89 contests for the Cardinals. On this list, he has the highest actual batting average, compared to some guys that have been in the league for a couple years now. Gorman was expected approximately 74 hits, about 10 more than his actual total. Taking that into account, his expected batting average is .262, which is quite impressive for someone who is currently known as a power hitter. Gorman will look to get his strikeout rate down in 2023, which will ultimately lead to the batting average and total hits continuing to rise.

Spencer Torkelson is another name to look out for on this list. Torkelson was also a rookie in 2022. Although he only batted a shade over .200, his expected batting average was near .250, due to him generating about 16 more expected hits. Torkelson, like Gorman, needs to get his strikeout rate down in order to become an impactful 3-5 hole hitter in the MLB, but him being slightly above average in this area is not cause for concern, considering it was only his rookie season.

Let's take a look at the bottom ten young players when it comes to expected hits versus actual hits per at-bat. These players may be off to a weaker start to their careers than their hitting splits may suggest.

Bottom 10 Players Under 25 in (Expected Hits - Hits) per AB

Last	First	Age	AB	PA	H	BA	xHits	xHits - H	(xHits - H)/AB	Normalized	xBA	BAdiff
Moniak	Mickey	24	106	112	18	0.170	4.482195	-13.517805	-0.1275265	0.0000000	0.0422849	-0.1277151
Trammell	Taylor	24	102	117	20	0.196	8.886956	-11.113044	-0.1089514	0.0852002	0.0871270	-0.1088730
Palacios	Richie	25	112	123	26	0.232	14.984555	-11.015445	-0.0983522	0.1338167	0.1337907	-0.0982093
Thompson	Bubba	24	170	181	45	0.265	32.831102	-12.168898	-0.0715818	0.2566075	0.1931241	-0.0718759
Steer	Spencer	24	95	108	20	0.211	13.412355	-6.587645	-0.0693436	0.2668734	0.1411827	-0.0698173
Burdick	Peyton	25	92	102	19	0.207	13.176604	-5.823397	-0.0632978	0.2946045	0.1432240	-0.0637760
Toglia	Michael	23	111	120	24	0.216	17.119112	-6.880888	-0.0619900	0.3006031	0.1542262	-0.0617738
Pasquantino	Vinnie	24	258	298	76	0.295	62.445580	-13.554420	-0.0525365	0.3439643	0.2420371	-0.0529629
Arraez	Luis	25	547	603	173	0.316	144.753321	-28.246679	-0.0516393	0.3480798	0.2646313	-0.0513687
Garcia	Dermis	24	116	125	24	0.207	18.051868	-5.948132	-0.0512770	0.3497415	0.1556195	-0.0513805

Figure 6: Bottom 10 Players Under 25 Years Old in $|xHits - truehits|$

Mickey Moniak was on the list above, but it is more appropriate to talk about him with the young players. The former number 1 overall pick has struggled to make an impact in the league, since being drafted in 2016, and this past season was no different. Though limited in plate appearances, Moniak was not able to generate many hits this season, tallying only 18 for the year. That being said, according to the model, Moniak was only expected about 4.5 hits for the entire season. Realistically, it would not be that low, but Moniak's inconsistency in hitting the ball in sweet spot is cause for concern. A once hopeful superstar has a lot of work to do to get up to the level scouts projected him to be.

Another name on this list to speak on is Luis Arraez. Arraez led the American League in batting average in 2022, and prevented Aaron Judge from being the second player to win the Triple Crown since 1967. According to the model, Arraez was expected about 28 less hits than he actually tallied. Taking this into account, his batting average would have bumped well below the .300 mark and Judge would be the first Triple Crown winner since Miguel Cabrera in 2012.

There are a variety of reasons for why some players may have inflated or deflated expected hits when comparing to their actual hits. The research and analysis for this is certainly not done here. An important next step would be to look at data from the 2021 season and see how players compare from one year to the next. This could bring light to many factors that could impact a player's down -or up- year. For example, maybe a player got hurt in 2022 and was not 100% healthy when playing. Maybe a player got traded after the 2021 season to a more favorable hitter's stadium. It would also be ideal to look at the young player's growth curves, and see how they have improved or worsened as the years have gone on. This will give insight into career projections for players that may not be proven yet.

Thank you for reading, please feel free to leave a comment with your thoughts on the topic!

May 24 Recap

I felt it was appropriate to add this to the end here, seeing as I wrote this article well before the 2023 season began. Now that we're about 50 games into the season, I decided to look at some of the players that appeared on the top 10 and bottom 10 list, and see how they are performing thus far. Below is the top 10 players under the age of 25 from the previous season, in respect to $xHits - actualHits$ per AB. The table examines their batting average and expected batting average from last year, and it compares it to their true batting average this season.

Player	2022 PA's	2022 Batting Avg	2022 xBA	2023 PA's	2023 Batting Avg
Josh Lowe	198	0.221	0.298	152	0.307
Cristian Pache	260	0.166	0.236	26	0.360
Taylor Walls	466	0.172	0.241	129	0.261
Trent Grisham	524	0.184	0.233	175	0.196
Geraldo Perdomo	500	0.195	0.242	130	0.315
Spencer Torkelson	404	0.203	0.247	190	0.229
Abraham Toro	352	0.185	0.228	170 (AAA)	0.274 (AAA)
Jose Barrero	174	0.152	0.194	118	0.231
Diego Castillo	283	0.206	0.247	169 (AAA)	0.329 (AAA)
Nolan Gorman	313	0.226	0.262	171	0.297

Despite Abraham Toro and Diego Castillo playing AAA this season, there have been significant increases in the batting average of every hitter on this list. Cristian Pache has failed to have many plate appearances this season. Taking Pache, Toro, and Castillo out of the equation, the other batters are hitting on average 0.069 above their batting average last season.

Now to look at the bottom 10 players from the previous season.

Player	2022 PA's	2022 Batting Avg	2022 xBA	2023 PA's	2023 Batting Avg
Mickey Moniak	112	0.170	0.042	32	0.419
Taylor Trammell	117	0.196	0.087	41	0.147
Richie Palacios	123	0.232	0.138	203 (AAA)	0.247 (AAA)
Bubba Thompson	181	0.265	0.193	56	0.163
Spencer Steer	108	0.211	0.141	192	0.257
Peyton Burdick	102	0.207	0.143	37	0.182
Michael Toglia	120	0.216	0.154	21	0.150
Vinnie Pasquantino	298	0.295	0.242	209	0.251
Luis Arraez	603	0.316	0.265	184	0.371
Dermis Garcia	125	0.207	0.156	138 (AAA)	0.248 (AAA)

This table tells me a couple of different things. First, it tells me that the predictions for the top 10 young players was more accurate. I still do think there is some accuracy in these results, but it is hard to determine when a player is not getting many plate appearances. In the cases of Spencer Steer and Luis Arraez, they have a lot of plate appearances, and have exceeded their numbers from last year. I think with young players, the statistical increase can be accredited to general progression from one season to the next. Typically, when a player is under the age of 25, you should expect them to be better the next season. This table also tells me that a higher parameter should be set on number of plate appearances for young players that are in the bottom tier in this statistic. With young players that are struggling early in their career, it's pretty typical for them to bounce between the Majors and Minors during the season. This will result in fewer plate appearances for the player, and it will be hard to gauge how they are truly progressing.

All this being said, I think it would be more beneficial to look at a few of the players that made the bottom 10 list which was not exclusive to the players 25 or under. Below are the players from the original bottom 10 list, that were not also featured on the young player's list.

Player	2022 PA's	2022 Batting Avg	2022 xBA	2023 PA's	2023 Batting Avg
Alex Call	131	0.237	0.124	189	0.218
Matt Carpenter	154	0.305	0.193	119	0.198
Eli White	117	0.200	0.095	17	0.071
Kris Bryant	181	0.306	0.204	191	0.281
Garrett Stubbs	121	0.264	0.171	33	0.226
Stuart Fairchild	110	0.247	0.159	114	0.232
Oscar Mercado	128	0.207	0.132	20	0.450

All of these players have seen decreases in batting average since last year, except for Oscar Mercado, who only has 20 plate appearances on the season. I think this list shows that there is accuracy in estimating both the top tier, and the bottom tier when it comes to estimated hit versus true hit difference.

I look forward to doing future analysis on this matter. It would be interesting to predict players future season output based on these metrics. Clearly this metric has worked - with pretty solid accuracy - in predicting this season's performance for individual players. Going season by season to detect trends for these players could be a next step in providing insightful information.

Thank you again for reading. Please feel free to leave a comment and let me know your thoughts on how this project came out.