Predicting HR rates using probit models

According to research done by Russell Carleton, HR rates stabilize for pitchers over a long period of time – over approximately 1320 batters. Of course, this is assuming that pitchers stay perfectly consistent over this time period – a condition rarely met in the real world.

I decided to take advantage of the pitch-by-pitch data and create an overarching model for predicting HRs, by pitch type and pitching hand, using the pitch indicators as predictors. My hope is that this type of model will be able to detect aberrant HR rates (for better or for worse) at a faster rate than using HR count data. In other words, when a pitcher gives up extra HRs, is there something in his mechanics which tips this off, or is he just getting unlucky? I'd like my model to be an added tool in answering this type of question.

Probit models predicting HR rates, by pitch type and pitcher handedness

Pitch type	FF (Four seamer)		FT (Two seamer)		SL (Slider)		CH (Change up)		SI (Sinker)		CU (Curveball)	
Pitching hand	R	L	R	L	R	L	R	L	R	L	R	L
# of observations	255694	91376	84520	40962	113404	40584	74682	39192	62725	23319	50132	18580
Pseudo R^2	.0261	.0311	.0467	.0497	.0687	.0539	.0747	.0835	.0528	.0521	.0832	.0816
releasevelocity	-5.90***	-4.56***	-3.19**		-8.74***	-3.88***	-4.62***	-3.14**	-2.39***		-2.38***	-2.82**
spinrate		-2.74**	-3.14**	-2.77**	-1.65~						-2.26*	
spinrate_rhb	-2.88**		-1.68~		1.76~				-2.92**			-1.92~
spindir			3.52***	2.90**			1.82~					
spindir_rhb	-2.93**	-3.81***	-3.42**	-4.30***					-1.94~	-2.56*	-1.76~	
px		2.32*			1.93~					-3.15***		
px_rhb	-19.71***	-8.69***	-10.89***	-5.07***	-18.85***	-8.18***	-13.68***	-5.98***	-12.52***	-5.54***	-13.70***	-6.95***
pz		3.90***	2.95**	1.66~					2.59**		2.86**	
szt	6.09***	3.06**	2.75**	2.94**	4.09***	1.77~	4.03***	1.75~	2.58**		3.63***	3.22**
szb								1.69~				-2.78**
xo		-1.83~			-1.71~		1.86~			3.14**		
xo_rhb		1.81~					1.77~				-2.92**	
zo	-2.04*	-4.07***	-2.81**				2.39*		-2.45*			
vxo		-2.01*					2.37*			3.16**		
vyo	-5.56***	-4.05***	-2.88**		-8.77***	-3.83***	-4.80***	-3.09**	-2.04*		-2.18*	-2.86**
vzo	-2.10*	-4.35***	-2.76**				2.34*		-2.43*			
ax					-1.74~		1.95~			2.97**		
ay		2.87**			-1.94~		-3.91***					-3.29**
az		-2.47*					3.11**		-1.76~		-1.87~	
rh_batter	4.04***	2.73**	3.17**	3.42**					2.09*	2.14*		
timesfaced	2.50*				3.63***				1.77~	2.23*		
outs	-3.21**	-3.10**	-2.69**	-2.62**					-2.30*			

***p<.001
**p<.01
*p<.05
~p<.1

The table lists results from 12 separate probit models – a separate model for each pitch listed, divided by pitching hand. The data was limited to swinging strikes, foul balls, and balls in play – I wanted to focus on the pitches where the batter made an attempt, successfully or not, to put the ball in play.

I coded the outcome as a 1 for a HR, and a 0 for everything else. Probit coefficients are tricky to interpret directly, so I instead have the corresponding z-values listed. I have only included z-values for coefficients that are at least marginally significant (at the 10% level).

All of the variables used come directly from the pitch data. I added interaction terms for batter handedness with the horizontal pitch data. There were additional controls added to the model that are omitted in the table above in the interest of space (i.e. pitch count data). And I used a bootstrapping procedure to give the models better external validity.

While the pseudo R^2 values are low, the sheer size of the database allows for some strong associations to be made. Not surprisingly, across all pitch types, faster pitches (releasevelocity) are associated with lower rates of HRs allowed. Pitching inside (px_rhb) is also strongly associated with HR prevention across all pitch types, although the strength of this association is stronger for same-handed matchups (righty-righty, lefty-lefty).

Through these models, I can now calculate an expected HR outcome probability per pitch type. I will use this data in three different ways in the examples that follow with Zack Greinke, James Shields, and Brandon McCarthy.

Zack Greinke

# of HRs allowed by pitch type

pitchType	Freq.	Percent	Cum.
CH	4	8.70	8.70
CU	8	17.39	26.09
FC	2	4.35	30.43
FF	20	43.48	73.91
FT	6	13.04	86.96
SL	6	13.04	100.00
Total	46	100.00	

Like many pitchers, Greinke is victimized by the long ball off his four-seam fb – the four seamer is responsible for 20 of the 46 HRs he has given up over the past three seasons. Given Greinke's sparkling low HR rate in 2015, we might expect that he found a way to minimize the damage off his four-seamer. The data shows otherwise, however.

# of HRs allowed by pitch type, 2013-14 (top) and 2015 (bottom)

pitchType	Freq.	Percent	Cum.
CH	4	12.50	12.50
CU	7	21.88	34.38
FC	2	6.25	40.63
FF	11	34.38	75.00
FT	3	9.38	84.38
SL	5	15.63	100.00
Total	32	100.00	

pitchType	Freq.	Percent	Cum.
CU	1	7.14	7.14
FF	9	64.29	71.43
FT	3	21.43	92.86
SL	1	7.14	100.00
Total	14	100.00	

The top table lists the pitch type for the HRs given up in the 2013 and 2014 seasons. The bottom table has the same data for the 2015 season. Surprisingly, Greinke was much more vulnerable in 2015 with his four-seamer than in the previous two seasons combined. Was this a one-off fluke? To examine, I checked the actual HR rates by pitch type to the projected HR rates using the probit models described earlier.

Actual and Predicted HR rates, by pitch type, 2013-14 (bottom) and 2015 (top)

	FF	FT	CH	CU	FC	SL
2015						
HR	.0144231	.0188679	0	.0126582	0	.0028011
Pred_HR	.0151916	.0138887	.0080299	.0133052	0	.0086172
Count	624	159	332	79	0	357
2013-14						
HR	.0109562	.006383	.0082474	.0230263	.0079681	.015015
Pred_HR	.0161125	.0155582	.0096369	.0149782	.0124844	.0108223
Count	1004	470	485	304	251	333

As seen in the data, Greinke changed his mix of pitches in 2015. He eliminated the cutter entirely, and upped his reliance on his change-up, four seamer, and especially his slider. While his four seamer was hit more often for a HR in 2015, this increased rate was more than negated by the unbelievable rate with which his change-up and slider prevented HRs (1 HR total from 689 pitches). While the data does suggest that Greinke's change-up and slider should have been more effective at preventing HRs in 2015, his actual HR rate off these pitches is remarkable and most likely unsustainable moving forward.

The data suggests Greinke was a bit lucky with HRs allowed in 2015. The expected outcome was ~18.48 HRs, or a HR/9 ratio of ~0.75 – right around his historical average. And while it's clear that Greinke's slider is exceptional at preventing HRs, there is a high correlation between slider usage and trips to the DL – he may need to rely less on the slider moving forward if he has his long-term health in mind.

James Shields

Up until the 2015 season, James Shields was the model of consistency. This consistency convinced the Royals to trade for Shields in the Wil Myers deal that was widely panned at the time (before it became known that Wade Davis would turn into Mariano 2.0). This consistency convinced the Padres to reward Shields with a $75 million contract, despite the risk that at 33 he was approaching the downslope of his career.

Unfortunately for the Padres, there is some evidence that suggests that Shields regressed in 2015. While his BABIP stayed constant and his K/9-rate shot up to the highest of his career, his walk and HR rates reached new highs as well. The bump in HRs to a 1.47 HR/9 rate (and a 1.74 HR/9 rate at Petco!) was well above Shields' average and entirely unexpected, and was primarily responsible for his disappointing season.

# of HRs allowed by pitch type, 2013-14 (top) and 2015 (bottom)

pitchType	Freq.	Percent	Cum.
CH	11	25.58	25.58
FC	7	16.28	41.86
FF	17	39.53	81.40
FT	4	9.30	90.70
KC	4	9.30	100.00
Total	43	100.00	

pitchType	Freq.	Percent	Cum.
CH	7	21.21	21.21
FC	6	18.18	39.39
FF	9	27.27	66.67
FT	6	18.18	84.85
KC	5	15.15	100.00
Total	33	100.00	

The top table lists Shields' pitch type for the HRs he gave up in the 2013 and 2014 seasons. The bottom table has the same data for the 2015 season. I examined the expected outcome of each type of pitch, with the exception of the knuckle curve (due to a lack of data across MLB).

Actual and Predicted HR rates, by pitch type, 2013-14 (bottom) and 2015 (top)

	FF	FT	CH	FC
2015				
HR	.027439	.0271493	.0185676	.0212766
Pred_HR	.0193573	.0156174	.0124805	.0165828
Count	328	221	377	282
2013-14				
HR	.0186199	.0113636	.0082474	.0088384
Pred_HR	.0184913	.0154521	.0126129	.0158632
Count	913	352	983	792

The data suggests that Shields' HR woes in 2015 might be due to some bad luck. While he lost a bit of velocity on his FB, the predicted HR values suggest that the effect was negligible. Excluding the knuckle curve, Shields gave up 28 HRs in 2015, while the predicted data gives an estimate of ~19.2 HRs – much closer to his historical average. If this data is to be believed, James Shields may be due for a bounce-back season in 2016.

Brandon McCarthy

Finally healthy, Brandon McCarthy seemingly enjoyed a breakout season in 2014. While his numbers on the surface were not flashy, he had career highs in IPs and K/9s, and a career low 2.87 xFIP rate suggested that better days were to come. The Dodgers took a gamble with a $48 million dollar contract.

Hindsight is 20/20, but we now know that the extra 2 mph that McCarthy added in velocity in the 2014 season may have contributed to his UCL injury in early 2015. Can we use the HR data to look for potential warning signs?

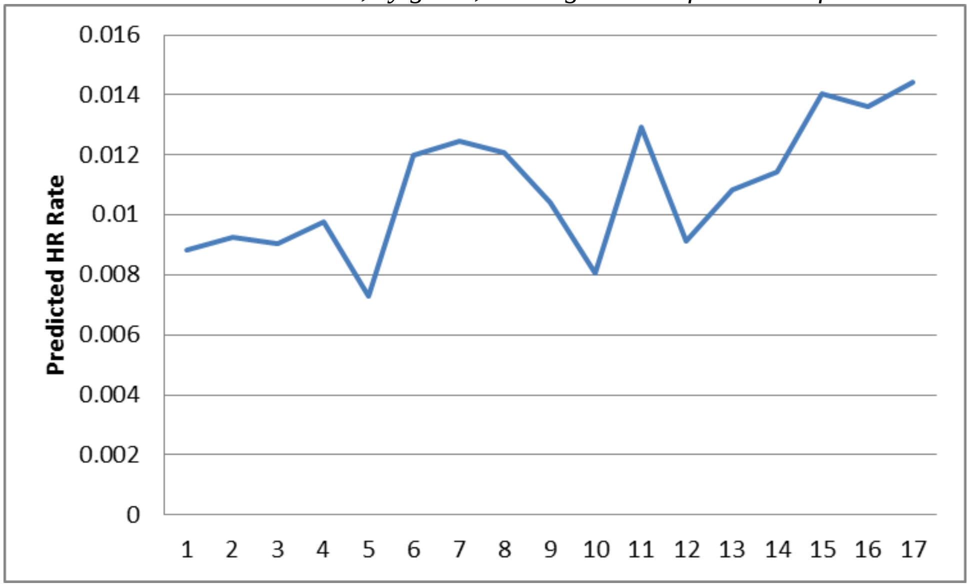
Predicted HR rates, by pitch type, listed sequentially per game starting with the first start after the 2014 ASB

	SI	FF	FC	CU
1	0.008833	0.012302	0.013538	0.007411
2	0.009254	0.009406	0.012018	0.00821
3	0.009033	0.014428	0.017556	0.006245
4	0.009765	0.011623	0.011062	0.00507
5	0.00731	0.011935	0.015357	0.005196
6	0.011997	0.010916	0.011039	0.009613
7	0.012465	0.012141	0.011387	0.010102
8	0.012093	0.010101	0.016612	0.012856
9	0.010431	0.012919	0.011141	0.008768
10	0.00806	0.012643	0.011401	0.010694
11	0.01291	0.009772	0.013373	0.010776
12	0.009147	0.012699	0.008616	0.009389
13	0.010858	0.011694	0.020618	0.008392
14	0.011448	0.012167	0.008097	0.006974
15	0.014058	0.009708	0.015044	0.014162
16	0.013622	0.009474	0.007994	0.013817
17	0.014439	0.013643	0.026003	0.012823

Above is a table with average predicted HR rate per pitch thrown in McCarthy's 17 starts after the 2014 ASB, listed sequentially. Starts 14-17 occurred in 2015.

McCarthy threw the sinker ~50% of the time in the 2014-15 seasons, so the first column (SI) may be most instructive. Notably, his sinker appears to trend up in home run tendencies before he hits noticeable highs starting with the second start of the 2015 season. The trend line is charted on the next page.

Predicted HR rates on Sinkers, by game, starting with the first start after the 2014 ASB



I do not know if UCL injuries build over time or are the result of freak accidents, but to the extent that any sort of warning signs may exist for such injuries, this example suggests that a close examination of the pitch data may give a team some advance warning.

**To the judges – Thank you for your time. This is the first time I have done a deep data dive into MLB pitch data - any and all feedback is very welcome and appreciated!

- Anthony Kim