Predicting Home Run Hitters

Renee Yaldoo

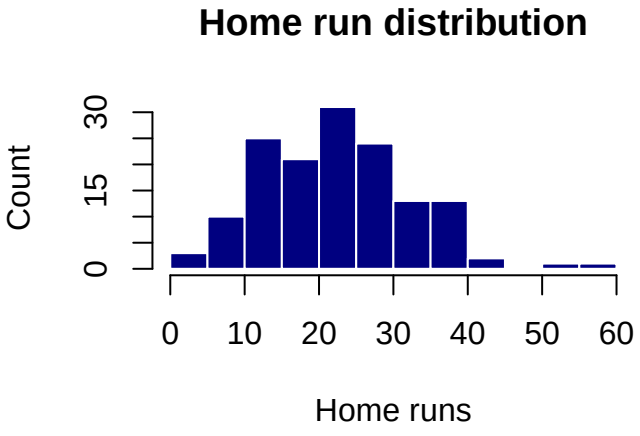
12/9/2017

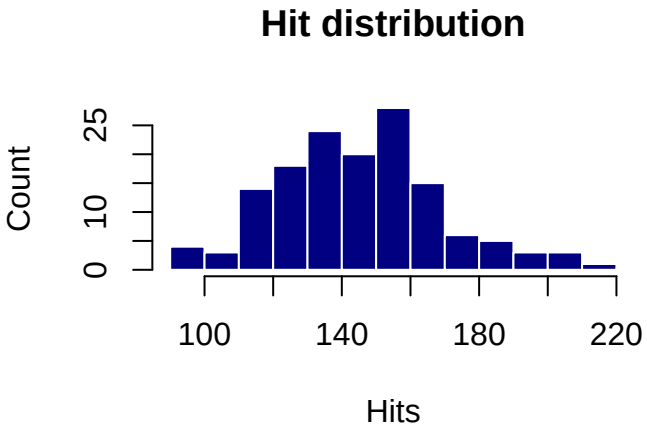
Introduction:

Baseball, a very popular sport and American pastime has dazzled and delighted fans since its first inception in 1839. While some fans live for those summer days on the diamond, others flock from all over to visit those famous stadiums. What drives those cheering fans wild? The home run hit! Players and fans alike have their superstitions for why certain players knock it out of the park, but can we determine which home run predictors produce the Babe Ruth of sluggers? My research question is: What is the best statistical model to determine which home run predictors ensures a player a home run hit in a season? To answer this question, we will build a multivariate linear model and choose the best model based on the following model criteria: R-Squared, R-Squared Adjusted, AIC, BIC, Press, Mallow's Cp, and cross-validation. My hypothesis is that RBI's (Runs Batted In) most accurately predicts the number of home runs a player will hit in a season. In addition to RBIs, the dataset include five other independent variables that will be included in the analysis. These are: batting average (AVG), hits (H), intentional walk (IBB), runs batted in (RBI), total bases (TB), pitches per plate appearance (P/PA). The dataset comes from ESPN's MLB Player Batting Stats, 2017. I have only used all the ranked players who are within the top 144 players of the dataset.

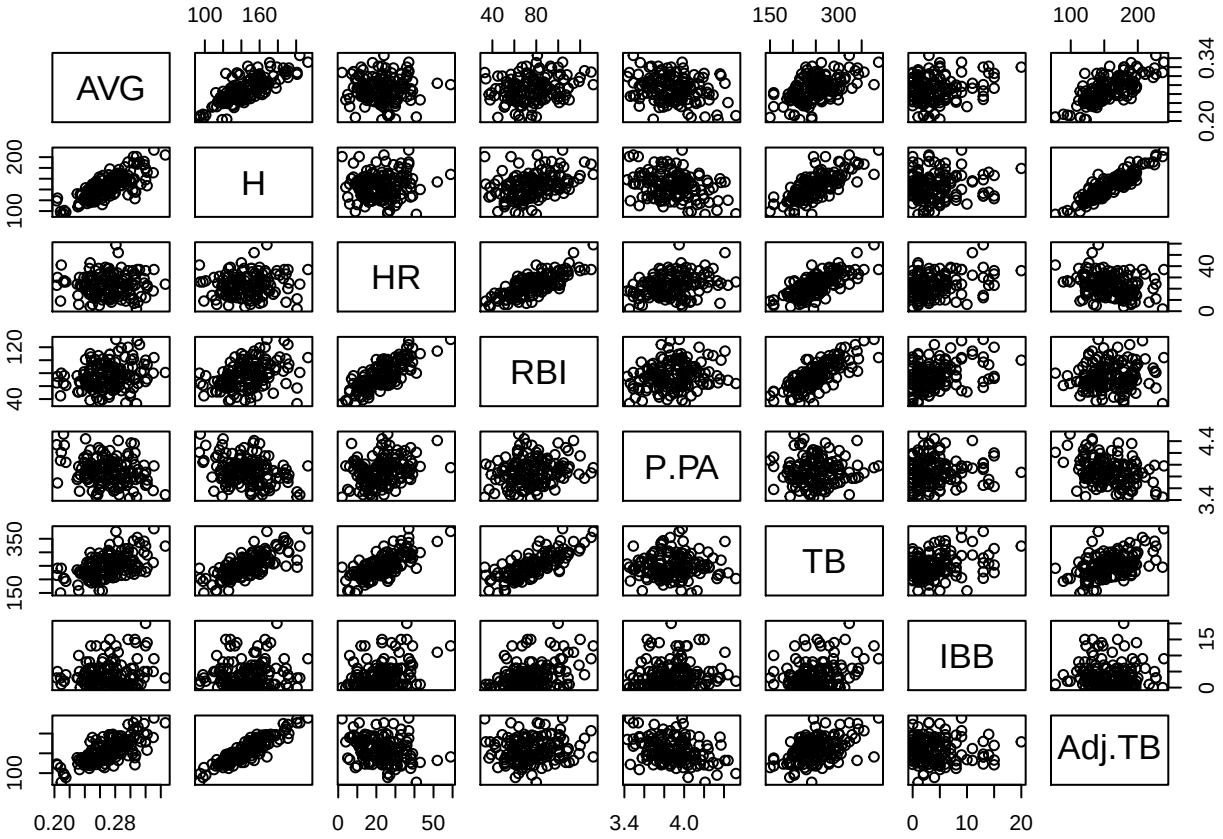
Pairwise correlation

Histograms of HR and H





Linear models for individual variables against HR



	AVG	H	HR	RBI	P.PA	TB	IBB	Adj.TB
AVG	1.00	0.77	0.01	0.22	-0.25	0.52	0.17	0.72
H	0.77	1.00	0.09	0.36	-0.29	0.73	0.06	0.91
HR	0.01	0.09	1.00	0.78	0.30	0.72	0.34	-0.26
RBI	0.22	0.36	0.78	1.00	0.20	0.77	0.36	0.08
P.PA	-0.25	-0.29	0.30	0.20	1.00	0.03	0.09	-0.34
TB	0.52	0.73	0.72	0.77	0.03	1.00	0.27	0.48
IBB	0.17	0.06	0.34	0.36	0.09	0.27	1.00	-0.05
Adj.TB	0.72	0.91	-0.26	0.08	-0.34	0.48	-0.05	1.00

We did an initial regression with all variables as predictors. In this first regression, total bases (TB) was heavily correlated with home runs (HR) and allowed us to predict home runs with almost perfect accuracy. When a player scores a home run, they have travel 4 bases. We then subtracted a player's home run from TB; this perfectly correlated with home runs.

The formula we used to take out the TB is the following: Adj.TB = TB - HR * 4 In the place of TB, we have used this adjusted TB (Adj.TB) variable.

Model selection part I: Stepwise, R-Square, Adjusted R-Square, AIC, BIC, PRESS, and Cp

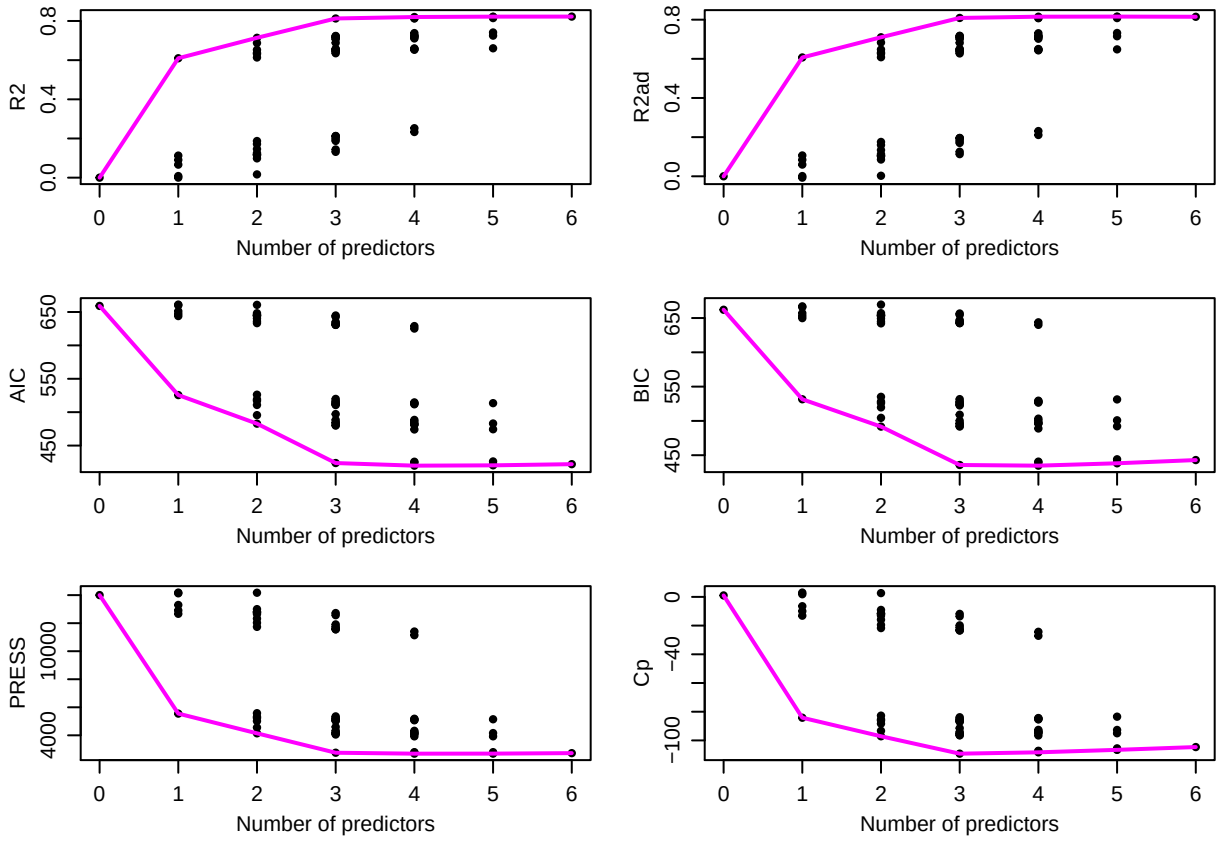
1.1 Forward Selection

Forward selection resulted in a best AIC of 419.92. The AVG and IBB variables were removed from the model and the best model is reported below.

```
lm(formula = HR ~ RBI + Adj.TB + H + P.PA, data = data)
```

We also ran ran backward variable selection and forward and backward stepwise selection. Each of these methods resulted in the same best model as that chosen by forward selection.

2 All Possible Regressions



Model Selection Results

R-Square

The table below shows that the more independent parameters we have for R-Squared, the higher our R-Squared value will be. The highest R-squared was 0.822. The best model was the full model with all variables included.

lm(formula = HR ~ AVG + H + RBI + P.PA + IBB + Adj.TB, data = data)

	p-1	X1	X2	X3	X4	X5	X6	R2
64	6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	0.8225131
63	5	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	0.8221917
48	5	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	0.8203667
47	4	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	0.8203263
56	5	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	0.8148534

R-Square Adjusted

For R-Squared Adjusted, this value will not necessarily increase as additional terms are introduced into the model. We want a model with the maximum Adjusted R-Square. The highest chosen R-squared adjusted was 0.816. And the model was

lm(formula = HR ~ H + RBI + P.PA + IBB + Adj.TB, data = data), where X1 = AVG is FALSE.

	p-1	X1	X2	X3	X4	X5	X6	R2ad
63	5	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	0.8157494
47	4	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	0.8151559
64	6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	0.8147399
48	5	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	0.8138582
55	4	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	0.8091876

AIC

For AIC, we want the model that gives us the lowest AIC. The model that is chosen for the lowest AIC is 419.92.

lm(formula = HR ~ H + RBI + P.PA + Adj.TB, data = data), where X1 = AVG, X5 = IBB and both are FALSE.

	p-1	X1	X2	X3	X4	X5	X6	AIC
47	4	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	419.9175
63	5	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	420.4146
48	5	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	421.8851
64	6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	422.1541
39	3	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	423.8447

BIC

As we can see for BIC, we want the model that gives us the lowest BIC. The model that is chosen for the lowest BIC is 434.78 and the model was the same model chosen by AIC.

	p-1	X1	X2	X3	X4	X5	X6	BIC
47	4	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	434.7665
39	3	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	435.7239
63	5	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	438.2335
55	4	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	439.3426
48	5	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	439.7040

PRESS

For PRESS, we want the model that gives us the lowest PRESS. The model that is chosen for the lowest PRESS is 2691.09. The model chosen by the PRESS metric was also the same as that chosen by BIC and AIC.

	p-1	X1	X2	X3	X4	X5	X6	PRESS
47	4	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	2691.085
63	5	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	2695.059
64	6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	2728.405
48	5	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	2731.124
39	3	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	2757.406

Cp

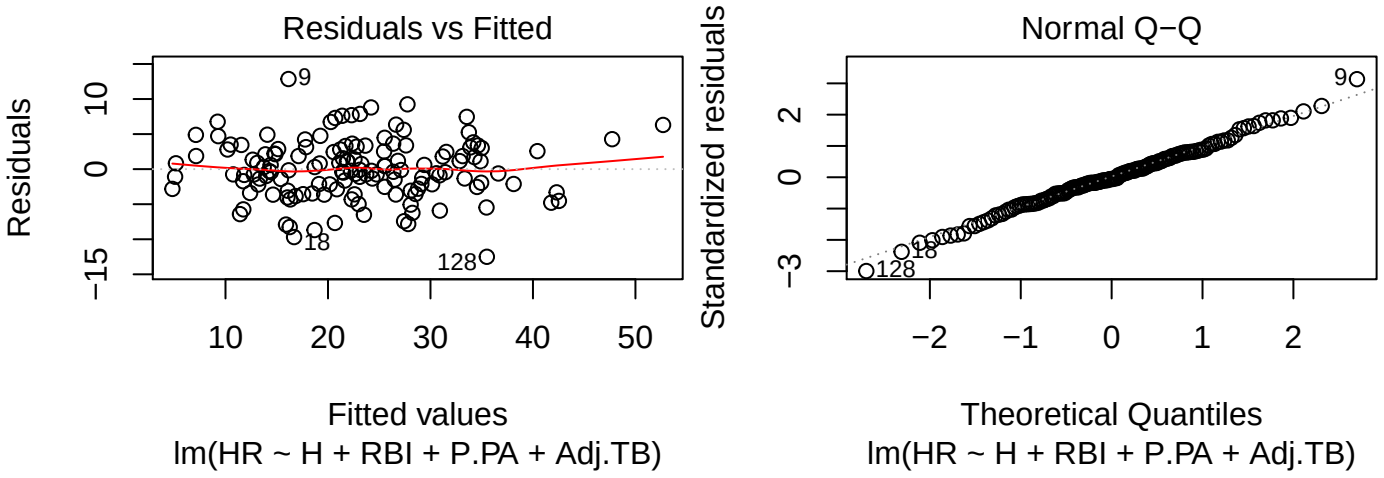
As we can see for Cp, we want the model that gives us the lowest Cp of -109.23. Based on the Cp metric the P.PA variable was removed, although it had been kept for previous models. The best model was:

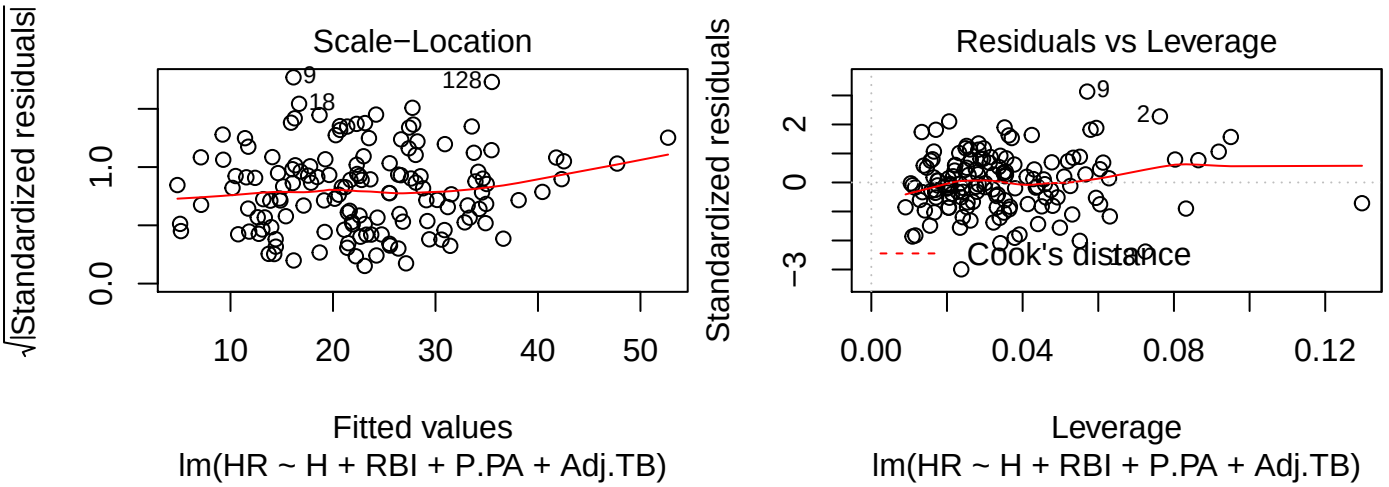
lm(formula = HR ~ H + RBI + Adj.TB, data = data),

Notice how we have a negative Cp value. We must beware of negative values of Cp. This could have been resulted because the MSE for the full model overestimates the true σ^2.

	p-1	X1	X2	X3	X4	X5	X6	Cp
39	3	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	-109.2270
47	4	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	-108.3067
55	4	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	-107.4771
40	4	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	-107.2338
63	5	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	-106.5734

Model Diagnostics





The normality plot shows that most points lie along the quantile-quantile line so the distribution of residuals is approximately normal. From the normality plot, we can see that 18, and 128 are close to -3 standard deviation away from the mean. And 9 is near 3 standard deviations from the mean.

For the residuals vs. fitted plot the three points identified in normality plot were also identified here as having large residual values. Along the fitted values there appears to be constant variance, which matches our model assumption. It also looks like the relationship between our predictors and the response is linear.

In the residuals vs. leverage there are a few points with high leverage but they don't conincide with the points with high residuals; therefore those points shouldn't have too big an effect on the overall model fit. All points were within the Cook's Distance of 0.5 so there are no points that are overly influential.

	Rank	PLAYER	TEAM	AVG	H	HR	RBI	P.PA	TB	IBB	Adjusted.TB
9	9	Jose Ramirez	CLE	0.318	186	29	83	3.99	341	5	225
18	18	Joe Mauer	MIN	0.305	160	7	71	4.36	219	3	191
128	128	Albert Pujols	LAA	0.241	143	23	101	3.91	229	5	137

If we look at our high residual points, player 9 had home runs much higher than those expected by the model and players 18 and 128 had home runs much lower than the model predicted. Player 18 also shows up in the high-leverage points. This is likely because his total number of home runs is lower on the scale of total home runs for all players, so the point has more of an effect on the model fit.

	Rank	PLAYER	TEAM	AVG	H	HR	RBI	P.PA	TB	IBB	Adjusted.TB
2	2	Charlie Blackmon	COL	0.331	213	37	104	3.98	387	9	239
13	13	Nolan Arenado	COL	0.309	187	37	130	3.86	355	9	207
14	14	Dee Gordon	MIA	0.308	201	2	33	3.45	245	0	237
18	18	Joe Mauer	MIN	0.305	160	7	71	4.36	219	3	191
20	20	Ender Inciarte	ATL	0.304	201	11	57	3.54	271	3	227
46	46	Aaron Judge	NYY	0.284	154	52	114	4.41	340	11	132
50	50	Giancarlo Stanton	MIA	0.281	168	59	132	3.95	377	13	141
140	140	Curtis Granderson	LAD/NYM	0.212	95	26	64	4.52	203	2	99

Model selection part II: Cross validation

Now we will take a look at cross validation and see which model performs best using cross validation. We used the bootstrap version of cross-validation using 100 iterations on training sets with 75% of the data and predictions made on test sets with the remaining 25%. The model with the lowest mean MSPR across all validation trials was the same model chosen as the best model using stepwise variable selection, AIC, BIC, and PRESS.

We tested the following four models which include all models chosen as the best models for the previous model selection metric.

model1: lm(formula = HR ~ H + RBI + Adj.TB, data = train)

model2: lm(formula = HR ~ H + RBI + P.PA + Adj.TB, data = train)

model3: lm(formula = HR ~ H + RBI + P.PA + IBB + Adj.TB, data = train)

model4: lm(formula = HR ~ H + RBI + P.PA + IBB + Adj.TB + AVG, data = train)

Cross Validation results

model1	model2	model3	model4
19.456	19.072	19.214	19.512

Model 2 had the lowest mean prediction error and this model matches the models selected by AIC, BIC and PRESS metrics.

Conclusion:

From the data we were given stepwise variable selection, AIC, BIC, PRESS, and Cross-Validation all gave the same best model for predicting Home-Runs. The model is

$$Homeruns = \beta_0 + \beta_1 Hits + \beta_2 RBIs + \beta_3 P.PA + \beta_4 AdjustedTB.$$

Looking at the standardized coefficients for this model, my hypothesis that RBI predict home run hitters is wrong. RBI's are positively related with home runs but the coefficient in our model is not that big relative when controlling for other factors in the model.

Standardized model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.00	0.04	0.00	1.00
H	1.12	0.12	9.02	0.00
RBI	0.46	0.05	8.71	0.00
P.PA	0.10	0.04	2.42	0.02
Adj.TB	-1.28	0.11	-11.21	0.00

The number of hits a player has is positively correlated with home runs; whereas the adjusted touched bases is negatively correlated with home runs. These variables are more predictive of a players number of home runs than RBIs. We can now successfully say we have solved the mystery on what makes a player, a Babe Ruth slugger!

Sources:

ESPN: http://www.espn.com/mlb/stats/batting)

MLB Sabermetric Statistics : (http://www.espn.com/mlb/stats/batting/_/type/sabermetric)