Conor Juengst

STAT 499

Dr. Lin – Fall 2015

On the Predictability of ERA and Other Measures of Pitcher Performance

Introduction:

The capacity to accurately forecast a player's future performance is crucial to the success of any professional sports franchise. In Major League Baseball, this ability is of particular significance as the free agency and salary arbitration systems result in personnel decisions involving fully-guaranteed contracts and large sums of money based largely on a player's past performance. Determining an appropriate valuation of a player should be solely based on what they are expected to contribute in the future – it is nonsensical to pay for past performance – and therefore it is a priority for clubs to pay foremost attention to statistics that best indicate how a player will perform in upcoming seasons.

This paper attempts to expound the primary relationships between a starting pitcher's statistics from past seasons and measures of that pitcher's performance in seasons that have not yet occurred. A multitude of techniques are utilized to explore these connections with this study restricting its attention to starting pitchers to avoid some of the small sample size issues inherent to relief pitchers.[1] Through this effort, dependencies between predictor variables are illuminated and the relative strength of their linear relationships with our selected response variables are quantified letting us determine which factors are most influential.

Data:

All data sets used for this analysis were exported from fangraphs.com using their custom leaderboard tool (which uses data supplied by Baseball Info Solutions). The selected data ranges from the 2006 to the 2015 MLB seasons providing a broader scope relative to an analysis of only the most

1 Relievers pitch a comparatively small number of innings per game/season which can result in misleading statistics and introduce noise into regression models.

recent seasons. Independent variables are selected by considering the strength of their year-to-year correlation and grouped into three-year segments to allow rate statistics to better stabilize. The dependent variables that this study is focused on are: ERA (Earned Run Average), RS/9 (Total Runs Scored Per 9 Innings), FIP (Fielding Independent Pitching), xFIP (Expected FIP), and SIERA (Skill Interactive ERA). Predictor and response data sets are merged together using FanGraphs' *playerid* identifier with inclusion in each contingent on reaching a minimum of 100 innings pitched per year. This process results in a final data frame containing 227 observations which will be used throughout this analysis.

Predictor (min. 300 IP)	**Response (min. 100 IP)**
2012-2014	2015
2009-2011	2012
2006-2008	2009

Methods:

This paper makes extensive use of ordinary least squares regression methods and diagnostics for studying the predictive potential and suitability of various independent variables. In addition, exploratory factor analysis is used for the purposes of dimension reduction and examining the latent relations between our predictors. Rotated factor patterns are then used to generate new variables for use in predictive models that allow for clear interpretation and low multicollinearity. Principal component regression methods are also employed for comparative purposes.[2]

Selection of Predictors:

The FanGraphs database contains a diverse array of pitching statistics covering a range of categories including batted ball, plate discipline, and pitch type data. Independent variables were selected for inclusion in predictor data sets primarily based on the strength of their year-to-year correlation with themselves (e.g. ground ball rates for pitchers correlate highly from one season to the

2 Principal Components are linear combinations of variables that are orthogonal to each other and therefore have zero multicollinearity.

next). This is particularly important since we are constructing predictive models. Notable variables excluded due to not meeting this criteria include HR/FB (Home Run to Fly Ball ratio), LOB% (Left on Base percentage), LD% (Line Drive percentage), and BABIP (Batting Average on Balls in Play). Also considered is the normality of our predictors and the significance of their linear relationships with our response variables.

Predictors/Independent Variables			
Standard/Rate Statistics	**Batted Ball**	**Plate Discipline**	**Pitch Type**
Batting Average Against	log[GB/FB][3]	O-Swing%	Fastball%[4]
K%	GB%	Z-Swing%	
BB%	FB%	Swing%	
K-BB%	Pull%	O-Contact%	
	Cent%	Z-Contact%	
	Oppo%	Contact%	
	Soft%	Zone%	
	Med%	F-Strike%	
	Hard%	SwStr%	

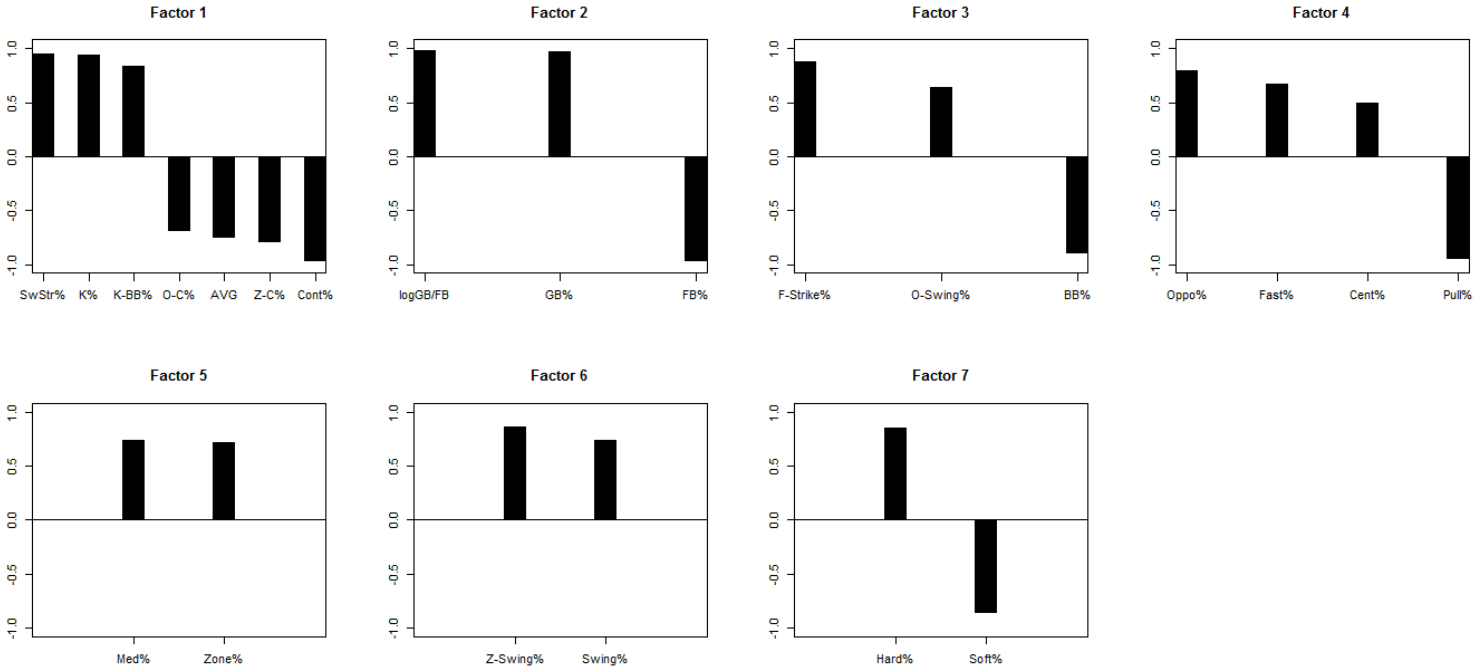
Creating Factor-Based Variables:

The primary goal of exploratory factor analysis is to explain the variability of a data set using a small number of underlying factors. This approach allows for reduced dimensionality (*m* factors instead of *n* independent variables) and alternative insights into how the variables are related to each other. Our 23 selected predictors are scaled down to 7 factors with each representing a different characteristic of a pitcher's performance.

Although we are constructing predictive models, our main objective is not necessarily to predict our response variables but instead to evaluate the specific components that effect measures of pitching performance in future seasons. Rather than implementing conventional principal components as predictors for linear modeling we are creating new variables using the most influential loadings from a

3 GB/FB appears to follow a log-normal distribution. See appendix for histograms.
4 Fastball% (percentage of total pitches that are fastballs) is used as a proxy for a pitcher's repertoire.

rotated factor structure.[5] This decision imparts our models with a more interpretable set of variables (respective to a principal component regression model) that are nearly absent of multicollinearity which helps in quantifying the impact of each on our response variables. A consequence of this approach is a potentially lower proportion of variance explained by our model compared to a similar principal component regression model. This is a result of our concentration on the large loadings in the factor pattern and the disregard of lesser loadings that contribute only slightly to the overall factor score.



Variable Interpretation:

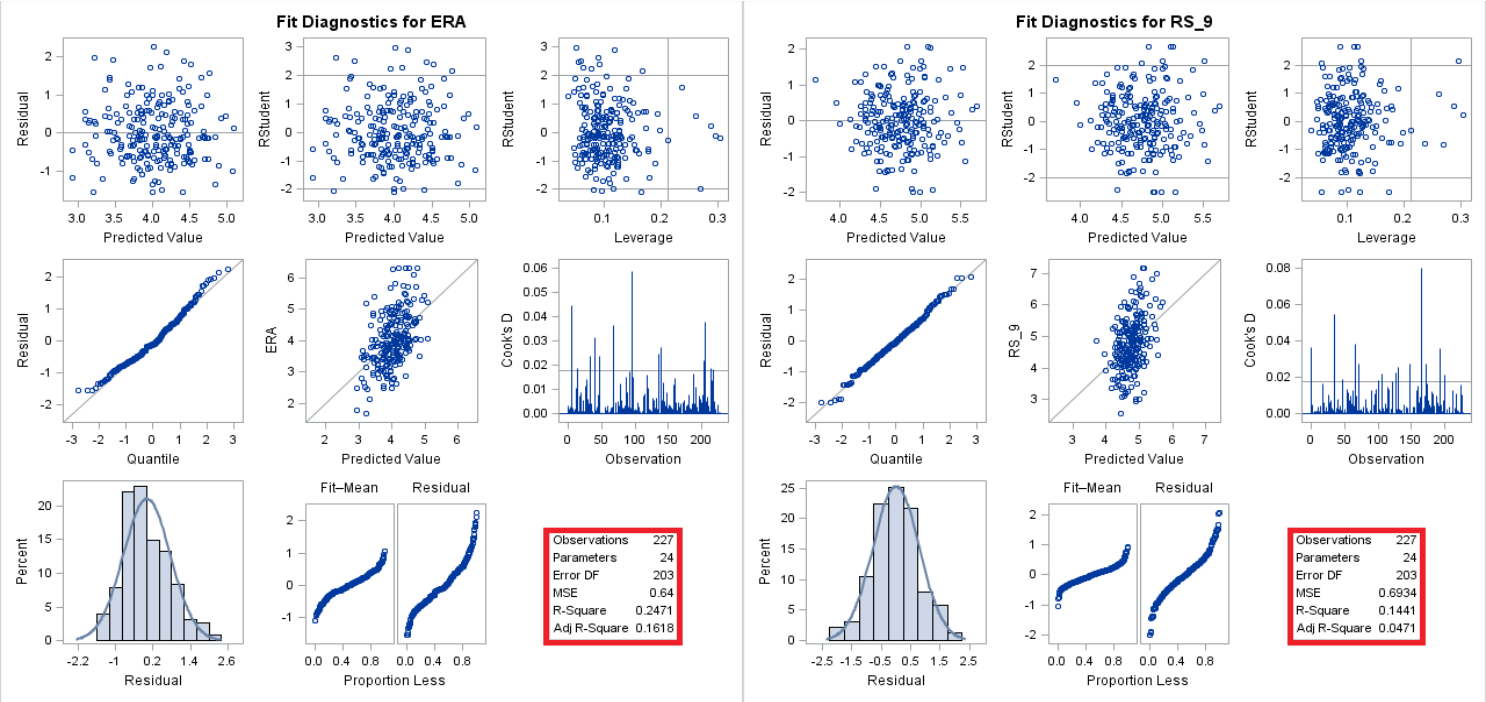
Each factor that we have generated can be thought of as a weighted grouping of independent variables that represents a distinct aspect of a pitcher's statistical profile. For example, the first and most complex of our factors appears to be an approximation of a pitcher's 'swing and miss' ability. The second factor rates a pitcher's proficiency for inducing ground balls, and the third can be thought of as a measurement of their command and plate discipline (e.g. starting at-bats with a strike and coaxing swings at pitches outside of the strike zone). The fourth factor is intriguing as it seems to reward a pitcher for preventing a batter from hitting a pitch to the pull field which has been shown to potentially

5 Factors are calculated with the factor procedure in SAS using the principal component method and a varimax rotation.

have a higher expected run value (Weinstein, 2014). Also noteworthy is how the proportion of fastballs that a pitcher throws influences this factor with fastball-heavy pitchers mitigating damage from hits towards the pull field. The last three factors are variables that simply relate two components to yield further insight into how our remaining predictors vary with each other and with the response variables.

Traditional Statistics – ERA and RS/9:

ERA is far from a perfect measure of a pitcher's performance but it is easily the most commonly used metric for diagnosing how well a player has pitched over a period of time. In comparison, RS/9 simplifies calculation by disregarding the notion of unearned runs and instead providing a raw average of total runs allowed by a pitcher per 9 innings. While it is more straightforward to compute, a pair of multiple linear regression models constructed with all of our independent variables indicates that RS/9 is much harder to predict compared to ERA when using this data. Correlation analysis yields a similar sentiment with RS/9 lacking meaningful linear relationships with any of our predictors – even ones that have significant correlations with ERA and our other response variables. If the ultimate goal is to approximate the effectiveness of a pitcher then RS/9 is seemingly too volatile to provide worthwhile information. ERA on the other hand demonstrates a clearer linear association with our predictors.



Repeating the procedure with a couple of principal component regression models (with principal components one through seven as predictors) furthers the theory that RS/9 does not vary linearly with this specific data set – perhaps signifying that it is a metric dependent on other measures such as defense and park factors. Replicating the process one more time with our factor-based variables again reiterates the premise of our data set not being suitable for modeling RS/9 while proving adequate at best for the prediction of ERA.

Observations	227	Observations	227
Parameters	8	Parameters	8
Error DF	219	Error DF	219
MSE	0.6495	MSE	0.7296
R-Square	0.1758	R-Square	0.0284
Adj R-Square	0.1495	Adj R-Square	-0.003
Observations	227	Observations	227
Parameters	8	Parameters	8
Error DF	219	Error DF	219
MSE	0.6573	MSE	0.73
R-Square	0.1658	R-Square	0.0279
Adj R-Square	0.1391	Adj R-Square	-0.003

Principal Component Model

Factor-Based Model

ERA RS/9

Defense Independent Pitching Statistics – FIP and xFIP:

Among our response variables, FIP and xFIP fall under the classification of DIPS (Defense Independent Pitching Statistics). These metrics concentrate on outcomes that the pitcher is principally responsible for: Strikeouts, walk, and home runs. All other outcomes of a plate appearance are influenced at least in part by outside factors (e.g. defense). By removing defense from the equation we get performance measures that are influenced far less by the team they play for and in turn offer a truer estimate of pitching ability in contrast with statistics like ERA and RS/9.

Intuitively, xFIP proves easier to forecast relative to FIP due to the normalized home run rate utilized in its calculation. The difference is slight though clear in both principal component and factor-based models and the linear relation in each is considerably stronger than our ERA regression models.

Observations	227	Observations	227
Parameters	8	Parameters	8
Error DF	219	Error DF	219
MSE	0.3741	MSE	0.2346
R-Square	0.3107	R-Square	0.3482
Adj R-Square	0.2887	Adj R-Square	0.3273

Principal Component Model

Observations	227	Observations	227
Parameters	8	Parameters	8
Error DF	219	Error DF	219
MSE	0.3725	MSE	0.2268
R-Square	0.3137	R-Square	0.3697
Adj R-Square	0.2918	Adj R-Square	0.3496

Factor-Based Model

FIP xFIP

Of special interest in this pair of models is the lower values of mean squared error and higher proportion of variation explained by our factor-based models compared to principal component models – particularly in the xFIP case. This suggests that our simplified variables are perhaps better suited for forecasting purposes than traditional principal components in certain cases such as predicting DIPS.

A Non-DIPS ERA Estimator – SIERA:

A successor to FIP and xFIP in the realm of ERA estimators, SIERA (Skill Interactive Earned Run Average) aims to estimate the true level of a pitcher's performance without disregarding the influence of balls in play. Calculation of this metric is much more involved and varies depending on the year and the source.[6] The result is an ERA estimator that boasts greater predictive power than its peers (in predicting next season's ERA) and offers a more precise representation of a pitcher's innate ability.

With its similarities to our defense independent ERA estimators, SIERA would seem apt for prognostication using factor-based variables in hopes of attaining similar benefits over principal component methods that we observed in our FIP and xFIP models. This is precisely the case as we again see a distinct improvement in the factor-based models when compared to the closely related principal component versions. In addition, the SIERA model maintains an even stronger linear relationship with our predictors than its defense independent relatives despite its increased complexity.

6 FanGraphs and Baseball Prospectus have alternate methods of calculation with differing parameters and adjustment constants. This paper utilizes the FanGraphs variant.

Observations	227
Parameters	8
Error DF	219
MSE	0.2244
R-Square	0.3688
Adj R-Square	0.3486

Principal Component Model

Observations	227
Parameters	8
Error DF	219
MSE	0.2175
R-Square	0.3882
Adj R-Square	0.3687

Factor-Based Model

SIERA

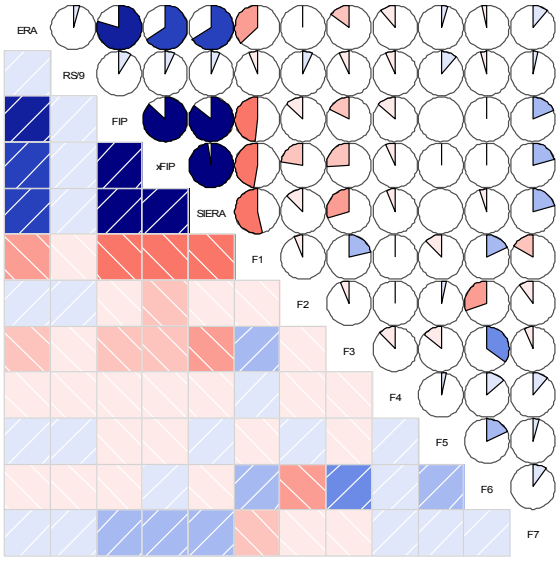
Model Comparison:

Analyzing the parameter estimates from each of our factor-based models and a correlation matrix containing our response variables and predictors highlights several intriguing relationships. While our ERA model was easily the least powerful of the group it still bears a modest resemblance to other four models, and in particular the FIP model (note Factors 3, 4, and 6). The SIERA and xFIP models also feature various parallels between their parameters and have comparable R^2 and MSE values suggesting that the metrics are nearly homogeneous in spite of SIERA's convoluted formula.

Parameters	ERA Model	FIP Model	xFIP Model	SIERA Model
Intercept	1.42295	2.05735	2.80732	2.96588
Factor 1	-1.879	-2.12477	-1.67282	-1.82072
Factor 2	.05223	-.23641	-.32411	-.18945
Factor 3	-2.06938	-2.01062	-2.83635	-2.93029
Factor 4	-1.13012	-1.13256	-.69382	-.61228
Factor 5	-.75086	-2.20716	-2.11698	-2.01374
Factor 6	2.46858	2.46217	2.56562	2.00347
Factor 7	1.3242	1.89075	1.62534	1.61025

This assertion is corroborated when looking at the correlation matrix where xFIP and SIERA have an extremely strong linear relationship (i.e. their Pearson's correlation coefficient is .9776). We can also discern the relative importance of each of our factors by looking at their correlation with our

response variables. Factor 1 is clearly the dominant factor in all scenarios with Factor 3 being a distant second. Factor 2, influenced by ground ball/fly ball rates, contributes to a lesser extent yet it is logical that its strongest correlation would be with xFIP (which makes use of a standardized home run rate).[7] Factors 5 and 6 appear to impact our response variables very little though Factor 7, our hard/soft contrast variable, shows a significant negative correlation with them.



Graphical representation of correlation matrix (using *corrgram* package in R).

Conclusion:

ERA and other measures of a pitcher's performance are not easily forecasted through traditional means. Standard multiple linear regression models offer decent results but interpreting the parameters is difficult due to problems with multicollinearity. Principal component regression models solve that issue but introduce another because each component is calculated as a linear combination of all independent variables creating predictors that are impossible to accurately interpret.

By crafting variables from a rotated principal component factor pattern we have resolved these complications and elucidated a few of the underlying relationships between pitching statistics from one season and performance metrics of the next. In addition, our new factor-based variables outperformed

7 Since xFIP uses the same HR/FB rate for each player, a pitcher is directly penalized for yielding more fly balls.

their principal component relatives at predicting FIP, xFIP, and SIERA while only experiencing a slight decrease in power in the case of ERA. In conclusion, these variables offer a new way to project a pitcher's future performance on par with current ERA estimators. They also allow for better prediction of the estimators themselves which are truer approximations of a pitcher's talent level.

Appendix:

ERA = 9 * (Earned Runs Allowed / Innings Pitched)

RS/9 = 9 * (Total Runs Allowed / Innings Pitched)

[an unearned run is one that would not have been scored without the aid of an error (either fielding error or passed ball)]
[an earned run is any run that does not qualify as unearned]

FIP = (13*HR + 3*(BB + HBP) – 2*K) / (Innings Pitched) + C

xFIP = (13*xHR + 3*(BB + HBP) – 2*K) / (Innings Pitched) + C

C = ERA – [(13*HR + 3*(BB + HBP) – 2*K) / (Innings Pitched)],

where all values are league averages from year that you are calculating

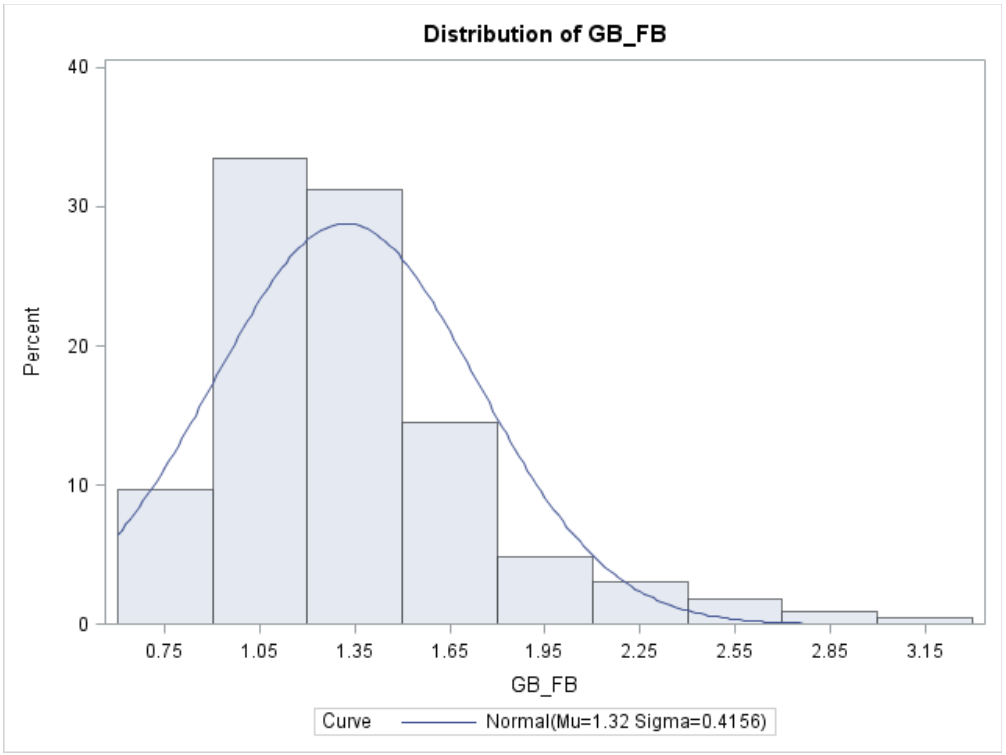
xHR = [(# of fly balls given up by pitcher) * (league average home run per fly ball rate)]

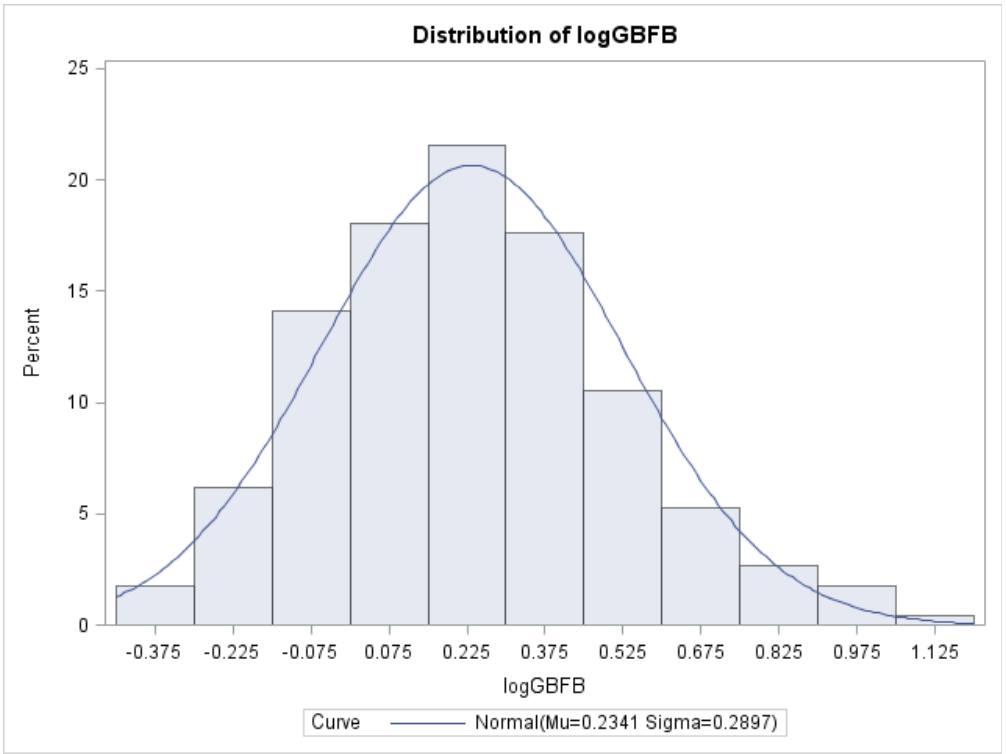
SIERA is calculated from strikeout, walk, and net ground ball (GB – FB) data:

SIERA calculation coefficients for 2010 season

Variable	Coefficient
K / PA	-15.518
$(K / PA)^2$	9.146
BB / PA	8.648
$(BB / PA)^2$	27.252
netGB / PA	-2.298
$(netGB / \|netGB\|)*(netGB / PA)^2$	-4.920
(K / PA)*(BB / PA)	-4.036
(K / PA)*(netGB / PA)	5.155
(BB / PA)*(netGB / PA)	4.546
Constant	5.534
Year coefficient (2010 is base year)	0
% innings as SP	0.367

netGB = [# of ground balls – # of fly balls]


Distribution of GB_FB
Percent
40
30
20
10
0
0.75
1.05
1.35
1.65
1.95
2.25
2.55
2.85
3.15
GB_FB
Curve
Normal(Mu=1.32 Sigma=0.4156)


Distribution of logGBFB
Percent
25
20
15
10
5
0
-0.375
-0.225
-0.075
0.075
0.225
0.375
0.525
0.675
0.825
0.975
1.125
logGBFB
Curve
Normal(Mu=0.2341 Sigma=0.2897)

References:

James, B. (1986-1988). *The Bill James Baseball Abstract*. Ballantine Books.

Johnson, R., & Wichern, D. (2014). *Applied Multivariate Statistical Analysis* (Sixth ed.). Prentice-Hall.

Kutner, M. (2005). *Applied Linear Statistical Models* (Fifth ed.). McGraw-Hill Irwin.

Petti, B. (2012, January 9). *What Starting Pitcher Metrics Correlate Year-to-Year?* Retrieved December 17, 2015, from http://www.beyondtheboxscore.com/2012/1/9/2690405/what-starting-pitcher-metrics-correlate-year-to-year

Staude, S. (2013, December 13). *Tool: Basically Every Pitching Stat Correlation*. Retrieved December 17, 2015, from http://www.fangraphs.com/blogs/tool-basically-every-pitching-stat-correlation/

Thorn, J., & Palmer, P. (1985). *The Hidden Game of Baseball: A Revolutionary Approach to Baseball and its Statistics*. Doubleday.

Weinstein, M. (2014, June 19). *Exploring Batted Ball Run Values and Spray*. Retrieved December 17, 2015, from http://www.hardballtimes.com/when-does-spray-data-become-reliable/