PENALIZED REGRESSION MODELS

FOR MAJOR LEAGUE BASEBALL METRICS

by

MUSHIMIE LONA PANDA

(Under the Direction of Cheolwoo Park)

ABSTRACT

Major League Baseball is a sport complete with a multitude of statistics to evaluate a player's performance and achievements. In recent years, traditional statistics are constantly being supplemented by more sophisticated modern metrics, to determine a player's predictive power. To address this issue, we use penalized regression models to determine which offensive and defensive metrics are consistent measures of a player's ability. Penalized linear regression techniques which have shrinkage and variable selection mechanisms, have been widely used to analyze high dimensional data. We use three popular regularized regression methods in our data analysis, Lasso, Elastic Net, and SCAD. We implement these regularized regression models on a set of thirty-one different offensive metrics and five defensive metrics. The results indicate that two defensive metrics stand out to distinguish players across time and the offensive metrics can be reduced to seven metrics, which is a substantial reduction in the dimensionality.

INDEX WORDS: baseball, penalized regression, principal component analysis

Penalized Regression Models

For Major League Baseball Metrics

by

MUSHIMIE LONA PANDA

B.S., LaGrange College, 2012

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2014

Penalized Regression Models
For Major League Baseball Metrics

by

Mushimie Lona Panda

Approved:

Major Professor: Cheolwoo Park

Committee: Kevin Byon
Jaxk Reeves

Electronic Version Approved:

Dr. Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

Penalized Regression Models

For Major League Baseball Metrics

Mushimie Lona Panda

Dedication

To my parents, there are not enough words to describe the gratitude in my heart for your boundless love and support. Your words of encouragement provided me with great strength which carried me through to complete this thesis. To my sister, thank you for having faith and belief in me that I could accomplish this momentous milestone. To my brother, you are very special to me and thank you for supporting me all the way. I am incredibly blessed to have an amazing family by my side in all that I do.

Acknowledgments

I would like to sincerely and wholeheartedly thank my advisor, Dr. Cheolwoo Park. After asking numerous faculty members in the department to be my advisor, Dr. Park kindly accepted my request without knowledge of my skills and abilities. His patience, guidance, and ability to effectively communicate with me in an obliging manner, has greatly been appreciated throughout the completion of this thesis. There is no way that with merely words, I could ever fully express my gratitude and how humble I am to have received his support.

I would also like to thank my advisory committee members, Dr. Jaxk Reeves and Dr. Kevin Byon. Dr. Reeves has showed great unwavering support and genuine interest in my academic studies and for that I am sincerely grateful for his contribution to my continuing success. Dr. Byon graciously accepted to be on my committee and has provided valuable comments and dynamic direction in the completion of this thesis.

Contents

Acknowledgements **v**

List of Figures **viii**

List of Tables **ix**

1 Introduction **1**

2 Data Description **4**

3 Penalized Least Squares Regression **8**

3.1 The Lasso 10

3.2 Smoothly Clipped Absolute Deviation 10

3.3 Elastic Net 11

3.4 Cross Validation 12

4 Principal Component Analysis **14**

4.1 Fielding Metrics 15

4.2 Hitting Metrics 18

5 Penalized Regression Analysis **23**

5.1 Defensive Metrics 24

5.2 Offensive Metrics 26

6 Summary **32**

Bibliography **34**

List of Figures

4.1 Scree Plot of Fielding Metrics 2002 - 2012 17
4.2 Scree Plot of Hitting Metrics 2002 - 2012 19

5.1 Fielding Metrics - Mean vs. Signal 25
5.2 Fielding Metrics - S.D. vs. Signal 26
5.3 Fielding Metrics - S.D. vs. Mean 26
5.4 Hitting Metrics - Mean vs. Signal 28
5.5 Hitting Metrics - S.D. vs. Signal 31
5.6 Hitting Metrics - S.D. vs. Mean 31

List of Tables

2.1 Defensive Measures 6

2.2 Offensive Measures 7

4.1 Descriptive Statistics for Fielding Metrics 2002 - 2012 16

4.2 Correlation Matrix for Fielding Metrics 2002 - 2012 16

4.3 Eigenvalues of the Correlation Matrix for Fielding Metrics 2002 - 2012 17

4.4 Eigenvectors for Fielding 2002 - 2012 18

4.5 Eigenvalues of the Correlation Matrix for Hitting Metrics 2002 - 2012 20

4.6 Eigenvectors for Hitting 2002 - 2012 21

4.7 Descriptive Statistics for Hitting Regular Season 2002-2012 22

5.1 Defensive Model Statistics 27

5.2 Offensive Model Statistics 29

5.3 Offensive Model Statistics 30

Chapter 1

Introduction

Baseball is rooted deep in American culture and known as America's favorite pastime. Major League Baseball (MLB) has shown an increased interest in statistics to measure a player's offensive ability and evaluate his performance on the field (Eder, 2013). A player's batting average was once defined as his worth, but this measure of hitting performance is now being supplemented by more complex metrics to predict a player's future performance (Davison, 2013).

With Major League Baseball becoming an ever advancing statistical game, there are newer sabermetrics concepts being created to better analyze players. Bill James, an renown American baseball writer, historian and statistician, is well known for coining the term sabermetrics and creating this cutting edge statistical approach. Sabermetrics uses statistical analysis to analyze baseball records and player performance. In 1977, James released his book called, *The Bill James Historical Baseball Abstract*. This book was widely influential in the world of baseball and took baseball research to an innovative level. The sabermetric method was illustrated in the 2011 film, *Moneyball*, based on the Michael Lewis' book by the same name. The film accounts the 2002 season of the Oakland Athletics team and their general manager Billy Beane's non-traditional sabermetric approach. Beane used this technique to analyze and scout players due to Oakland's limited payroll. This statistical method changed the game of baseball through evaluating the consistency and predictive performance of different offensive metrics, used to estimate the true ability of a player.

Major League Baseball has always been a sport where analysis of players is heavily dependent on statistics but recently, traditional statistics have been put aside as more sophisticated metrics are created and researched for player analysis. Regardless of the procedure used the goal remains the same, to measure a player's true ability and see which metrics are most reliable and consistent. There has been previous research on statistical analysis and baseball data. Neal et al. (2010) used linear regression models to estimate a MLB player's batting average in the second half of the season based on his performance in the first half. Their linear models consistently outperformed other Bayesian estimators in estimating batting averages based on the same amount of data. Yang and Swartz (2004) considered a two-stage Bayesian model for predicting the probability of winning a game in MLB. They looked into various factors such as team strength including the past performance of the two teams, the batting ability of the two teams and the starting pitchers. It was concluded that the proposed two-stage Bayesian model is effective in predicting the outcome of games in MLB.

We will evaluate offensive and defensive metrics to determine which metrics are consistent measures of player ability. McShane et. al (2011) also tried a similar method using a Bayesian approach, but only for offensive metrics. To do this, we use penalized linear regression models to determine which offensive and defensive metrics are most useful to distinguish players. We implement regularized regression models on thirty-one offensive metrics and five defensive metrics.

This thesis is organized as follows: in Chapter 2, we introduce the data description. In Chapter 3, we review the literature of penalized regression which includes the Lasso, Elastic Net, and SCAD. In Chapter 4, we explore the correlation between offensive metrics and defensive metrics with a principal component analysis. In Chapter 5, we discuss the results of our penalized regression analysis based on three criteria. In Chapter 6, we give a summary and suggest future work.

Chapter 2

Data Description

Our data comes from a Major League Baseball statistics and analysis website called Fangraphs.com. We have five defensive metrics outlined in Table 2.1 and forty-five offensive metrics outlined in Table 2.2. When selecting which players to include in our datasets, we set the minimum games played to be 10 for offensive and the minimum games played in the field for defense to be 100 games. The data for offense contains 24,110 player-seasons from 4,474 unique players spanning the 1973-2012 seasons. Data for fourteen of the forty-five offensive metrics were not available before the 2002 season; therefore, we fit the model on 7,249 player-seasons from 1,857 unique players.[1] Furthermore, the data for defense contains 5,585 player-seasons from 1,211 unique players during the 1973-2012 seasons. For 2002-2012 seasons we fit the model on 1,637 player-seasons from 476 unique players using the same five metrics.

McShane et. al (2011) used a hierarchical Bayesian variable selection model to determine which offensive metrics are most predictive within players. This Bayesian method is used for separating out players that are consistently distinct from the overall population on each offensive metric. For a particular offensive metric, Y_{ij} denotes the metric value for player i during season j. Therefore Y_{ij} denotes the player-seasons for each player's performance on a

[1] These metrics are BUH, FB, LD, GB, IFFB, IFH, GB/FB, LD%, GB%, FB%, IFFB%, HR/FB, IFH%, and BUH%.

given metric. The authors model the player-seasons for each metric as,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{2.1}$$

where $i = 1, ..., p$, p is the number of players, and ϵ_{ij} is assumed to be normally distributed with mean 0 and variance σ^2. The parameter μ is the overall population mean for the given metric and α_i is the player-specific differences from the population mean μ. Zeroed players have $\alpha_i = 0$ and non-zeroed players have $\alpha_i \neq 0$. They denote p_1 as the unknown proportion of players that are in the non-zeroed group. This proportion p_1, is the model parameter for evaluating the overall reliability of an offensive metric, as it gives the probability that a random chosen player shows consistent differences from the population mean. Therefore, metrics with high signal should have a high p_1 meaning a large fraction of players differ from the overall league mean.

We look at players who are consistently different from the population mean and those players who are not in a penalized setting. The ANOVA model in (2.1) can be rewritten as the following regression model for each metric:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{p-1,i} + \varepsilon_i \tag{2.2}$$

where $i = 1, ..., n$, n is the total number of observations, and ϵ_i is assumed to be normally distributed with mean 0 and variance σ^2. Our objective is to put the players in two groups, zeroed players and non-zero players. Zeroed players have $\beta_i = 0$, therefore, these players are close to the overall mean. The non-zeroed players have $\beta_i \neq 0$, which shows that these players are unique players. We are interested to see which metrics are high signal or low signal. To determine which metrics are high signal or low signal, we look at the fraction of non-zero players divided by the total number of players (McShane et. al, 2011). If the fraction is high, this implies that a metric will be more informative and have greater predictive power than the overall mean. If the fraction is low, then the metric has little predictive power and can be used to predict the overall league mean. We will also use the absolute mean and standard deviation to determine which metrics are set apart from the others. We consider these criteria because it is possible that the regression coefficients could be close to zero, even though the fraction is high. We will determine which offensive and defensive metrics are most consistent within players across time, given a metric contains a large amount of signal. Since the ordinary least squares estimator does not automatically set the coefficient to zero, i.e. does not select variables, we will apply penalized regression approaches to find informative metrics.

Table 2.1: Defensive Measures

Metric	**Description**
PO	putout
A	assists
E	errors
DP	double plays
FP	fielding percentage
	(PO+A)/(PO+A+E)

Table 2.2: Offensive Measures

Metric	Description
1B	singles
2B	doubles
3B	triples
AB	at-bats
AVG	batting average
BABIP	batting average for balls in play
BB	bases on balls (walks)
BB/K	walk to strikeout ratio
BB%	walk rate
BUH	bunt hit
BUH%	bunt hit rate
CS	caught stealing
GB	ground ball
GB/FB	ground ball to fly ball ratio
GDP	grounded into a double plays
PA	plate appearances
R	runs scored
H	hits
HR	home runs
RBI	runs batted in
OBP	on-base percentage
OPS	on-base percentage plus slugging (OBP+SLG)
IBB	intentional walks
HBP	hit by pitch
ISO	isolated power (SLG-AVG)
FB	fly ball
LD	line drive
IFFB	infield fly ball
IFH	infield hit
K%	strikeout rate
LD%	line drive rate
GB%	ground ball rate
FB%	fly ball rate
IFFB%	infield fly ball rate
HR/FB	home runs to fly ball ratio
IFH%	infield hit rate
SAC	sacrifice bunts
SB	stolen bases
SF	sacrifice flies
SLG	slugging percentage
SO	strikeouts
SPD	speed
wOBA	weighted on-base average
wRAA	weighted runs above average
wRC	weighted runs created

Chapter 3

Penalized Least Squares Regression

Linear regression attempts to model the relationship between the response variable and the explanatory variables by fitting a linear equation to observed data. A linear regression model is written $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{y}$ is an $n \times 1$ vector, $\boldsymbol{X}$ is an $n \times p$ matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameter coefficients. In our dataset, the explanatory variables are the players and the response variable is the different offensive and defensive metrics. When trying to fit a regression line to a dataset, a common approach is the method of least squares. We pick the coefficients β to minimize the residual sum of squares (RSS),

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{3.1}$$

which produces the least squares estimator $\hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$. Although this is a frequently used technique, these estimates are not always the best because of prediction accuracy and interpretation. The OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients. The OLS estimates have low bias and low variability when the relationship between the explanatory variables are correlated and the number of observations n is larger than the number of predictors p.

When $p > n$, the least squares cannot be used. The second reason is interpretation. With a large number of predictors, we often like to determine a smaller subset that exhibits the strongest effects (Tibshirani 1996). Hence, the model would be easier to interpret by setting the coefficients to zero, for unimportant variables.

There are two popular approaches for improving ordinary least square estimates, subset selection and ridge regression. Subset selection is a discrete process that the predictors are either added or deleted from the model. In subset selection, we look for the best subset. By choosing only a subset of predictor variables, subset selection produces a model that is interpretable but often has high variance. Therefore, other methods such as shrinkage methods are preferred. Ridge regression (Hoerl and Kennard, 1970) has a penalty on the size of the regression coefficients in order to shrink the coefficients toward 0. This is a continuous process shrinkage method that adds the ℓ_2 penalty term, $|\beta|^2 = \sum_{j=1}^{p} \beta_j^2$. The ridge regression estimator is $\hat{\boldsymbol{\beta}}^{ridge} = (\boldsymbol{X'X} + \lambda \mathbf{I})^{-1}\boldsymbol{X'y}$ where λ is a tuning parameter. If λ equals zero then the estimator is least squares. As $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}^{Ridge} \to 0$, which means the larger the λ, the more the coefficients shrinks. Ridge regression is more stable than subset selection because it uses a continuous process. However, the shrunken coefficients never have the value zero. Thus by using this penalty, there are no sparse solutions which does not give an easily interpretable model. There are several penalized regression methods that improve OLS and select relevant variables simultaneously.

Penalized linear regression techniques which have shrinkage and variable selection mechanisms have been widely used to analyze high dimensional data. The purpose of shrinkage is to prevent overfit arising due to either collinearity of the covariates or high-dimensionality. We use three popular regularized regression methods in our data analysis, Lasso, Elastic Net, and SCAD, which are capable of improving prediction ability and selecting important variables. These techniques will be applied to each offensive and defensive metric regression model consisting only of indicators for each player.

3.1 The Lasso

The least absolute shrinkage and selection operator (Lasso), proposed by Tibshirani (1996), is a penalized least squares regression that uses the ℓ_1 penalty on the regression coefficients. Unlike Ridge regression, the penalty of Lasso makes some of the coefficients to have the value zero. The Lasso estimate is defined as

$$\hat{\boldsymbol{\beta}}^{Lasso} = \underset{\beta}{\operatorname{argmin}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|_1, \tag{3.2}$$

where λ is a tuning parameter. The constraint term, $|\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|$, is called the ℓ_1 penalty. The ℓ_1 penalization is used on the regression coefficients, which tends to continuously shrink some coefficients toward zero as λ increases, and sets other coefficients exactly equal to zero if λ is sufficiently large. The Lasso regression penalty produces sparse solutions and performs both continuous shrinkage and automatic variable selection. By using the Lasso and shrinking some coefficients and setting others to zero, we are left with a smaller subset of variables that retain important features to explain the model. The Lasso tries to retain the good features of both subset selection and Ridge regression. However, the Lasso has a problem with optimization when the number of predictor variables are larger than the number of observations. At most, the number of variables that are selected can be equal to the number of observations. The glmnet function in the R package glmnet was used for Lasso.

3.2 Smoothly Clipped Absolute Deviation

The smoothly clipped absolute deviation (SCAD) penalty, is a non-convex penalty function designed to diminish the bias created when using the Lasso penalty (Fan and Li, 2001). It is a quadratic spline function with knots at λ and $a\lambda$. Fan and Li proposed that a good penalty function should result in an estimator with the following three properties: i) unbiasedness ii) sparsity iii) continuity. All three of these desirable properties are achieved by the SCAD penalty. The SCAD penalty is given by

$$p_{\lambda}^{scad}(\boldsymbol{\beta}_j) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \leq \lambda, \\ -\left(\dfrac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \text{if } \lambda < |\beta_j| \leq a\lambda, \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| > a\lambda \end{cases} \tag{3.3}$$

This corresponds to a quadratic spline function with knots at λ and $a\lambda$. The SCAD estimator $\hat{\boldsymbol{\beta}}^{SCAD}$ is the minimizer of

$$\underset{\beta}{\operatorname{argmin}} |\mathbf{y} - \mathbf{X}| + \sum_{j=1}^{p} p_{\lambda}^{scad}(\boldsymbol{\beta}_j). \tag{3.4}$$

SCAD penalized regression sets small coefficients to zero, while retaining the large coefficients as they are. Therefore, SCAD is continuous, produces sparse solutions and unbiased coefficients for large coefficients. The ncvreg function in the R package ncvreg was used for SCAD.

3.3 Elastic Net

The elastic net (Zou and Hastie, 2005) is a regularization regression technique similar to the Lasso that simultaneously does automatic variable selection and continuous shrinkage. This method combines the ℓ_1 penalty of Lasso and the ℓ_2 penalty of the ridge regression. Although both methods are shrinkage methods, the effects of the ℓ_1 and ℓ_2 penalization are quite different in practice. Applying an ℓ_2 penalty term tends to result in all small but non-zero regression coefficients, whereas applying an ℓ_1 penalty term tends to result in many regression coefficients with comparatively little shrinkage. Combining ℓ_1 and ℓ_2 penalties tends to give a result in between, with fewer regression coefficients set to zero than in a pure ℓ_1 setting, and more shrinkage of the other coefficients (Chaturvedi et al., 2012).

The naïve elastic net criterion for any fixed non-negative λ_1 and λ_2 is defined as

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1 \tag{3.5}$$

where

$$|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2 \quad \text{and} \quad |\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|.$$

The naïve elastic net estimator $\hat{\boldsymbol{\beta}}^{enet}$ is the minimizer of $L(\lambda_1, \lambda_2, \boldsymbol{\beta})$. This procedure can be viewed as a penalized least squares method. Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, then finding $\hat{\boldsymbol{\beta}}$ is equivalent to the optimization problem

$$\hat{\boldsymbol{\beta}}^{enet} = \underset{\beta}{\operatorname{argmin}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + (1 - \alpha)|\boldsymbol{\beta}|_1 + \alpha |\boldsymbol{\beta}|^2.$$

The function $(1 - \alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2$ is called the elastic net penalty which is a convex combination of the Lasso and Ridge penalty. When $\alpha = 1$, the naïve elastic net becomes Ridge regression and when $\alpha = 0$ the Lasso penalty. For all $\alpha \in [0, 1)$, the elastic net penalty function is singular (without first derivative) at 0 and is strictly convex for all $\alpha > 0$, therefore having the characteristics of both the Lasso and Ridge regression. By using the elastic net penalty, it retains the variable selection property while correcting for extra shrinkage. The glmnet function in the R package glmnet was used for Elastic Net.

3.4 Cross Validation

Cross validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model (Liu et al., 2008). This is a model evaluation technique for estimating the performance of a predictive model. One common type of cross validation for estimating a tuning parameter λ, is K-fold cross validation. In K-fold cross validation, the data is first partitioned into k roughly equal sized parts or folds. Of the k folds, one fold of the data is held out for validation (testing set) while the remaining $k-1$ folds are used as training data (learning set) to fit the model. So the learning set computes the model while the testing set assesses the test error. This process is then repeated k times with each of the k folds used only once as the validation data. By using this method, all the data in the dataset are eventually used for both training and validation. A common choice for K-fold cross validation is $k = 10$, which is the size chosen for our dataset.

For the tuning parameter, we want to choose λ that minimizes the mean squared error.

$$\text{MSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{3.6}$$

This was accomplished by using the glmnet and ncvreg package in R, that produces the value of λ that gives the minimum mean cross-validated error for each model. We fit the penalty models to the dataset using the λ that minimizes the cross-validation error.

Chapter 4

Principal Component Analysis

Before we move on to our main analysis using penalized regression, we present principal component analysis on offensive and defensive metrics. Principal Component Analysis (PCA) is a method of data reduction that transforms a set of observations of correlated variables, into a new set of uncorrelated variables, with successively smaller variances called principal components. This method is used for restating the information in an original set of variables in terms of a new set of variables where the first principal component is selected to contain as much as possible of the variability of the original set of variables, the second principal component has the second highest, and so on. For example, the first principal component, Y_1 , is

$$Y_1 = \begin{pmatrix} c_{11} c_{12} c_{13} \dots c_{1k} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \end{pmatrix} = c_{11}x_1 + c_{12}x_2 + c_{13}x_3 \dots + c_{1k}x_k.$$

The principal components are linear combinations of the original variables. We can think of $c_{11}, c_{12}, \dots, c_{1k}$ as weights for each of the original variables $x_1, x_2, \dots, x_k$. This means a larger weight would indicate a greater importance of that variable in the first principal component and a weight near zero would indicate that the variable is not important. Given k variables, we can construct k principal components, each of which will be a linear combination of the original set of variables.

There are two common ways to determine how to select the number of principal components. One way to do this is to look at the cumulative proportion of total variance which is usually taken to be around 80% or 90%. A scree plot is another way to select the number of principal components. This plot is a graph of eigenvalues λ_j of a correlation matrix against $j, j = 1, ..., k$. With the eigenvalues ordered from largest to the smallest, the number of principal components to be selected is determined at the point that the slope of the graph is steep to the left of j, but also not too steep to the right. However, interpretation of the scree plot is rather subjective since it involves selecting the number of principal components based on the visual appearance of the plot. We will see this may not be so clear when we look at the hitting metrics scree plot.

4.1 Fielding Metrics

For this analysis, we use the reduced dataset of 1,637 player-seasons which begin in 2002. The summary statistics are in Table 4.1. In Table 4.2 we have the correlation matrix for fielding. First, we analyze which variables are correlated for the fielding metrics. In this dataset, the variables are the metrics PO, A, E, DP, and FP. From Table 4.2, we see the metric putouts and fielding percentage have a positive correlation with $r = 0.557$. So as putouts increases, fielding percentage increases too. There is a strong positive correlation between errors and assists with $r = 0.686$. Also, assists and double plays have a high correlation with $r = 0.638$. This means that as assists and double plays increase so will errors. There is a strong negative correlation between errors and fielding percentage with $r = -0.787$. Since this is a negative correlation coefficient, this implies that as errors increases, fielding percentage decreases.

Table 4.1: Descriptive Statistics for Fielding Metrics 2002 - 2012

Variable	N	Mean	Std Dev	Minimum	Maximum
PO	1637	421.26	354.23	53	1596
A	1637	165.06	165.00	0	561
E	1637	8.66	6.10	0	34
DP	1637	43.45	45.35	0	175
FP	1637	0.98	0.01	0.90	1

Table 4.2: Correlation Matrix for Fielding Metrics 2002 - 2012

	PO	A	E	DP	FP
PO	1.000	-0.343	-0.236	0.361	0.557
A	-0.343	1.000	0.686	0.638	-0.404
E	-0.236	0.686	1.000	0.396	-0.787
DP	0.361	0.638	0.396	1.000	0.052
FP	0.557	-0.404	-0.787	0.052	1.000

In Table 4.3, we have the eigenvalues of the correlation matrix, the difference between the consecutive eigenvalues, the proportion of variation explained by each of the principal components and the cumulative proportion of the variation explained by each of the principal components. Eigenvalues are the variances of the principal components. Since we conducted our principal components analysis on the correlation matrix, the variables are standardized which means that the each variable has a variance of 1. If all the eigenvalues are added together, it will equal a total of 5 which is the total variance of the original variables. From Table 4.3, we examined the eigenvalues to determine how many principal components should be considered. We see that about 53% of the variation is explained by the first eigenvalue, about 32% of the variation is explained by the second eigenvalue, and so forth. The cumulative percentage is obtained by adding the successive proportions of variation explained in the proportion column. For example, $0.527 + 0.319 = 0.846$ is about 85% of the variation explained by the first two eigenvalues together and about 97% of the variation is accounted for in the the first three eigenvalues. We can conclude that two principal components explain about 85% of the total variation. However, if we used the scree plot in Figure 4.1, we would use four principal components. Since the scree plot is rather subjective, we used the proportion of total variance taken at 85%, and therefore two principal components. In Table 4.3, the variance of Y_1 values will be 2.635, the variance of Y_2 values will be 1.594, the variance of Y_3 values will be 0.605, and so on. These values are nothing but the eigenvalues.

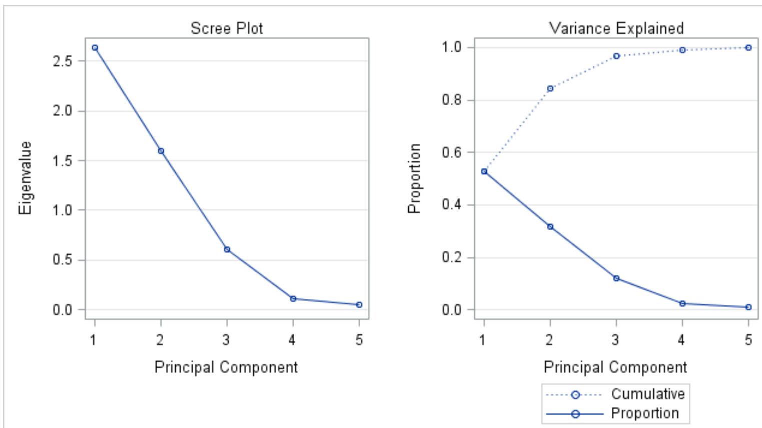

Figure 4.1: Scree Plot of Fielding Metrics 2002 - 2012

Table 4.3: Eigenvalues of the Correlation Matrix for Fielding Metrics 2002 - 2012

	Eigenvalue	Difference	Proportion	Cumulative
1	2.635	1.041	0.527	0.527
2	1.594	0.989	0.319	0.846
3	0.605	0.488	0.121	0.967
4	0.118	0.070	0.024	0.991
5	0.047		0.010	1.000

Next, we look at the eigenvectors, which are the coefficients for the individual variables in the respective principal components. Interpretation of the principal components is based on finding which variables strongly contributes to each component. From Table 4.3, we know that the first principal component accounts for about 53% of the total variation in the data. We also observe that the first eigenvector has mostly positive elements except for putout and fielding percentage. The first principal component seems to measure infielders and outfielders overall fielding performance. Given that the metrics A, E, and DP are positive, an infielder with many assists tend to have many double plays and errors as well. Meanwhile, PO and FP are decreasing for outfielders. The second principal component explains about 32% of the total variability. The second eigenvector has positive coefficients and therefore the second principal component is a measure of a fielders positive performance. We can see that the metrics PO and DP are big contributors to a fielders performance.

Table 4.4: Eigenvectors for Fielding 2002 - 2012

	Prin1	Prin2	Prin3	Prin4	Prin5
PO	-0.3035	0.5765	-0.5933	0.231407	-0.41228
A	0.5242	0.2507	0.4878	0.406263	-0.50934
E	0.5654	0.0643	-0.4351	0.436197	0.54468
DP	0.2585	0.6934	0.1321	-0.61213	0.24552
FP	-0.4967	0.3464	0.4509	0.465226	0.462212

4.2 Hitting Metrics

We will now look at the results of the principal component analysis for the hitting metrics. We would like to include all metrics, therefore we will work with the reduced dataset of 7,249 player-seasons which begin in 2002. The summary statistics for hitting are in Table 4.7. In the correlation matrix there are forty-five offensive metrics being analyzed. There are quite a few variables with high correlations so we will only mention a few. It is clear that there is a strong correlation between OBP and AVG. As a batters on-base percentage increases, so will his batting average. OPS has a high correlation with AVG, OBP, and SLG. This is reasonable since OPS is calculated by (OBP+SLG). The metrics ISO and SLG have a positive correlation which is expected since ISO is (SLG-AVG).

In Table 4.5, we have the eigenvalues of the correlation matrix for hitting. We will examine the appropriate percentage of total variation from Table 4.5 to determine how many principal components should be considered. We see that about 48% of the variation is explained by the first eigenvalue, about 10% of the variation is explained by the second eigenvalue, about 7% of the variation is explained by the third eigenvalue and so on. We can also look at the cumulative percentage. About 58% of the variation is explained by the first two eigenvalues together, about 65% of the variation is accounted for in the the first three eigenvalues, about 70% of the variation is accounted for in the first four eigenvalues and so forth. From the output, we can conclude that seven principal components explain about 79% and eight principal components explain about 81% of the total variation. Since both are roughly close to about 80% of the variation we will select seven principal components. However, if we used the scree plot in Figure 4.2 to determine the number of principal components, we would maybe use ten. By looking at the plot, one can see this is simply a rough estimate. In Table 4.5, the variance of Y_1 values will be 21.465, the variance of Y_2 values will be 4.695, the variance of Y_3 values will be 3.066, and so on. If all of the eigenvalues are added together, it will equal a total of 45 which is the total variance of the original variables.

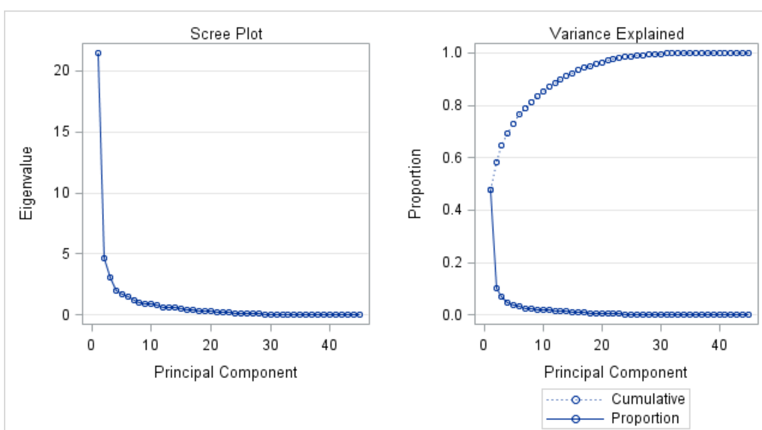

Figure 4.2: Scree Plot of Hitting Metrics 2002 - 2012

Now we look at the eigenvectors for the coefficients of the principal components. The first principal component explains about 48% of the the total variation and thus may represent how a hitter performs at the plate such as number of hits and runs. The second principal component explains about 10% of the the total variation and seems to measure a hitter's sacrifice bunts (SAC) and decreasing isolated power (ISO) and home runs to fly ball ratio (HR/FB). The third principal component may represent a hitter's batting average for balls in play (BABIP) and his speed (SPD). The fourth principal component seems to measure a player's ground ball to fly ball ratio (GB/FB) and his fly ball rate (FB%). The fifth principal component may represent as a player's strikeout ability (K%, BB/K). The sixth principal component measures a player's plate discipline (BB%). The seventh principal component seems to measures a player's infield hit rate (IFH%), infield fly ball rate (IFFB%), and his line drive rate (LD%).

Table 4.5: Eigenvalues of the Correlation Matrix for Hitting Metrics 2002 - 2012

	Eigenvalue	Difference	Proportion	Cumulative
1	21.465	16.769	0.477	0.477
2	4.695	1.629	0.104	0.581
3	3.066	1.036	0.068	0.650
4	2.030	0.367	0.045	0.695
5	1.663	0.142	0.037	0.732
6	1.521	0.359	0.034	0.765
7	1.161	0.171	0.026	0.791
8	0.990	0.028	0.022	0.813
9	0.963	0.077	0.021	0.835
10	0.886	0.102	0.020	0.854
11	0.784	0.140	0.017	0.872
12	0.644	0.028	0.014	0.886
13	0.616	0.038	0.014	0.900

Table 4.6: Eigenvectors for Hitting 2002 - 2012

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
AB	0.2043	0.1163	-0.0776	-0.0218	-0.0011	-0.0706	-0.0154	-0.0203
PA	0.2057	0.1082	-0.0836	-0.0120	-0.0107	-0.0443	-0.0263	-0.0142
H	0.2070	0.1022	-0.0545	0.0208	0.0010	-0.0726	-0.0072	-0.0227
1B	0.1975	0.1449	-0.0064	0.0250	-0.0445	-0.1120	-0.0045	-0.0362
2B	0.1993	0.0562	-0.0949	0.0118	0.0092	-0.0858	-0.0234	-0.0215
3B	0.1220	0.1757	0.1668	-0.0813	0.1105	0.1360	-0.0651	-0.1106
HR	0.1787	-0.0738	-0.2206	0.0289	0.1496	0.0914	0.0195	0.0607
R	0.2065	0.0835	-0.0605	0.0127	0.0355	0.0483	-0.0190	-0.0093
RBI	0.2003	0.0021	-0.1675	0.0342	0.0539	-0.0240	0.0025	0.0080
BB	0.1889	0.0002	-0.1196	0.0746	-0.0884	0.1921	-0.1049	0.0124
IBB	0.1215	-0.0659	-0.1520	0.1587	-0.1112	0.2017	-0.0694	0.1043
SO	0.1768	0.0499	-0.1361	-0.0345	0.1994	0.0509	-0.1184	0.0110
HBP	0.1351	0.0311	-0.0784	-0.0240	0.0317	-0.0309	0.0450	0.0695
SF	0.1610	0.0380	-0.1333	-0.0358	-0.0538	-0.1249	-0.0156	-0.0617
SAC	-0.0268	0.2765	0.0721	0.0058	-0.0085	-0.0441	-0.1215	0.1580
GDP	0.1678	0.0530	-0.1446	0.0648	-0.0813	-0.2027	0.0588	-0.0492
SB	0.1118	0.2338	0.2073	-0.0931	0.0789	0.2412	-0.0550	-0.0583
CS	0.1213	0.2195	0.1764	-0.0834	0.0551	0.1662	-0.0671	-0.0325
BUH	0.0515	0.2466	0.2575	-0.1174	0.0232	0.1771	-0.1238	0.2304
AVG	0.1589	-0.1347	0.2771	0.1164	-0.0069	-0.1999	0.0845	0.0312
OBP	0.1650	-0.1808	0.2311	0.1171	-0.1309	0.0008	0.0299	0.0194
SLG	0.1701	-0.2210	0.1303	0.0347	0.1460	-0.0287	0.0902	0.0428
OPS	0.1744	-0.2134	0.1749	0.0685	0.0424	-0.0181	0.0698	0.0352
ISO	0.1442	-0.2491	-0.0228	-0.0428	0.2470	0.1242	0.0763	0.0439
BABIP	0.1043	-0.1201	0.3210	0.1874	0.1508	-0.2323	-0.0716	0.0632
FB	0.2007	0.0609	-0.1258	-0.0908	-0.0320	-0.0770	0.0123	-0.0168
LD	0.1992	0.1108	-0.0466	0.0160	-0.0560	-0.1350	-0.0596	-0.0085
GB	0.1917	0.1661	-0.0265	0.0351	-0.0551	-0.0993	0.0444	-0.0539
IFFB	0.1605	0.0596	-0.1239	-0.2069	-0.0663	-0.0985	0.1616	0.1835
IFH	0.1583	0.1987	0.0882	-0.0211	0.0396	-0.0318	0.2012	-0.1263
GB/FB	-0.0767	0.1840	-0.0599	0.4369	0.0807	0.0648	0.1433	0.1281
BB/K	0.1115	-0.0875	0.0720	0.1111	-0.5283	0.2397	0.0295	-0.0230
BB%	0.0985	-0.1767	0.0110	0.0830	-0.3105	0.4005	-0.1145	-0.0202
K%	-0.1127	-0.0016	-0.1687	0.0189	0.4289	0.1418	-0.2266	0.0586
LD%	0.0536	-0.0782	0.1900	0.1125	-0.0067	-0.3409	-0.4669	0.2243
GB%	-0.0915	0.2443	-0.0365	0.4188	0.0416	0.1164	0.2652	0.0023
FB%	0.0693	-0.2216	-0.0673	-0.5186	-0.0414	0.0653	-0.0258	-0.1286
IFFB%	-0.0179	0.0116	-0.0469	-0.2146	-0.1029	-0.0231	0.4054	0.7066
HR/FB	0.0997	-0.2246	-0.0767	0.0508	0.3445	0.1654	0.0835	0.1079
IFH%	0.0238	0.0099	0.2016	-0.0651	0.1436	-0.0059	0.4994	-0.2483
BUH%	0.0477	0.1087	0.1708	-0.0885	0.0421	0.1804	-0.1185	0.3679
wOBA	0.1728	-0.2097	0.1925	0.0717	0.0093	-0.0145	0.0658	0.0269
wRC	0.2086	0.0232	-0.0925	0.0672	0.0231	0.0430	-0.0239	0.0045
wRAA	0.1380	-0.1622	-0.0761	0.2084	0.0917	0.2135	-0.0132	0.0374
Spd	0.0824	0.1244	0.2977	-0.1328	0.1067	0.2042	0.0368	-0.1239

Table 4.7: Descriptive Statistics for Hitting Regular Season 2002-2012

Variable	N	Mean	Std Dev	Minimum	Maximum
AB	7249	250.33	199.89	7	716
PA	7249	281.13	224.71	10	778
H	7249	65.95	58.26	0	262
1B	7249	43.71	38.54	0	225
2B	7249	13.26	12.48	0	56
3B	7249	1.38	2.12	0	23
HR	7249	7.59	9.51	0	58
R	7249	33.80	31.74	0	143
RBI	7249	32.22	31.68	0	156
BB	7249	23.89	24.43	0	232
IBB	7249	1.94	3.89	0	120
SO	7249	48.89	38.65	0	223
HBP	7249	2.56	3.45	0	30
SF	7249	2.03	2.34	0	16
SAC	7249	2.29	3.10	0	24
GDP	7249	5.74	5.75	0	32
SB	7249	4.32	8.03	0	78
CS	7249	1.70	2.62	0	24
BUH	7249	1.02	2.39	0	38
AVG	7249	0.23	0.07	0	0.56
OBP	7249	0.30	0.08	0	0.61
SLG	7249	0.36	0.13	0	0.93
OPS	7249	0.65	0.20	0	1.42
ISO	7249	0.13	0.08	0	0.59
BABIP	7249	0.28	0.07	0	0.67
FB	7249	72.79	64.17	0	282
LD	7249	40.19	35.43	0	160
GB	7249	88.24	74.36	0	405
IFFB	7249	7.65	7.94	0	65
IFH	7249	5.42	5.88	0	57
GB/FB	7249	1.73	1.92	0	52
BB/K	7249	0.44	0.32	0	5.66
BB%	7249	0.07	0.04	0	0.38
K%	7249	0.21	0.09	0	0.77
LD%	7249	0.19	0.06	0	1
GB%	7249	0.46	0.11	0	1
FB%	7249	0.34	0.10	0	1
IFFB%	7249	0.11	0.10	0	1
HR/FB	7249	0.08	0.08	0	1
IFH%	7249	0.06	0.05	0	1
BUH%	7249	0.14	0.24	0	1
wOBA	7249	0.29	0.08	0	0.57
wRC	7249	33.85	35.08	-10	184
wRAA	7249	0.18	12.80	-38	110
Spd	7249	3.39	2.03	0	10

Chapter 5

Penalized Regression Analysis

To evaluate the offensive and defensive metrics, we examined the signal, or the fraction of players who differ from the overall league mean. When this fraction is high, a metric contains a large amount of information and will have a greater predictive power than the league mean. We calculated the proportion of players that were fitted with non-zero coefficients by the Lasso, Elastic Net and SCAD penalties. Each component of the $\hat{\beta}^{Lasso}, \hat{\beta}^{enet}$, and $\hat{\beta}^{SCAD}$ vector corresponds to the individual coefficient of a given player, and we are interested in which of these individual coefficients are fitted to be different from zero. The trend in the three penalized regression models is SCAD provides lower fractions because it tends to choose a smaller-sized model compared to the other two models. SCAD sets more coefficients to zero while the Lasso and Elastic Net regression models have a large percentage of non-zero coefficients. The previous paper by McShane et. al (2011) only looked at the fraction using the Lasso but we used other criteria and penalty functions to identity which metrics stand out. For our analysis, we looked at the absolute mean and standard deviation as well as the signal, to determine which metrics are consistent measures of a players performance. We calculated the three criteria as follows: Let p be the number of players.

$$\begin{aligned} \text{Signal} &= \frac{\#\text{ of nonzero } \hat{\beta}_j}{p} && (5.1) \\ \text{Mean} &= \frac{1}{p}\sum_{j=1}^{p}|\hat{\beta}_j| && (5.2) \\ \text{S.D.} &= \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \bar{\hat{\beta}})^2 && (5.3) \end{aligned}$$

where $\bar{\hat{\beta}}$ is the sample mean of the estimated coefficients. The absolute mean of the regression coefficients measures the magnitude of the nonzero coefficients for each model. The larger the absolute mean, the more information on the given metric. The standard deviation of the three penalized regression models tells us how far the estimates coefficients are spread from the mean. If the standard deviation is low, the players are very close to the coefficients and each other. Hence, the offensive or defensive metric would not be unique. If the standard deviation is high, this indicates that the players are spread out over a large range of values and therefore a very informative metric. In our analysis, all the variables are standardized for direct comparison.

5.1 Defensive Metrics

We used the fraction values of three different penalized methods, the absolute mean, and standard deviation to determine which defensive measures stand out. In Figure 5.1, we have a plot of the absolute mean versus the signal for the three different penalized models for each defensive metric. We see the metrics A and PO have a high signal and mean while FP has a high signal but low mean. In Figure 5.2, we have a plot of the standard deviation versus signal. We see once again that the metrics PO and A have a high signal and standard deviation. Finally, in Figure 5.3, we see the metrics PO and A have a relatively high mean and standard deviation.

The values plotted in these graphs can be found in Table 5.1. Throughout the plots, the metrics PO and A are consistent in all three criteria. For the Lasso regression model for PO, the signal is 0.82, the mean is 180.244 and the S.D. is 267.82. For the Lasso regression for A, the signal is 0.964, the mean is 130.035, and the S.D. is 114.924. The metric DP has high signal but the mean and standard deviation do not have high values like the two previously mentioned metrics. The metric E and FP have high signals but small means and standard deviations. Both of the metrics PO and A are related to retiring a batter which is the objective of the defense. A fielder is credited with a putout whenever he puts out a batter or tags out a runner. Also, a fielder is credited with an assist any time he throws or deflects a batted or thrown ball in such a way that a putout results. (MLB.com). The best metrics to determine defensive ability are PO and A.

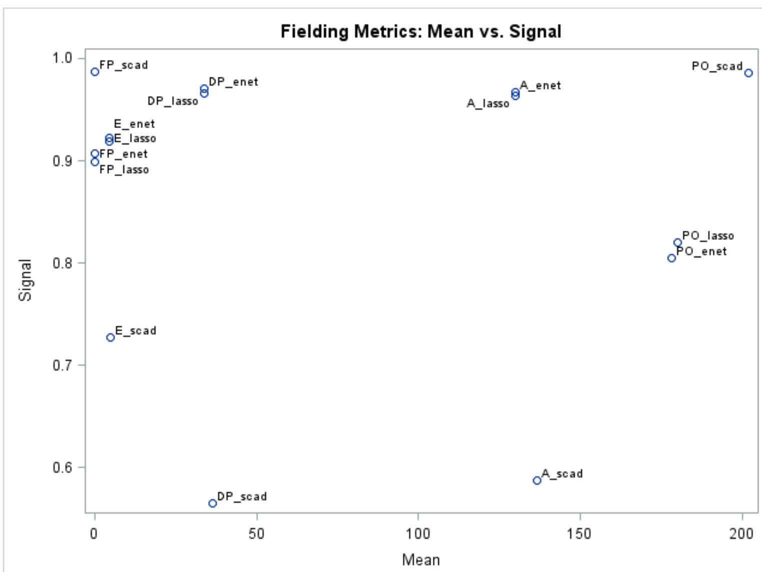

Figure 5.1: Fielding Metrics - Mean vs. Signal

In Table 5.1 we have the defensive model statistics. The first row of each metric is the proportion of players who differ from the overall mean for each penalty, the second row is the absolute mean of the regression coefficients, and the third row is the standard deviation.

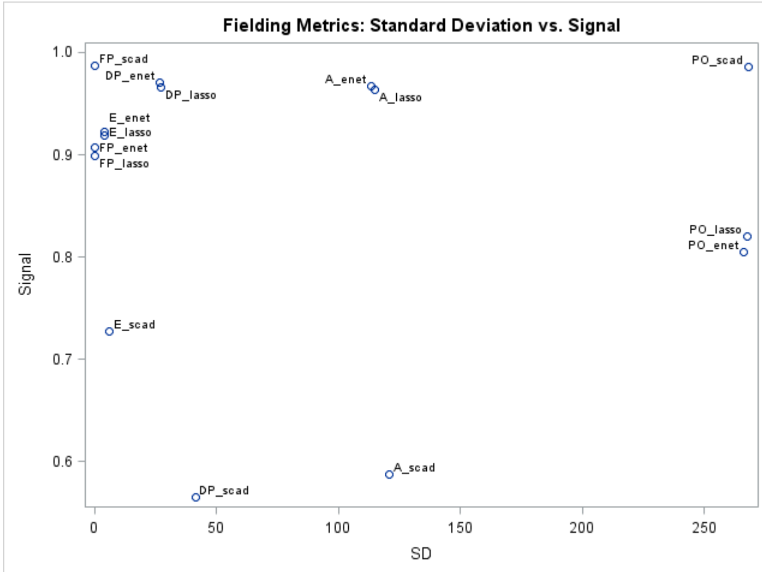

Figure 5.2: Fielding Metrics - S.D. vs. Signal

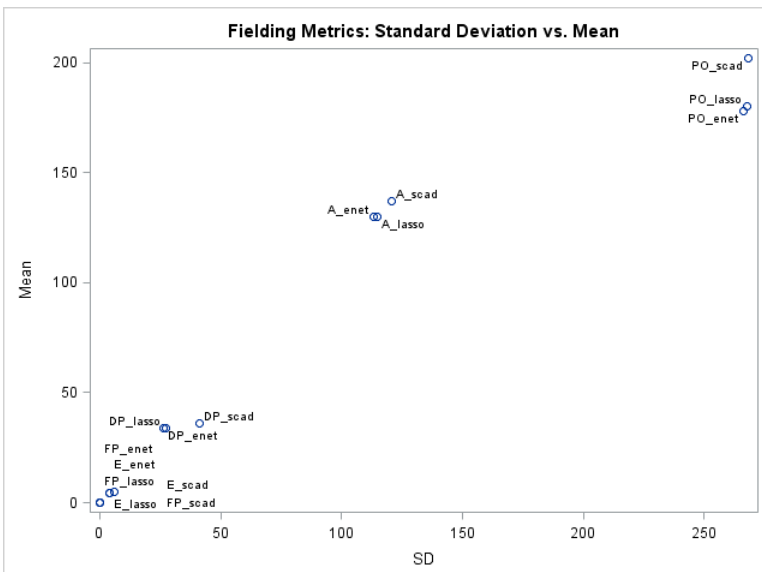

Figure 5.3: Fielding Metrics - S.D. vs. Mean

5.2 Offensive Metrics

In Figure 5.4, we have a plot of the absolute mean versus signal for offensive metrics. We see from this plot that the metrics H, R, BB, RBI, 1B, SO and wRC have a high mean and signal.

Table 5.1: Defensive Model Statistics

Metric		Lasso	Elastic Net	SCAD
PO	Signal	0.820	0.805	0.986
	Mean	180.244	178.078	201.977
	S.D.	267.820	266.298	268.161
A	Signal	0.964	0.967	0.588
	Mean	130.035	130.000	136.817
	S.D.	114.924	113.174	120.678
E	Signal	0.922	0.919	0.727
	Mean	4.345	4.336	5.022
	S.D.	4.239	4.210	6.051
DP	Signal	0.966	0.970	0.565
	Mean	33.854	33.664	36.293
	S.D.	27.194	26.512	41.204
FP	Signal	0.899	0.907	0.987
	Mean	0.0088	0.0090	0.0108
	S.D.	0.0099	0.0098	0.030

The metrics K%, BB/K, SPD, and HR have high signal but small means. In Figure 5.5, we have a plot of the standard deviation versus signal. Here we see the metrics H, R, BB, 1B, SO, and wRC have a high S.D. and signal. In Figure 5.6, we have a plot of the standard deviation versus the absolute mean. We see the metrics H, 1B, R, RBI, BB, SO, and wRC have a high S.D. and mean. There are seven metrics that are consistent throughout the plots for our three criteria: BB, H, RBI, R, SO, 1B, and wRC. The metrics BB and SO are related to plate discipline, H measures hitting power, the metrics R, RBI and 1B measures runs scored, and the metric wRC relates to a player's run contribution to his team. The metric wRC is an improved version of Bill James' Runs Created (RC) statistic, which attempted to quantify a player's total offensive value and measure it by runs (Fangrahs.com). For the Lasso regression for the metric BB, the signal is 0.943, the mean is 10.989 and the S.D. is 10.972. For the Elastic Net regression for H, the signal is 0.948, the mean is 29.551 and the S.D. is 22.861. For the Elastic Net regression for R, the signal is 0.941, the mean is 14.98 and the S.D. is 12.938. For the SCAD regression for RBI, the signal is 0.992, the mean is 18.344 and the S.D. is 21.923. For the SCAD regression for 1B, the signal is 0.989, the mean is 26.982 and the S.D. is 28.2. For the Lasso regression for SO, the signal is 0.911, the mean is 15.958 and the S.D. is 16.36. Lastly, for the SCAD regression for wRC, the signal is 0.99, the mean is 20 and the S.D. is 23.364. The offensive model statistics are found in Table 5.2 and Table 5.3.

McShane et. al (2011) identified a set of nine metrics as high signals metrics: K/PA, GB/BIP, HR/FB, SPD, HR/PA, BB/PA, and GB/BIP. The authors decided to further reduce this to five metrics because (i) K/PA and K measures strikeouts, (ii) HR/FB, HR/PA and ISO measures power hitting, and (iii) GB/BIP and FB/BIP measure the tendency to hit grounders as opposed to fly balls. Therefore, their set of five offensive metrics are K/PA, SPD, ISO, BB/PA and GB/BIP. In our analysis, the four metrics K%, Spd, ISO and BB% are high signal in terms of the fraction of players who differ from the overall league mean. If we only used the fraction, we have similar results to McShane et. al (2011), but since we also consider the absolute mean and standard deviation, we have different results.

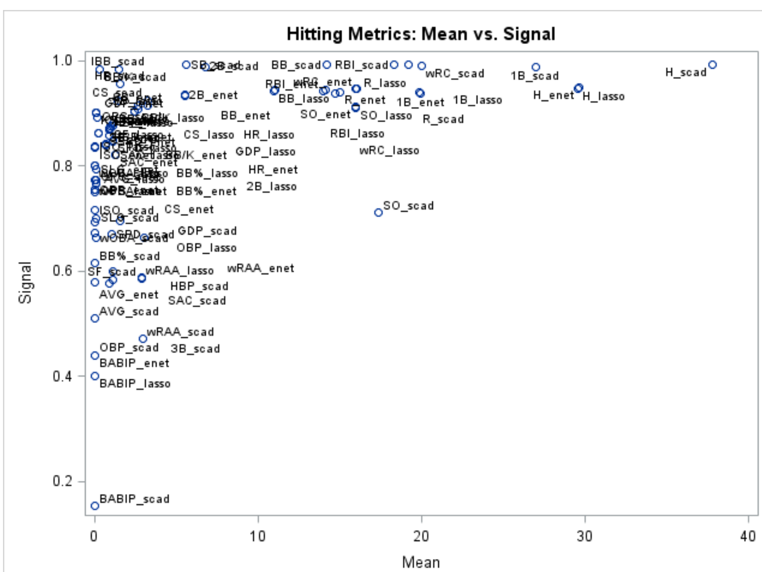

Figure 5.4: Hitting Metrics - Mean vs. Signal

Table 5.2: Offensive Model Statistics

Metric		Lasso	Elastic Net	SCAD
AB	Signal	0.942	0.949	0.986
	Mean	104.540	106.560	133.920
	S.D.	75.860	75.210	145.132
AVG	Signal	0.674	0.693	0.511
	Mean	0.032	0.033	0.031
	S.D.	0.043	0.044	0.056
BABIP	Signal	0.402	0.439	0.154
	Mean	0.016	0.017	0.006
	S.D.	0.035	0.040	0.025
BB	Signal	0.943	0.944	0.992
	Mean	10.989	11.021	14.192
	S.D.	10.972	10.898	16.696
BB/K	Signal	0.863	0.863	0.984
	Mean	0.203	0.203	0.267
	S.D.	0.246	0.246	0.361
BB%	Signal	0.801	0.800	0.616
	Mean	0.021	0.021	0.017
	S.D.	0.022	0.022	0.030
CS	Signal	0.881	0.876	0.924
	Mean	1.075	1.077	1.383
	S.D.	1.595	1.588	2.072
GDP	Signal	0.903	0.903	0.663
	Mean	2.423	2.427	3.004
	S.D.	2.230	2.217	3.721
H	Signal	0.948	0.949	0.993
	Mean	29.551	29.641	37.815
	S.D.	22.861	22.671	41.182
HBP	Signal	0.869	0.869	0.601
	Mean	0.904	0.904	1.113
	S.D.	1.458	1.452	1.814
HR	Signal	0.916	0.924	0.956
	Mean	3.233	3.240	1.567
	S.D.	4.608	4.585	5.712
IBB	Signal	0.873	0.876	0.983
	Mean	0.936	0.938	1.456
	S.D.	1.627	1.620	1.972
ISO	Signal	0.837	0.836	0.716
	Mean	0.038	0.038	0.038
	S.D.	0.034	0.034	0.054
K%	Signal	0.901	0.902	0.901
	Mean	0.069	0.070	0.078
	S.D.	0.077	0.077	0.106
OBP	Signal	0.756	0.755	0.579
	Mean	0.043	0.043	0.042
	S.D.	0.054	0.053	0.070
OPS	Signal	0.766	0.766	0.893
	Mean	0.104	0.103	0.138
	S.D.	0.123	0.123	0.183

Table 5.3: Offensive Model Statistics

Metric		Lasso	Elastic Net	SCAD
PA	Signal	0.950	0.950	0.991
	Mean	117.371	119.620	155.237
	S.D.	85.692	84.934	162.869
R	Signal	0.938	0.941	0.993
	Mean	14.728	14.980	19.227
	S.D.	13.045	12.938	21.470
RBI	Signal	0.945	0.943	0.992
	Mean	14.117	13.972	18.344
	S.D.	13.521	13.601	21.923
SAC	Signal	0.821	0.821	0.696
	Mean	1.294	1.292	1.582
	S.D.	1.607	1.600	2.252
SB	Signal	0.909	0.916	0.992
	Mean	2.629	2.632	5.576
	S.D.	5.222	5.211	5.716
SF	Signal	0.859	0.872	0.583
	Mean	0.867	0.894	1.104
	S.D.	0.872	0.868	1.450
SLG	Signal	0.773	0.794	0.701
	Mean	0.065	0.067	0.071
	S.D.	0.071	0.072	0.107
SO	Signal	0.911	0.914	0.712
	Mean	15.958	15.987	17.374
	S.D.	16.360	16.255	24.688
SPD	Signal	0.849	0.846	0.670
	Mean	1.087	1.085	1.040
	S.D.	0.972	0.969	1.511
wOBA	Signal	0.774	0.751	0.663
	Mean	0.044	0.043	0.048
	S.D.	0.054	0.053	0.074
wRAA	Signal	0.588	0.587	0.471
	Mean	2.864	2.861	2.916
	S.D.	5.218	5.204	6.352
wRC	Signal	0.948	0.948	0.990
	Mean	15.996	16.047	20.005
	S.D.	14.604	14.492	23.364
1B	Signal	0.941	0.939	0.989
	Mean	19.860	19.910	26.982
	S.D.	15.488	15.372	28.199
2B	Signal	0.934	0.935	0.989
	Mean	5.515	5.530	6.745
	S.D.	4.727	4.690	7.985
3B	Signal	0.842	0.840	0.576
	Mean	0.655	0.656	0.859
	S.D.	0.957	0.952	1.297

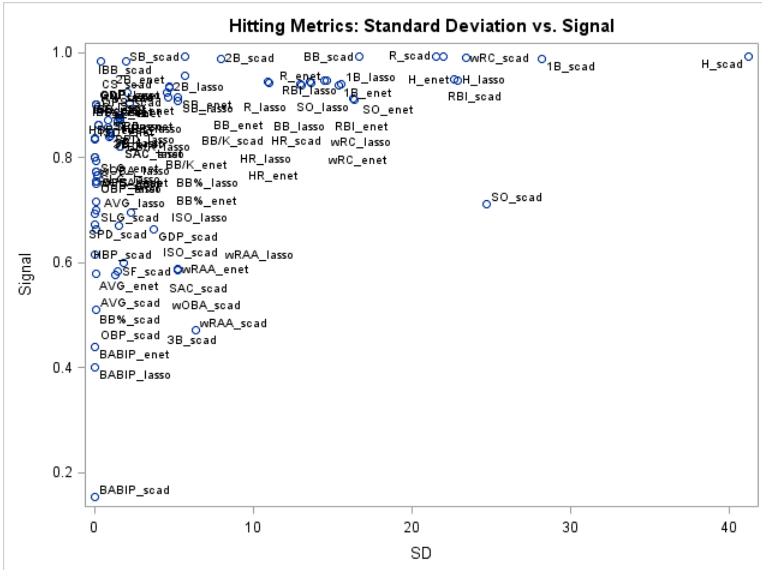

Figure 5.5: Hitting Metrics - S.D. vs. Signal

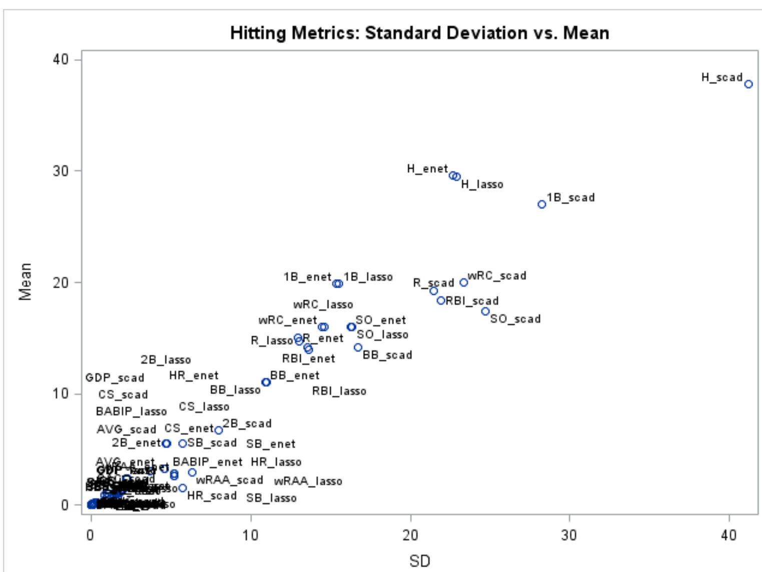

Figure 5.6: Hitting Metrics - S.D. vs. Mean

Chapter 6

Summary

We discussed penalized regression methods that allow us to determine which offensive and defensive metrics are useful to distinguish players over time. A good metric is one that provides a consistent measure of player ability, while a poor metric is indistinguishable from player to player. A high signal metric has a large fraction of players who differ from the overall mean, while a low signal metric performs close to the overall mean. Our principal component analysis suggests, only a small subset of offensive metrics are needed to obtain much of the information contained in the full set of forty-five metrics. Also, with this analysis we reduced the five defensive metrics to two principal components. We evaluate each of the metrics utilizing three criteria. We used the Lasso, Elastic Net, and SCAD regression models, the absolute mean of the regression coefficients for the given penalty, and the standard deviation.

For five defensive metrics, we identified two metrics which stand out has unique metrics. These metrics had a large fraction of players that differ from the overall mean, a large absolute mean and a large standard deviation. A number of the thirty-one offensive metrics demonstrate some degree of signal but we identify seven metrics which stand out throughout all three criteria. This set of seven, provides a substantial reduction in the dimensionality for hitting metrics.

In future research, the actual prediction of each metric would be valuable information. We would look at a player and analyze ten player seasons to predict the next player season. For our particular dataset, we withhold the 2012 season values for each player, estimate the models using the values from the 2002-2011 seasons, and then used the predicted models to forecast the 2012 season values.

Bibliography

[1] Breheny, Patrick. (2013). Regularization paths for SCAD- and MCP-penalized regression models. R Package. Version 3.1-0.

[2] Chaturvedi, Nimisha, Jelle Goeman, and Rosa Meijer. (2012). "L1 and L2 Penalized Regression Models," February 14, 2014. http://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf

[3] Davison, Drew. (2013). "Advanced baseball statistical analysis: Game-changer or fuzzy math?" March 23, 2014. http://www.star-telegram.com/2013/05/11/4843118/advanced-baseball-statistical.html

[4] Eder, Steve. (2013)." Era of Modern Baseball Stats Brings WAR to Booth." The New York Times, March 22, 2014.

http://www.nytimes.com/2013/04/02/sports/baseball/baseball-broadcasts-introduce-advanced-statistics-but-with-caution.html?_r=0

[5] Fangraphs.com. September 18, 2013. www.fangraphs.com

[6] Fan, J. and R.Li, (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, 96, 1348-1360.

[7] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani.(2013). Lasso and elastic-net regularized generalized linear models. R Package. Version 1.9-5.

[8] Hoerl, A.E. and R.W Kennard. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, 12, 55-67.

[9] Liu, H., P. Refaeilzadeh, and L. Tang (2008): "Cross-Validation," November 5, 2013.

`http://www.cse.iitb.ac.in/~tarung/smt/papers_ppt/ency-cross-validation.pdf`

[10] McShane, Blakeley B., Alexander Braunstein, James Piette, and Shane T. Jensen.(2011). "A Hierarchical Bayesian Variable Selection Approach to Major League Baseball Hitting Metrics". *Journal of Quantitative Analysis in Sports*, 7, Iss.4, Article 2.

[11] MLB.com, January 16, 2014.

`http://mlb.mlb.com/mlb/official_info/baseball_basics/abbreviations.jsp`

[12] Neal, Dan, James Tan, Feng Hao, and Samuel S. Wu. (2010). "Simply Better: Using Regression Models to Estimate Major League Batting Averages," *Journal of Quantitative Analysis in Sports*, 6, Iss.3, Article 12.

[13] Yang, Tae Young and Tim Swartz. (2004). A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball. *Journal of Data Science*, 2, 61-73.

[14] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, B*, 58, 267-288.

[15] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, B*, 67, 301-320.