# 126 Data Project, Step 4

Sam Ream, Valeria Lopez, Skyler Yee

## Introduction

Using the "History of Baseball" data set, we analyzed how our predictors (singles, doubles, triples, home runs, walks, intentional walks, hit by pitches, stolen bases, BMI, and batting hand) affected the runs scores by individual players. We sampled player statistics randomly from games played between 2000-2015, which allowed us to get an accurate representation of the population of all players who played between 2000 and 2015. Using both Ridge Regression and LASSO, we shrunk the size of some predictors to obtain estimates with smaller variance for higher precision.

## Colinearity

**Correlation Table**

|  | T | HR | S | W | D | IW | SB | HBP |
|---|---|---|---|---|---|---|---|---|
| TRIPLE (T)) | 1 | 0.47 | 0.76 | 0.59 | 0.68 | 0.45 | 0.8 | 0.49 |
| HOME_RUNS (HR) | 0.47 | 1 | 0.79 | 0.88 | 0.89 | 0.78 | 0.36 | 0.72 |
| SINGLES (S) | 0.76 | 0.79 | 1 | 0.85 | 0.95 | 0.72 | 0.68 | 0.74 |
| WALKS (W) | 0.59 | 0.88 | 0.85 | 1 | 0.9 | 0.8 | 0.5 | 0.71 |
| DOUBLE (D) | 0.68 | 0.89 | 0.95 | 0.9 | 1 | 0.74 | 0.55 | 0.75 |
| INT_WALKS (IW) | 0.45 | 0.78 | 0.72 | 0.8 | 0.74 | 1 | 0.37 | 0.58 |
| STOLEN_BASES (SB) | 0.8 | 0.36 | 0.68 | 0.5 | 0.55 | 0.37 | 1 | 0.4 |
| HIT_BY_PITCH (HBP) | 0.49 | 0.72 | 0.74 | 0.71 | 0.75 | 0.58 | 0.4 | 1 |

**Variance Inflation factor Table**

```
## (Intercept)        TRIPLE    HOME_RUNS      SINGLES        WALKS       DOUBLE
##    2.000005      3.838656     8.727293    21.457634     7.646515    28.534174
##   INT_WALKS STOLEN_BASES HIT_BY_PITCH
##    3.297618      3.437006     2.498759
```

## Analysis

When looking at our complete set of predictors, At_Bat, Doubles, and Triples are highly correlated to one another and also have the largest effect on the variance of our model by a significant margin. As a result, we concluded that they can not be included in the model of all predictors unless we are willing to sacrifice the fit and predictive accuracy of the model. If we wanted to train a model that uses all of the predictors, we would want to use a shrinkage method such as a Ridge or Lasso regression.

# Ridge Regression
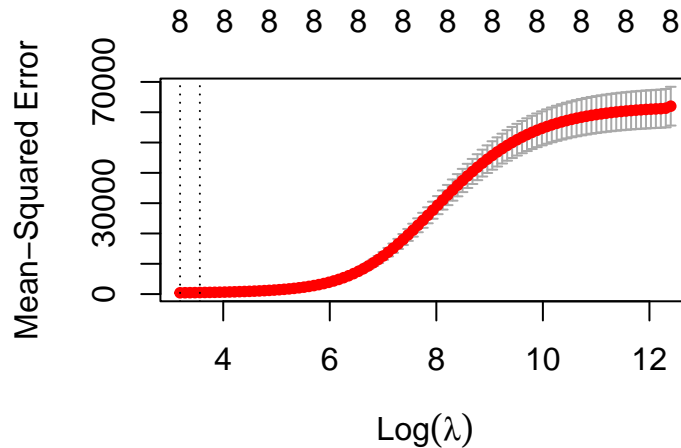
**Optimal Lambda - Ridge Regression**



Figure 1: The relationship between MSE and Log(lambda)

We found that the MSE was minimized when $\lambda$ is equal to: 24.2822337

## Model Analysis

**R-Squared Analysis**    When Lambda equals 24.2822337, the R-Squared is 0.9934422. This implies that the model explains approximately 99.34% of the variation in the response in our training data set.

**Coefficient Analysis**

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                    s1
## (Intercept) 3.7553556
## TRIPLE      1.9420330
## HOME_RUNS   0.6668767
## SINGLES     0.1773687
## WALKS       0.2396912
## DOUBLE      0.5568166
## INT_WALKS   0.3818829
## STOLEN_BASES 0.4693847
## HIT_BY_PITCH 0.6851189
```

In observing our coefficients, we can see that Triples have the largest effect (an increase of 1.6434 expected runs per Triple) on the expected number of Runs and that every predictor contributes some information to the model.

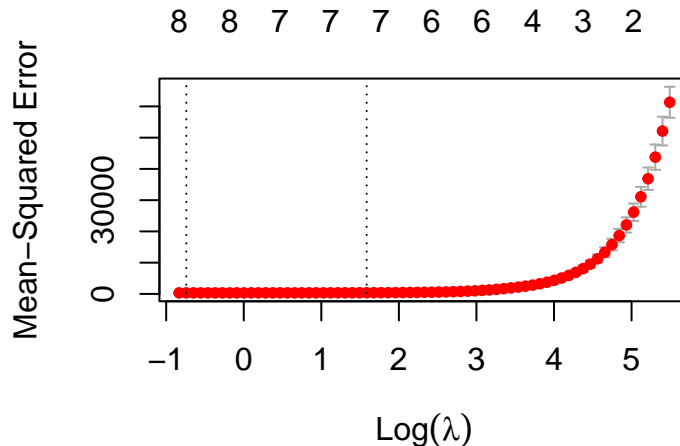# Lasso Regression

**Optimal Lambda - Lasso Regression**



Figure 2: The relationship between MSE and Log(lambda)

We found that the MSE was minimized when $\lambda$ is equal to: 0.4766701

## Model Analysis

**R-Squared Analysis**   When Lambda equals `best_lambda`, the R-Squared is 0.9954788. This implies that the model explains approximately 99.55% of the variation in the response in our training data set.

**Coefficient Analysis**

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                     s0
## (Intercept)  -2.9213599
## TRIPLE         1.1578347
## HOME_RUNS      0.8031178
## SINGLES        0.2778234
## WALKS          0.2594099
## DOUBLE         0.4598648
## INT_WALKS     -0.2650777
## STOLEN_BASES  0.4176051
## HIT_BY_PITCH  0.2236940
```

Through observing our coefficients, we can observe that Triples and Home Runs have a much larger effect on the expected number of Runs than any other predictor in a lasso regression model. Intentional walks also decrease the expected number of runs and At Bats can be removed without consequence. Since a lasso regression punishes an increase in predictors more harshly, this is likely why At Bats were removed from the model in the lasso regression while they were left in (with a very small value) in the Ridge Regression.
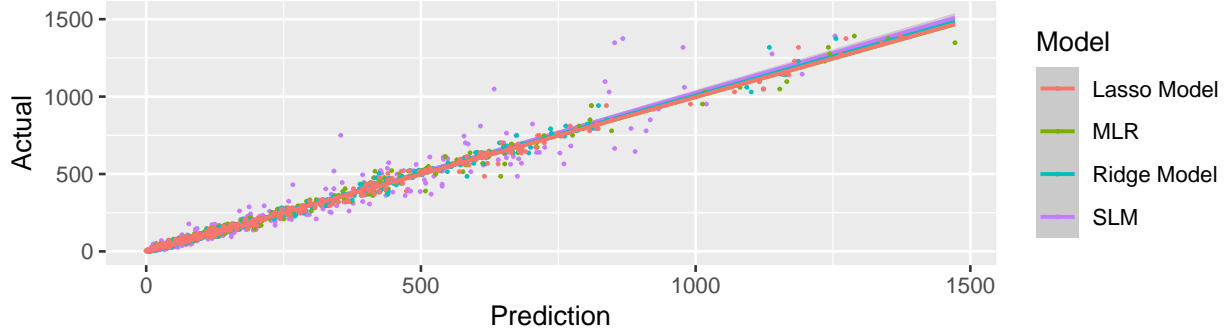
3

# Comparison of our Models



Figure 3: A comparison of the Linear Models

## MSE and $R^2$ By model

| Model Type | $R^2$ | MSE |
|---|---|---|
| SLR | 0.9393236 | 4125.3574196 |
| MLR | 0.9919045 | 550.4106253 |
| Ridge Regression | 0.9937332 | 426.0792005 |
| Lasso Regression | 0.9943709 | 382.7208519 |

The graph and table above were generated for several different sets of 500 new random observations from our original dataset. In each case, the patterns displayed were consistent with those show above. The Lasso Regression has the lowest Mean Square Error and the Highest $R^2$ of the models–with the Ridge Regression close behind. On the other hand, the Singular linear regression has the lowest $R^2$ and MSE. The Multi-Linear Regression has a slightly lower $R^2$ and MSE than the Lasso and Ridge Regressions. However, it is very close to the shrinkage models and fits the data well.

# Investigation - Principle Component Analysis

We chose Principal Components Analysis for our innovation because this method is used when there are a large number of predictors. The goal of this method is to replace our predictors with a smaller number of linear combinations of the predictors. We are essentially transforming our data into a lower-dimensional space while collating highly correlated variables together, allowing us to more easily understand and visualize our data. For example if we have $X_1, X_2, ..., X_k$ predictors with k being large or at least $k \geq 2$, we want to replace k with $k_0 < k$ linear combinations of our predictors.

Let $\mathbf{X}' = (X_1, X_2, ...X_k)$ and $\mathbf{u}'$ be a $p \times 1$ vector of constants such that $\mathbf{u}'_1 \mathbf{u}_1 = 1$. The first principal component will be the linear combination $Z_1 = \mathbf{u}'\mathbf{X}$ such that the variance of $Z_1 = \mathbf{u}'_1 \text{Var}(\mathbf{X})\mathbf{u}_1$ is as large as possible to retain as much as the variation in the predictors as possible. If $\text{Var}(\mathbf{X})$ is known, then $\mathbf{u}_i$'s are the eigenvectors that corresponds to the $k_0$ largest eigenvalues of $\text{Var}(\mathbf{X})$. If $\text{Var}(\mathbf{X})$ is unknown, like in our case, we replace the variance matrix with the sample covariance matrix.

First we normalize the data by dividing by the sample standard deviation. Then, we use the sample correlation matrix to compute our eigenvalues $(\hat{\lambda}_j)$ and select the values with the highest variances. Now, we compute the corresponding eigenvectors $(\hat{\mathbf{u}}_j)$ and multiply them by $\mathbf{X}'$. These are our principal components $Z_j$.
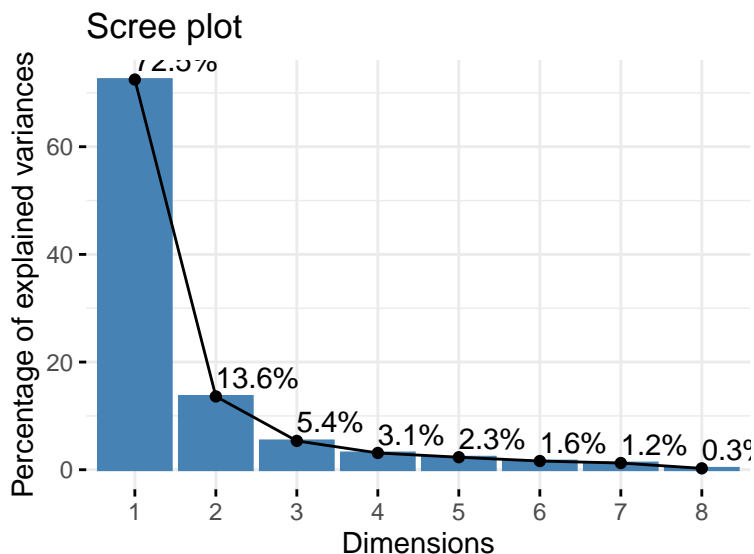
**Visualizations**



Figure 4: Component contributions to the total variance

This plot displays the eigenvalues and shows us that the first two components contribute the most the total variance.
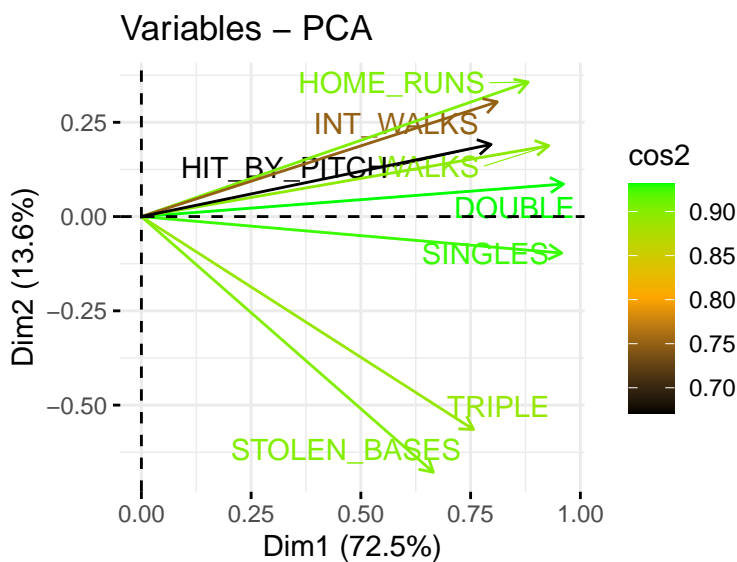


Figure 5: Biplot combined with cos2

In this plot, predictors that are grouped together are correlated to each other. In our case, home runs, intentional walks, hit by pitches, walks, doubles and singles are positively correlated together in both components. Additionally, stolen bases and triples are correlated to each other positively in component one and negatively in component two. Finally, from this plot we can see that predictors that are farther from the origin and closer to green (representing a high square cosine) are better represented. For our predictors, they are all represented fairly similarly, with doubles being the most and hit by pitches and intentional walks falling behind.

**Analysis**

By looking at our calculations and visualizations, we can see that the first two components capture almost all of the variance in our predictors, so we would select these for our model. The weights of component one are fairly evenly distributed over the predictors while component two puts more weight on stolen bases and triples.

**Problems that Could Arise**

If the data is not a random sample from the population, then the variables will be measured on some arbitrary scale that depends on the sampling design since the sample standard deviations used to standardize the variables will not align with the population. Our sample is a random sample from our population, so we do not run into this issue.

# Conclusion

If we want to include as many of our predictors as possible, there are issues with collinearity among some predictors. To mitigate this issue, a Ridge or Lasso regression can be employed, with the Lasso regression with a $\lambda$ value equal to 0.4766701 being the best fit. However, for ease of explanation one could consider using the Multi-Linear Regression we created in step 3 as it has much fewer predictors and the fit and predictive accuracy of the model is very similar to the models generated through Shrinkage methods. In our innovative step we used principal component analysis to demonstrate that we can take linear combinations of our predictors to decrease our number of regressors to two while still getting a well fit model.