

126 Data Project, Summary

Sam Ream, Valeria Lopez, Skyler Yee

Step 1

We selected “The History of Baseball” data set, which is a record of all the statistics of baseball players who have played in the MLB up until 2015. We sampled 500 players who played between 2000-2015 to capture the current state of the game. Our population is all of the MLB players from 2000-2015. We did not want to generalize the whole history of the MLB because the rules of the game and the strategies used by teams and players are always changing. Our predictors are singles, doubles, triples, home runs, walks, stolen bases, hit by pitches, intentional walks, BMI, and batting hand. Our objective was to predict the expected number of runs that future players will have on record, given observations from their past performances.

Almost all of our predictors were highly positively correlated to one another. Additionally, most of our predictors were highly correlated to our response variable, Runs.

Step 2-3

```
##  
## Durbin-Watson test  
##  
## data:  batting$RUNS ~ batting$DOUBLE  
## DW = 2.0821, p-value = 0.8207  
## alternative hypothesis: true autocorrelation is greater than 0
```

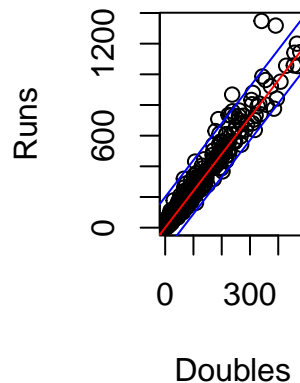


Figure 1: Scatter plot of the relationship between Runs and Doubles

- **Linearity:** All the points on the relationship plot above are arranged in a very linear way without transformations
- **Constant Variance:** Almost all of the points have a similar distance from a proposed straight line.

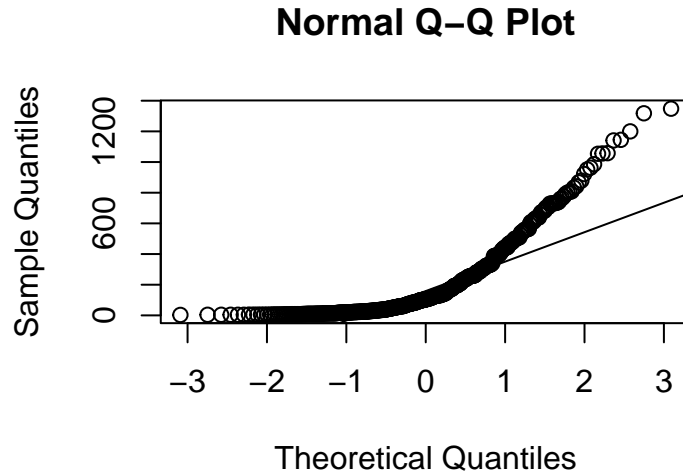


Figure 2: QQ-Plot showing that the data is not normally distributed

- **Independence:** With the knowledge that one batter hitting the ball well enough to get a double does not affect the likelihood of the next batter doing the same, we know that the predictors are independent of one another.

- **Normality:** While our errors do not appear to be normally distributed, our large sample size allows us to leverage the Central Limit Theorem to make meaningful analysis.

Hypothesis Testing

Significance Test

$$H_0 : \beta_i = 0$$

$$H_a : \beta_1 \neq 0$$

$$\alpha = 0.05$$

Test Statistic = 102.942

P Value ≈ 0

We reject H_0 at 0.05 level. Thus, the amount of doubles a player hits is a significant predictor of how many runs the player scores.

Fit of Model

The R^2 value of our model is 0.9551, which means that the model explains 95.51% of the variance of the recorded events. Additionally, the residual plot in figure 3 shows how the data points share a similar spread which implies that the model is a good fit.

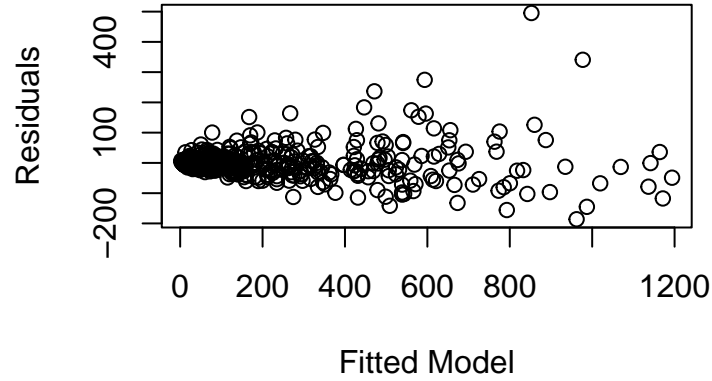


Figure 3: Residual Plot of the fitted model

Computational Models

For our computational models, we used the predictors: Total Intentional Walks, Singles, Triples, Stolen Bases, and Home Runs obtained in a career. We selected these predictors because of their low correlation in addition to their interesting relation to obtained Runs.

Model 1 - Full Model (Ω)

$$\mathbb{E}[Y] = \text{Intercept} + \text{Intentional Walks} + \text{Singles} + \text{Triples} + \text{Stolen Bases} + \text{Home Runs} + \epsilon$$

Model 2 - Reduced Model (ω)

$$\mathbb{E}[Y] = \text{Intercept} + \text{Singles} + \text{Triples} + \text{Home Runs} + \epsilon$$

Comparison:

$H_0 : \beta \in \omega$: The Reduced Model is sufficient

$H_\alpha : \beta \in \Omega \omega \in w$: The Reduced Model is not sufficient

```
## Analysis of Variance Table
##
## Model 1: RUNS ~ INT_WALKS + SINGLES + TRIPLE + STOLEN_BASES + HOME_RUNS
## Model 2: RUNS ~ SINGLES + TRIPLE + HOME_RUNS
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      494 386485
## 2      496 448675 -2    -62189 39.745 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

As we rejected H_0 in favor for H_α , we can determine that the reduced model does not model the data well enough to justify the reduction in predictors. As such, we decided to use model 1, the full model, as our computational model.

Statistical Model

We used a stepwise search to create the best model for our data. For a size of 4 predictors the variables home runs, singles, walks, and stolen bases create a well fit model.

Final Model Selection

Between the two models we created, the statistical model and computational model, we selected the statistical model. The reason behind this selection is that the statistical model has a larger R_{adj}^2 value and we want to explain as much of the variance as possible in our model.

Step 4

Collinearity 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

As we observed the collinearity between certain predictors to be high as seen in the table above, we implemented a Ridge and Lasso regression on the entire set of predictors to make a model that fits well and utilizes all predictors. While the Lasso regression did not eliminate any predictors, it did result in the best fit of all the previous models as can be demonstrated in the plot and table below.

Coefficients of the Ridge Model

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                               s1
## (Intercept)  3.2023432
## TRIPLE      1.9086140
## HOME_RUNS   0.6733704
## SINGLES     0.1804050
## WALKS       0.2409483
## DOUBLE     0.5607536
## INT_WALKS   0.3426712
```

```
## STOLEN_BASES 0.4687444
## HIT_BY_PITCH 0.6620003
```

In observing our coefficients, we can see that Triples have the largest effect (an increase of 1.9420 expected runs per Triple) on the expected number of Runs and that every predictor contributes some information to the model.

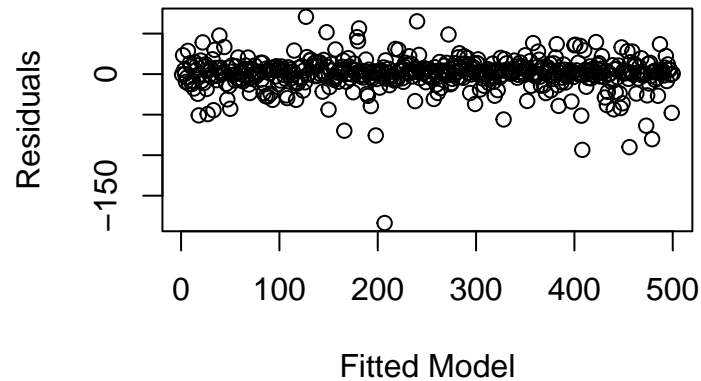


Figure 4: Residual plot of the Ridge Regression

Coefficients of the Lasso Model

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept) -2.9755404
## TRIPLE      1.1596216
## HOME_RUNS    0.8057241
## SINGLES      0.2781733
## WALKS        0.2599091
## DOUBLE       0.4581973
## INT_WALKS    -0.2765872
## STOLEN_BASES 0.4178732
## HIT_BY_PITCH 0.2241790
```

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 decreases the expected number of runs and no predictor can be removed without consequence.

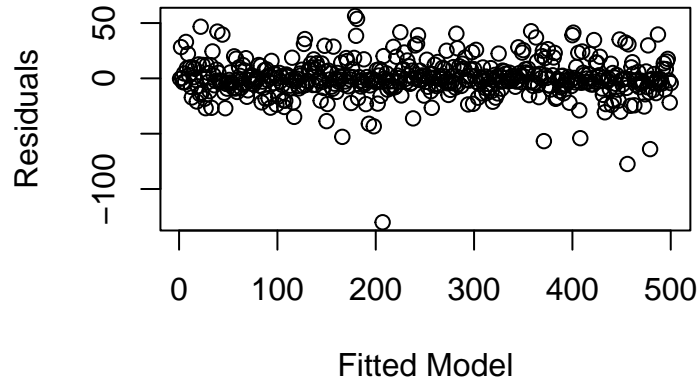


Figure 5: Residual plot of the Lasso Regression

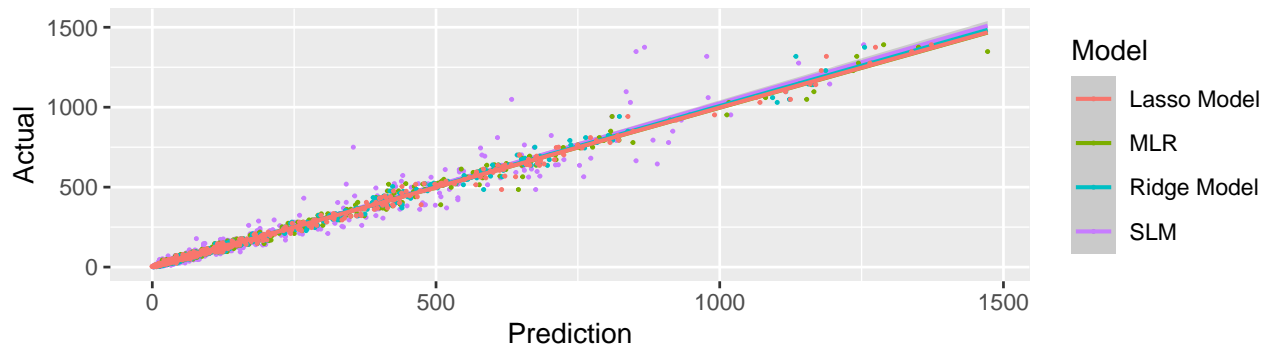


Figure 6: A comparison of the Linear Models

Model Type	R^2	MSE
SLR	0.9393236	4125.3574196
MLR	0.9919045	550.4106253
Ridge Regression	0.9937332	426.0792005
Lasso Regression	0.9943832	381.8830127

Analysis

As some predictors had a very high level of collinearity, we elected to use shrinkage methods to fit a model to all predictors. The graph and table above were generated for several different sets of 500 new random observations from our original data set. 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—so it might be a good idea to use it if ease of explainability is important.

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.

Visualizations

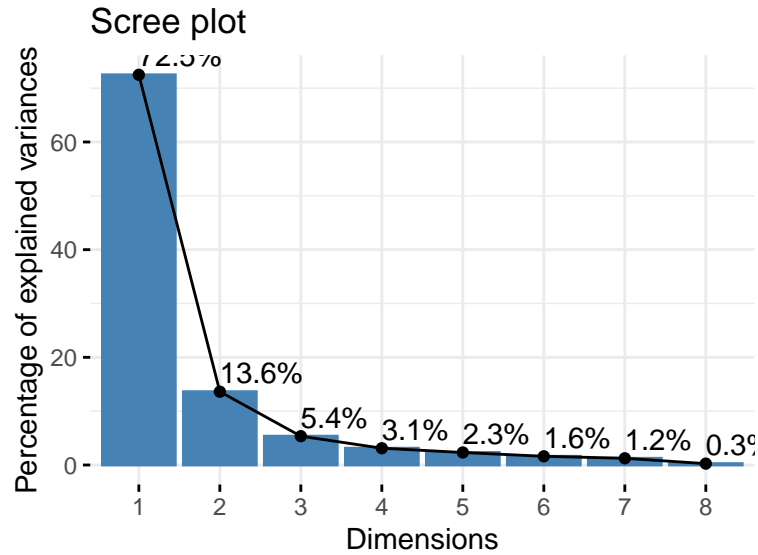


Figure 7: Component contributions to the total variance

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

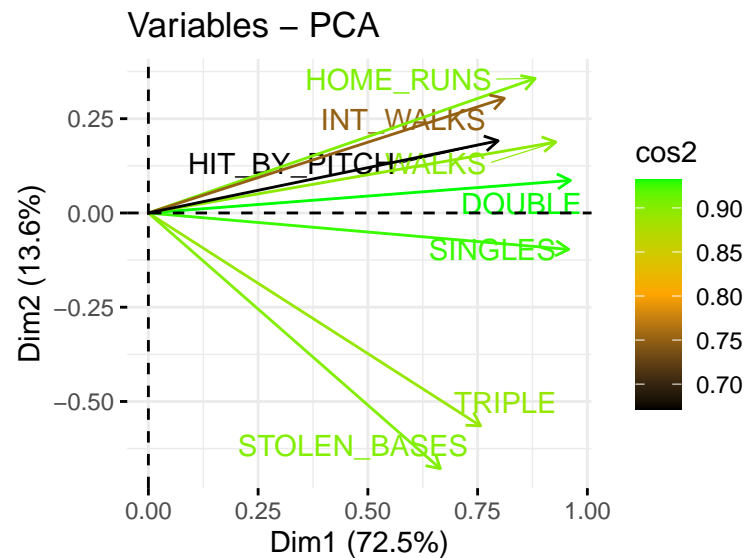


Figure 8: Biplot combined with cos2

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.

Points of interest

We noted that in both of our Ridge and Lasso regressions that the number of Triples that a player scored had the largest effect on the number of Runs that they would go on and achieve. This makes sense as getting to third base in one hit would significantly increase the odds of making it to the home base. Interestingly enough though, Doubles and Home runs both had a very similar effect—implying that training players to get good at getting to the second base will result in similar results of training them to hit home runs more consistently.

Additionally, we would like to note that it may be worthwhile to conduct a similar experiment investigating the expected chance of achieving a run instead of the number of runs scored. Some predictors might be inflated as a player who is not particularly skilled may be given an opportunity to bat more than usual—which could affect this data. Since we only polled information from the MLB in recent years, it is our belief that this will not be the case as all players are of a high skill level.

Summary

To best model the expected number of Runs that a player will have on record given observations of their performance, We would suggest using the Multi-Linear Regression or Lasso Regression. While a Singular linear regression with Doubles is not the worst model, it can be significantly improved on. The MLR is not as good of a fit as the Lasso Regression, but it is much easier to explain and can help players understand what to focus on if they want to increase their number of runs. On the other hand, the Lasso Regression is an incredibly well fit model that can accurately predict the expected number of runs better than any other proposed model. We believe that the Lasso Regression was so successful because we had a lot of predictors included in our data and the Lasso Regression excels when this is the case. While principle component analysis did result in some very interesting results, we found that for the purposes of this analysis it did not increase accuracy or make the model easier to comprehend for the end users.