# 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**

| | AB | T | HR | S | W | D | IW | SB | HBP |
|---|---|---|---|---|---|---|---|---|---|
| AT_BAT (AB) | 1 | 0.73 | 0.87 | 0.98 | 0.9 | 0.98 | 0.74 | 0.62 | 0.76 |
| TRIPLE (T) | 0.73 | 1 | 0.47 | 0.76 | 0.59 | 0.68 | 0.45 | 0.8 | 0.49 |
| HOME_RUNS (HR) | 0.87 | 0.47 | 1 | 0.79 | 0.88 | 0.89 | 0.78 | 0.36 | 0.72 |
| SINGLES (S) | 0.98 | 0.76 | 0.79 | 1 | 0.85 | 0.95 | 0.72 | 0.68 | 0.74 |
| WALKS (W) | 0.9 | 0.59 | 0.88 | 0.85 | 1 | 0.9 | 0.8 | 0.5 | 0.71 |
| DOUBLE (D) | 0.98 | 0.68 | 0.89 | 0.95 | 0.9 | 1 | 0.74 | 0.55 | 0.75 |
| INT_WALKS (IW) | 0.74 | 0.45 | 0.78 | 0.72 | 0.8 | 0.74 | 1 | 0.37 | 0.58 |
| STOLEN_BASES (SB) | 0.62 | 0.8 | 0.36 | 0.68 | 0.5 | 0.55 | 0.37 | 1 | 0.4 |
| HIT_BY_PITCH (HBP) | 0.76 | 0.49 | 0.72 | 0.74 | 0.71 | 0.75 | 0.58 | 0.4 | 1 |

**Variance Inflation factor Table**

```
##  (Intercept)       AT_BAT       TRIPLE    HOME_RUNS      SINGLES        WALKS
##     2.000434   193.307023     4.202315    16.711230    90.879729     8.023297
##       DOUBLE    INT_WALKS STOLEN_BASES HIT_BY_PITCH
##    36.986176     4.052985     3.449440     2.503874
```

### 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

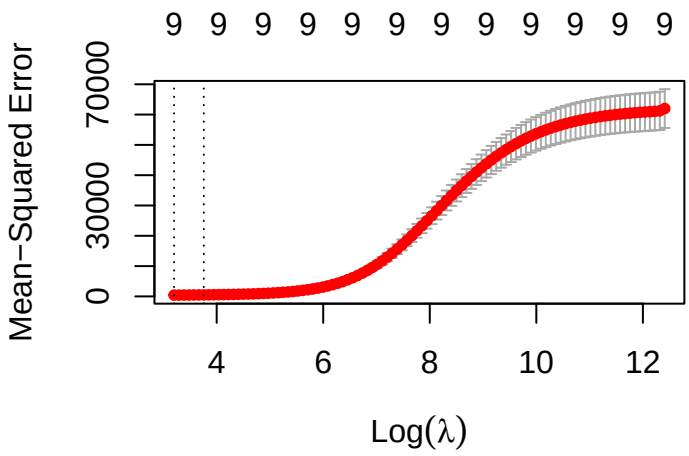


Figure 1: The relationship between MSE and Log(lambda)

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

```
## [1] 24.53741
```

### Model Analysis

#### R-Squared Analysis

```
## [1] 0.9936345
```

When Lambda equals 24.5374055, the R-Squared is 0.9936345. This implies that the model explains approximately 99.36% of the variation in the response in our training data set.

#### Coefficient Analysis

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  0.27979948
## AT_BAT       0.02421806
## TRIPLE       1.64344654
## HOME_RUNS    0.56755451
## SINGLES      0.13430865
## WALKS        0.21506443
## DOUBLE       0.44115090
## INT_WALKS    0.40320761
## STOLEN_BASES 0.43156691
## HIT_BY_PITCH 0.59305714
```

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

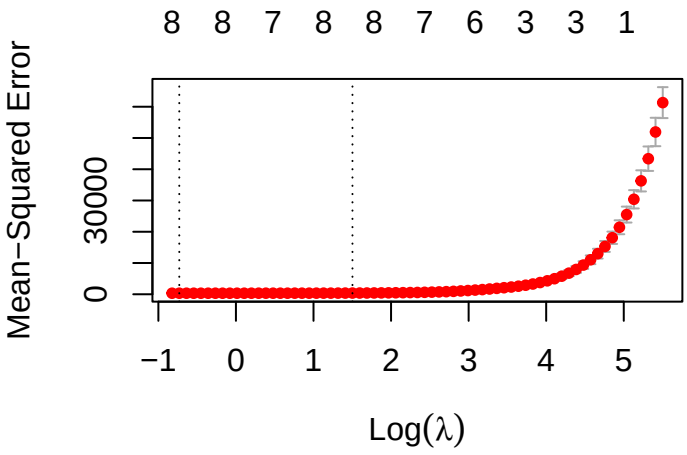


Figure 2: The relationship between MSE and Log(lambda)

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

```
## [1] 0.4816792
```

## Model Analysis

**R-Squared Analysis** When Lambda equals `best_lambda`, the R-Squared is 0.9954734. This implies that the model explains approximately 99.55% of the variation in the response in our training data set.

**Coefficient Analysis**

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)  -2.9446270
## AT_BAT        .
## TRIPLE        1.1593205
## HOME_RUNS     0.8101005
## SINGLES       0.2809539
## WALKS         0.2594500
## DOUBLE        0.4480666
## INT_WALKS    -0.2722396
## STOLEN_BASES  0.4146871
## HIT_BY_PITCH  0.2197922
```

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.

# Comparison of our Models

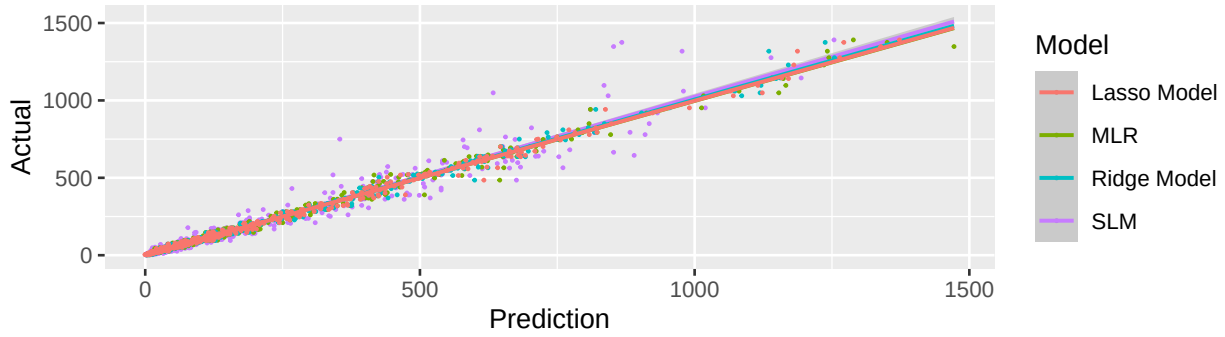


Figure 3: A comparison of the Linear Models

**MSE and $R^2$ By model**

| Model Type | $R^2$ | MSE |
|---|---|---|
| SLR | 0.9393236 | 4125.3574196 |
| MLR | 0.9919045 | 550.4106253 |
| Ridge Regression | 0.9933724 | 450.6093155 |

| Model Type | $R^2$ | MSE |
|---|---|---|
| Lasso Regression | 0.9943301 | 385.4925752 |

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

Waiting for sam's work

## 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 0.4816792 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.

INCLUDE BIT ABOUT SAMS WORK