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.

Ridge Regression

Optimal Lambda - Ridge Regression

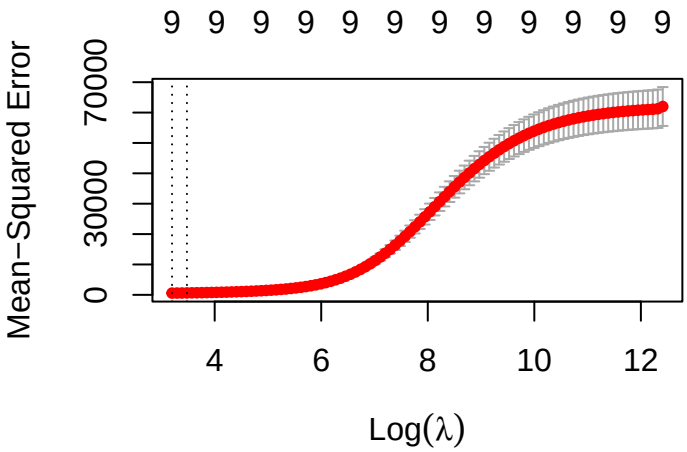

Figure 1: The relationship between MSE and Log(lambda)

We found that the MSE was minimized when λ is equal to:

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

Model Analysis

R-Squared Analysis When Lambda equals 24.53741, the R-Squared is 0.9914. This implies that the model explains approximately 99.14% of the variation in the response values..

MSE Analysis

Lasso Regression

Optimal Lambda - Lasso Regression

```
## [1] 0.3643727
```

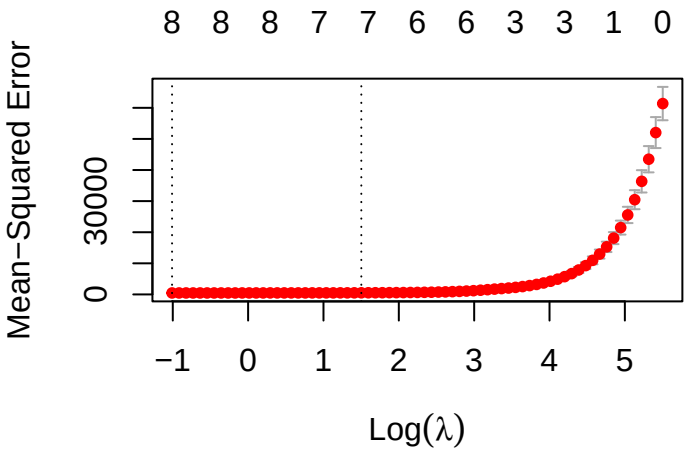

Figure 2: The relationship between MSE and Log(lambda)

We found that the MSE was minimized when λ is equal to:

```
## [1] 0.3643727
```

Model Analysis

R-Squared Analysis When Lambda equals 0.3643727, the R-Squared is 0.9935687. This implies that the model explains approximately 99.36% of the variation in the response values.

MSE Analysis

Comparison of our Models

Investigation - Principle Component Analysis

Conclusion