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

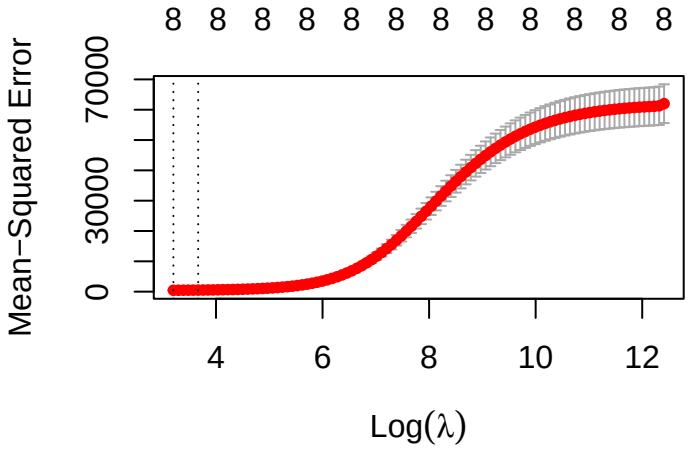

Figure 1: The relationship between MSE and Log(lambda)

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

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

Model Analysis

R-Squared Analysis

```
## [1] 0.9931578
```

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..

Coefficient Analysis

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  1.67374671
## AT_BAT       0.02678647
## HOME_RUNS    0.54511668
## SINGLES      0.14637655
## WALKS        0.21679832
## DOUBLE       0.46391340
## INT_WALKS    0.35821606
## STOLEN_BASES 0.59104211
## HIT_BY_PITCH 0.57095025
```

Lasso Regression

Optimal Lambda - Lasso Regression

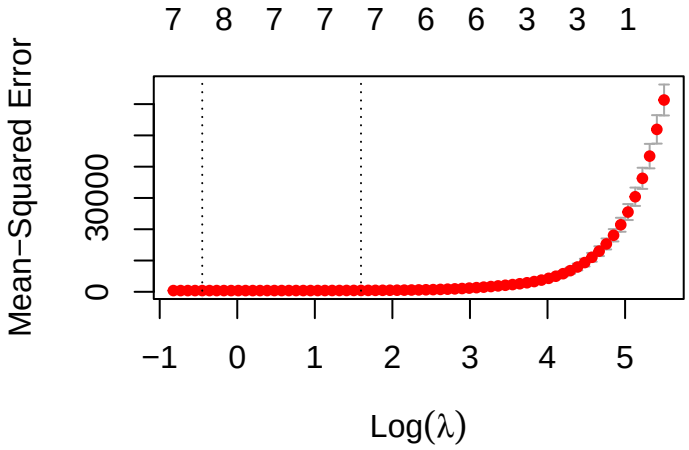

Figure 2: The relationship between MSE and Log(lambda)

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

```
## [1] 0.6367516
```

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.

Coefficient Analysis

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)  -1.4489150
## AT_BAT        .
```

```
## HOME_RUNS     0.7820053
## SINGLES       0.2940449
## WALKS         0.2542213
## DOUBLE        0.4828622
## INT_WALKS    -0.2476265
## STOLEN_BASES  0.5439248
## HIT_BY_PITCH  0.2033609
```

Comparison of our Models

There is a plot below this

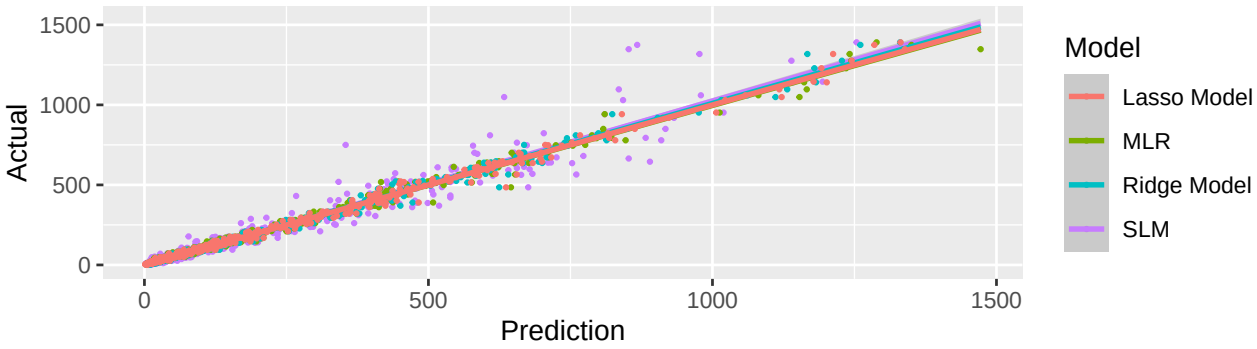

Figure 3: A comparison of the Linear Models

Model Type	R^2	MSE
SLR	0.9393236	4125.3574196
MLR	0.9919045	550.4106253
Ridge Regression	0.9927685	491.6653673
Lasso Regression	0.9935948	435.4840011

There is a plot above this

Investigation - Principle Component Analysis

Waiting for sam's work

Conclusion

GDRIVE - PSTAT126_PROJ_STEP_4_CONC

(Everything is still temp above, will finalize tomorrow well before 3pm)