

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

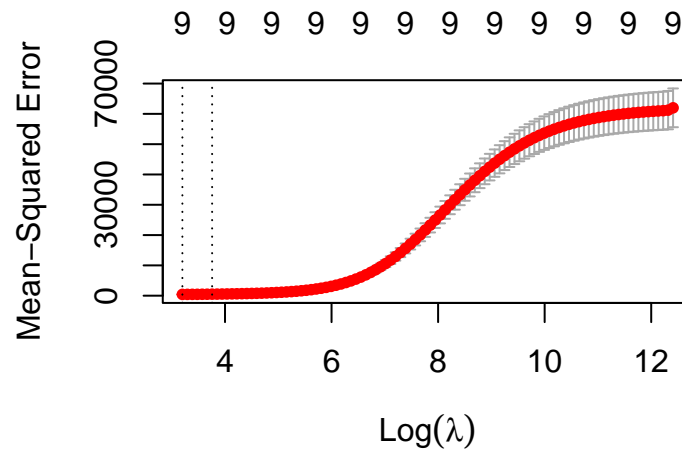


Figure 1: The relationship between MSE and Log(lambda)

We found that the MSE was minimized when λ 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

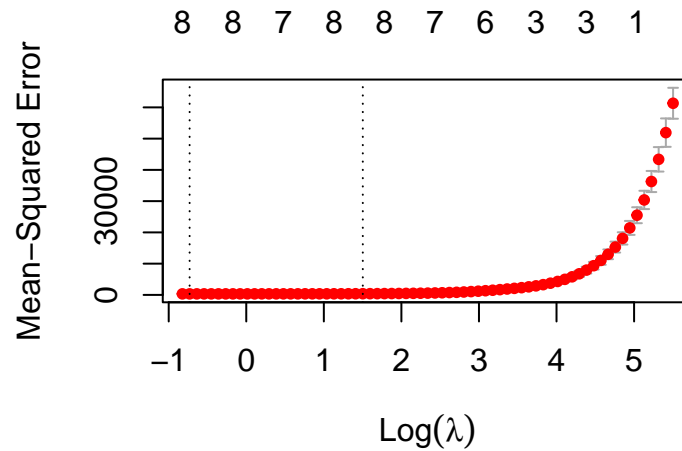


Figure 2: The relationship between MSE and Log(lambda)

We found that the MSE was minimized when λ 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

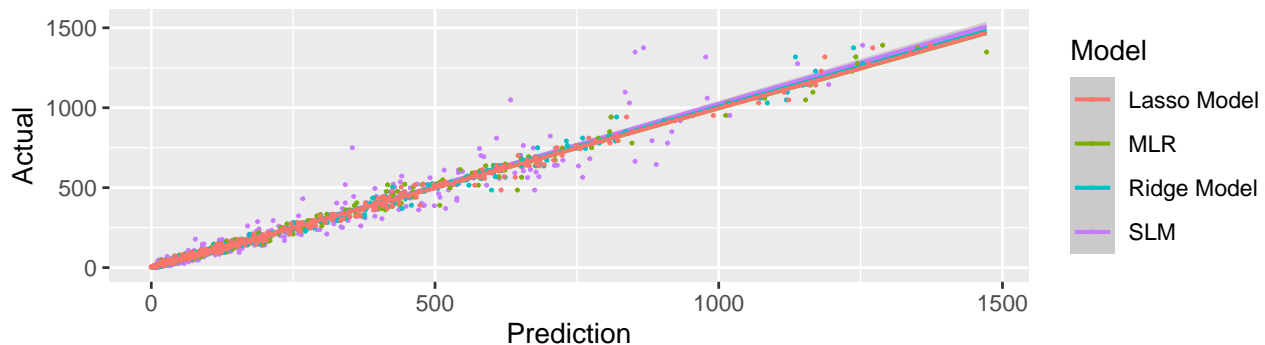


Figure 3: A comparison of the Linear Models

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
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

Innovation Step: Principal Components 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, \dots, 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 $\mathbf{X}' = (X_1, X_2, \dots, X_k)$ 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, compute the PCA using the sample correlation matrix, and assess the cumulative proportion of each principal component to see which

principal components explain the most of the total variance. Then look at the loading matrix to see how these principal components relate to each column of our data.

Calculations

```
## player_id AT_BAT RUNS HOME_RUNS TRIPLE DOUBLE SINGLES WALKS INT_WALKS
## 1 abernbr01 868 97 8 5 36 163 60 1
## 2 abreubo01 6651 1141 223 30 457 1188 1160 84
## 3 adamsma01 1084 123 39 5 63 192 64 6
## 4 adamsru01 884 125 17 7 46 148 85 2
## 5 aguilch01 230 26 3 2 13 35 16 1
## 6 ahmedni01 491 58 10 6 19 74 32 1
## STOLEN_BASES HIT_BY_PITCH BMI HAND
## 1 21 7 H R
## 2 319 28 0 L
## 3 4 3 B L
## 4 15 5 0 L
## 5 2 0 0 R
## 6 4 1 0 R

## HOME_RUNS TRIPLE DOUBLE SINGLES WALKS INT_WALKS STOLEN_BASES HIT_BY_PITCH
## 1 8 5 36 163 60 1 21 7
## 2 223 30 457 1188 1160 84 319 28
## 3 39 5 63 192 64 6 4 3
## 4 17 7 46 148 85 2 15 5
## 5 3 2 13 35 16 1 2 0
## 6 10 6 19 74 32 1 4 1

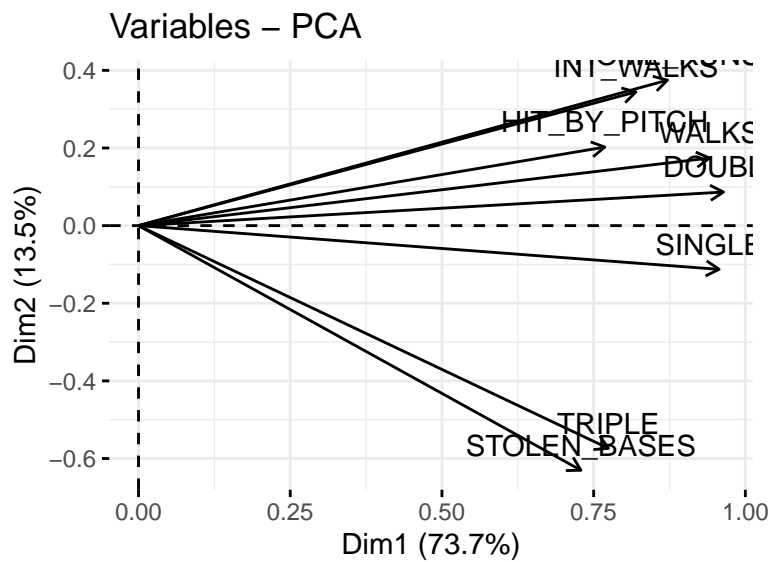
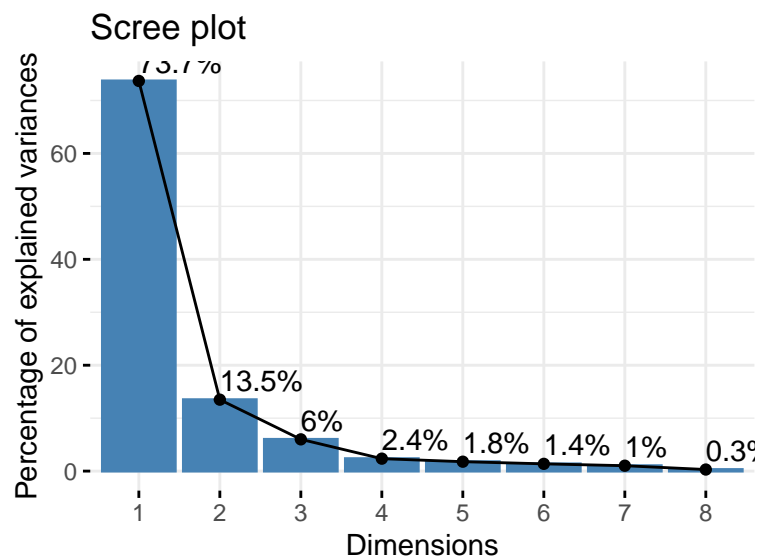
## HOME_RUNS TRIPLE DOUBLE SINGLES WALKS INT_WALKS
## [1,] -0.5241498 -0.23485367 -0.4359369 -0.3088904 -0.4200702 -0.4654091
## [2,] 2.5225738 1.66648754 3.8248574 2.7936954 5.1835827 3.3284386
## [3,] -0.0848548 -0.23485367 -0.1626794 -0.2211099 -0.3996933 -0.2368641
## [4,] -0.3966126 -0.08274637 -0.3347304 -0.3542941 -0.2927144 -0.4197001
## [5,] -0.5950039 -0.46301461 -0.6687119 -0.6963353 -0.6442163 -0.4654091
## [6,] -0.4958082 -0.15880002 -0.6079880 -0.5782857 -0.5627086 -0.4654091
## STOLEN_BASES HIT_BY_PITCH
## [1,] -0.1128655 -0.3679012
## [2,] 4.5572164 0.5406037
## [3,] -0.3792796 -0.5409498
## [4,] -0.2068940 -0.4544255
## [5,] -0.4106224 -0.6707362
## [6,] -0.3792796 -0.6274741

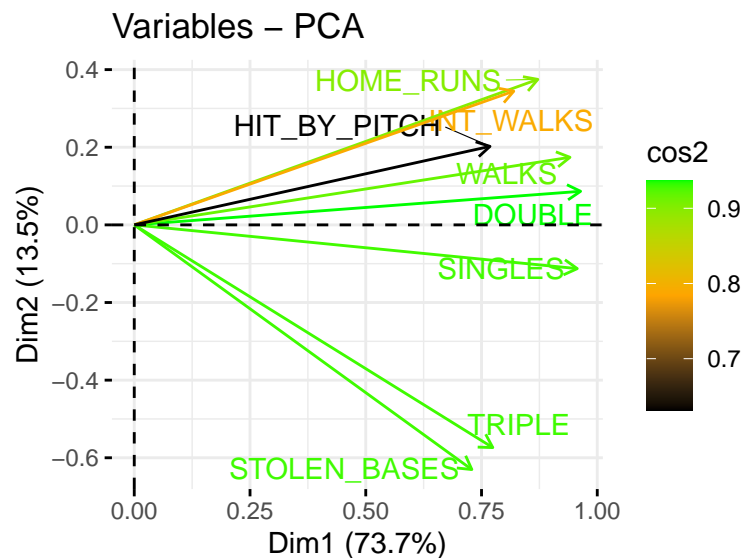
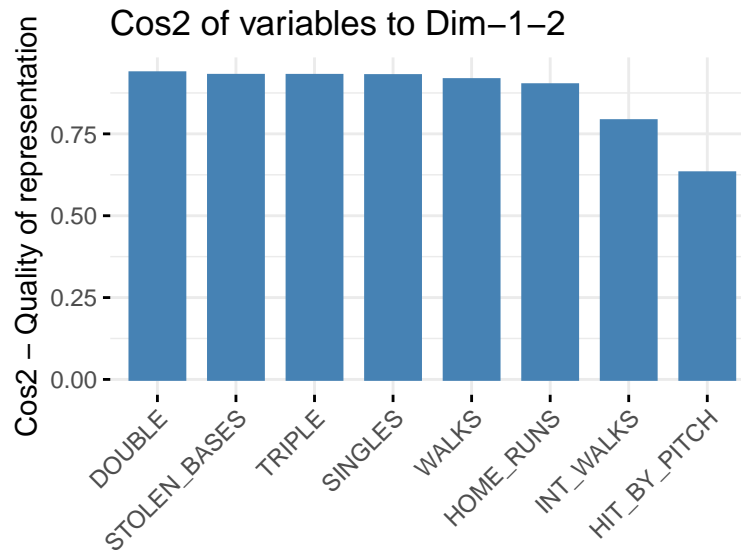
## Importance of components:
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Standard deviation 2.4255705 1.0378322 0.69213845 0.43327716 0.37672042
## Proportion of Variance 0.7368979 0.1349068 0.06000196 0.02351316 0.01777533
## Cumulative Proportion 0.7368979 0.8718046 0.93180659 0.95531975 0.97309509
## Comp.6 Comp.7 Comp.8
## Standard deviation 0.33096666 0.28546021 0.154215291
## Proportion of Variance 0.01371981 0.01020635 0.002978752
## Cumulative Proportion 0.98681489 0.99702125 1.000000000
```

NULL

##		Comp.1	Comp.2
##	HOME_RUNS	0.3593405	0.3611824
##	TRIPLE	0.3192515	-0.5528387
##	DOUBLE	0.3974176	0.0833358
##	SINGLES	0.3944617	-0.1085914
##	WALKS	0.3879672	0.1677413
##	INT_WALKS	0.3379293	0.3319424
##	STOLEN_BASES	0.3005838	-0.6073543
##	HIT_BY_PITCH	0.3167755	0.1948786

Visualizations





Analysis

By looking at our calculations and visualizations, we can see that

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 λ 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