126 Data Project, Step 4

Sam Ream, Valeria Lopez, Skyler Yee

```
##
## Call:
## lm(formula = RUNS ~ HOME_RUNS + TRIPLE + DOUBLE + SINGLES + WALKS +
##     INT_WALKS + STOLEN_BASES + HIT_BY_PITCH, data = batting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.879  -6.647   1.320   6.232 120.103
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.53064    1.01346  -3.484 0.000539 ***
## HOME_RUNS     0.83843    0.03132  26.772  < 2e-16 ***
## TRIPLE        1.18244    0.13439   8.798  < 2e-16 ***
## DOUBLE        0.43452    0.04107  10.581  < 2e-16 ***
## SINGLES       0.28335    0.01094  25.909  < 2e-16 ***
## WALKS         0.26418    0.01124  23.503  < 2e-16 ***
## INT_WALKS    -0.39823    0.06512  -6.116 1.96e-09 ***
## STOLEN_BASES  0.41945    0.02767  15.158  < 2e-16 ***
## HIT_BY_PITCH  0.22759    0.05000   4.552 6.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.75 on 491 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9955
## F-statistic: 1.368e+04 on 8 and 491 DF,  p-value: < 2.2e-16
```

intro:

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.

#Conclusion The data was what we had anticipated because the models had high R squared values, which indicates a large quantity of the variability can be explained by the regression. Using Ridge Regression, which aims to minimize SSE, we saw an R-squared value of 0.9914. Using LASSO regression, which shrinks the less important coefficients to zero, we got an R-squared value of 0.9936.

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 $X_1, X_2, ..., X_k$ 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 $\mathbf{X}' = (X_1, X_2, ...X_k)$ 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 adamsda02    140   10         2      1      5      19     9         0
## 4 alcanar01    304   36        10      2     11      36    22         0
## 5 alfoned01   2435  305        59      4    118     477   251        15
## 6 alomaro01   1852  280        40     20     92     376   210        12
##   STOLEN_BASES HIT_BY_PITCH BMI HAND
## 1           21            7   H    R
## 2          319           28   0    L
## 3            0            2   0    R
## 4            9            2   H    B
## 5           19           24   0    R
## 6           58            9   0    B
```

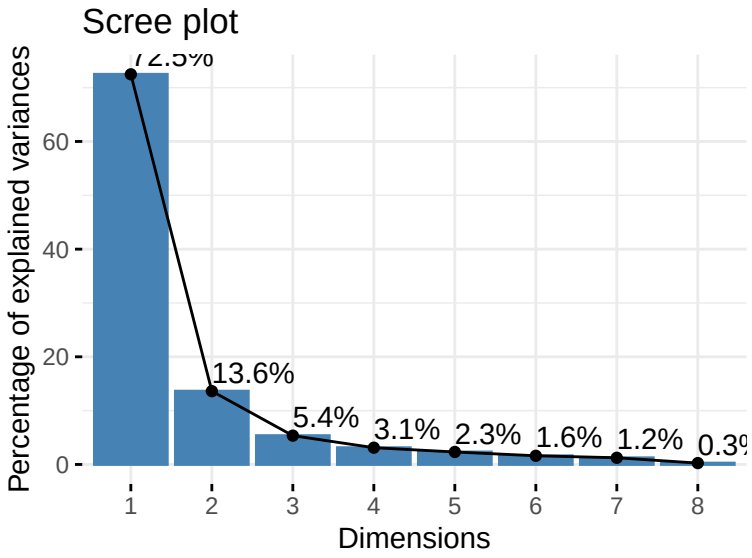
```
##   HOME_RUNS TRIPLE DOUBLE SINGLES WALKS INT_WALKS STOLEN_BASES HIT_BY_PITCH
## 1         8      5     36     163    60         1           21            7
## 2       223     30    457    1188  1160        84          319           28
## 3         2      1      5      19     9         0            0            2
## 4        10      2     11      36    22         0            9            2
## 5        59      4    118     477   251        15           19           24
## 6        40     20     92     376   210        12           58            9
```

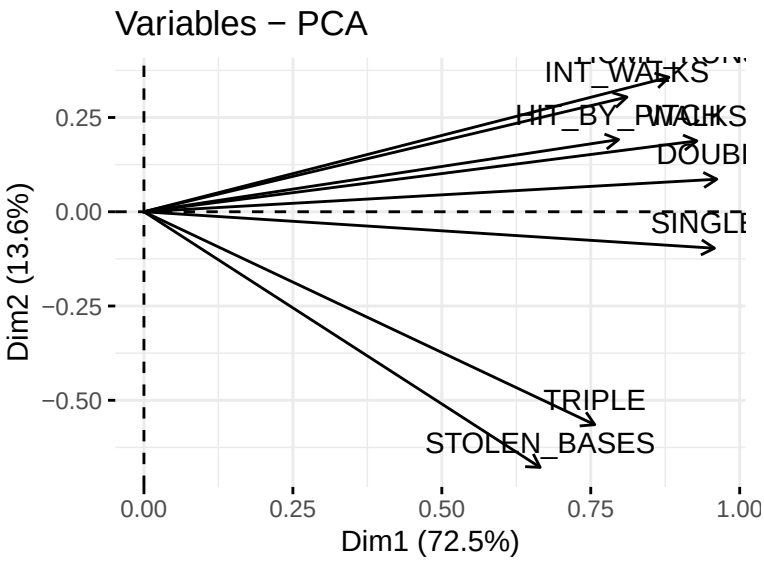
```
##       HOME_RUNS     TRIPLE     DOUBLE    SINGLES      WALKS  INT_WALKS
## [1,] -0.5749622 -0.2555350 -0.4687530 -0.3386818 -0.4613892 -0.4998066
## [2,]  2.4639500  2.0309271  3.8466857  2.8879086  5.5005516  3.4684302
## [3,] -0.6597690 -0.6213689 -0.7865169 -0.7919784 -0.7378064 -0.5476167
## [4,] -0.5466932 -0.5299105 -0.7250142 -0.7384642 -0.6673471 -0.5476167
## [5,]  0.1458961 -0.3469935  0.3717838  0.6497566  0.5738205  0.1695345
## [6,] -0.1226590  1.1163423  0.1052721  0.3318194  0.3516027  0.0261043
##      STOLEN_BASES HIT_BY_PITCH
## [1,]  -0.08354722   -0.3672401
## [2,]   5.84715121    0.5183872
## [3,]  -0.50148235   -0.5781038
## [4,]  -0.32236729   -0.5781038
## [5,]  -0.12335057    0.3496963
## [6,]   0.65281466   -0.2828947
```

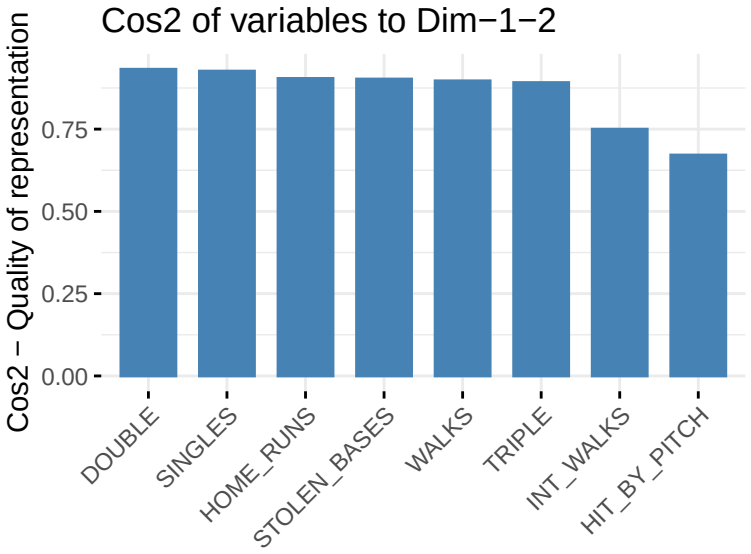
```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     2.405503 1.0427551 0.65380416 0.49875764 0.43051008
## Proportion of Variance 0.724755 0.1361896 0.05353956 0.03115721 0.02321379
## Cumulative Proportion  0.724755 0.8609446 0.91448416 0.94564137 0.96885516
##                           Comp.6     Comp.7      Comp.8
## Standard deviation     0.35906419 0.31510718 0.142971113
## Proportion of Variance 0.01614818 0.01243644 0.002560213
## Cumulative Proportion  0.98500335 0.99743979 1.000000000
```

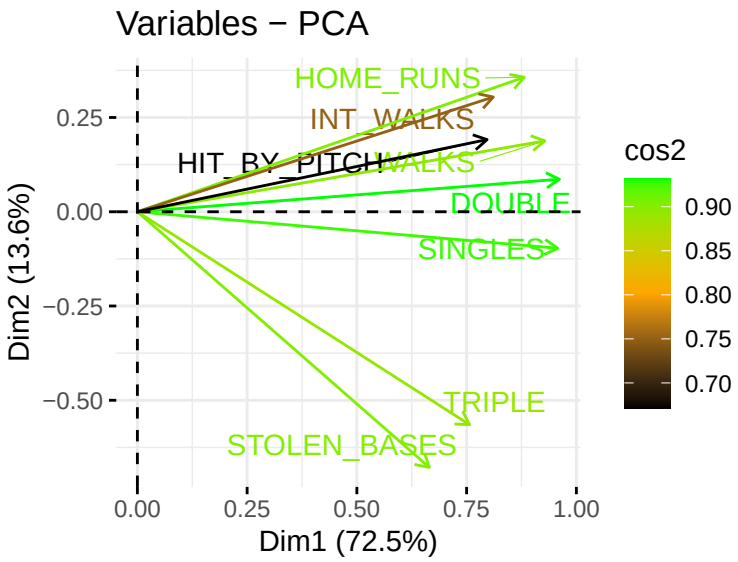
```
##                  Comp.1      Comp.2
## HOME_RUNS    0.3663597  0.34214705
## TRIPLE       0.3145657 -0.54158422
## DOUBLE       0.3997034  0.08274283
## SINGLES      0.3980834 -0.09279209
## WALKS        0.3858354  0.18026231
## INT_WALKS    0.3370378  0.29176878
## STOLEN_BASES 0.2765011 -0.65026645
## HIT_BY_PITCH 0.3311790  0.18356250
```

Visualizations




Variables − PCA
INT_WALKS
HIT_BY_PITCH
WALKS
DOUBL
SINGLE
TRIPLE
STOLEN_BASES
0.25
0.00
−0.25
−0.50
Dim2 (13.6%)
0.00
0.25
0.50
0.75
1.00
Dim1 (72.5%)


Cos2 of variables to Dim−1−2
Cos2 − Quality of representation
0.75
0.50
0.25
0.00
DOUBLE
SINGLES
HOME_RUNS
STOLEN_BASES
WALKS
TRIPLE
INT_WALKS
HIT_BY_PITCH



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.