126 Data Project, Step 2

Sam Ream, Valeria Lopez, Skyler Yee

Introduction

For our project, we are looking at the "History of Baseball" data set, which is a record of all the stats of baseball players who have played in the MLB up until the year 2015. We are interested in seeing which of our predictors (singles, doubles, triples, home runs, walks, intentional walks, hit by pitches, stolen bases, player BMIs, and batting hand) contribute the most to runs scored by individual players. So far, we have seen that our sample aligns with overall MLB stats (eg. the batting average for our sample is 0.246 compared to the MLB's 0.250) and all of our model assumptions hold, which allowed us to construct a simple linear model with doubles as our predictor. Now, we will be selecting predictors to construct potential multiple linear models and selecting the best one.

Analysis of Variables

When investigating the predictors, we noted it did not appear that we needed to use any transformation to make the data more linear. Additionaly, it is clear that some predictors were highly correlated. An example of this is Singles and Doubles which have a correlation of 0.95. As a consequence of this, we elected to use only one of these two variables in our hand-made model and did the same with other predictors with similar levels of correlation.

We decided to not include interaction variables because different values of our categorical variables (BMI and handedness) do not drastically affect the response. We also felt it was not necessary for any of our non-categorical variables as there are no interactions that we believe to be interesting.

Computational Models

For our computational models, we used the predictors: Total Intentional walks, Singles, Triples, Stolen Bases, and Home_Runs obtained in a career. We selected these predictors because of their low correlation in addition to their interesting relation to obtained Runs. To help prevent over-correlation, we also elected to create a reduced model using only predictors related to hitting the ball and compared the two to see if we could use a smaller model.

Model 1 - Full Model (Ω)

$$\mathbb{E}[Y] = \text{Intercept} + \text{Intentional Walks} + \text{Singles} + \text{Triples} + \text{Stolen Bases} + \text{Home Runs} + \epsilon$$

Model 2 - Reduced Model (ω)

$$\mathbb{E}[Y] = \text{Intercept} + \text{Singles} + \text{Triples} + \text{Home Runs} + \epsilon$$

Comparison:

$H_0 : \beta \in \omega$: The Reduced Model is sufficient

$H_\alpha : \beta \in \Omega\omega \in w$: The reduced Model is not sufficient

```
## Analysis of Variance Table
##
## Model 1: RUNS ~ INT_WALKS + SINGLES + TRIPLE + STOLEN_BASES + HOME_RUNS
## Model 2: RUNS ~ SINGLES + TRIPLE + HOME_RUNS
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    494 386485
## 2    496 448675 -2    -62189 39.745 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

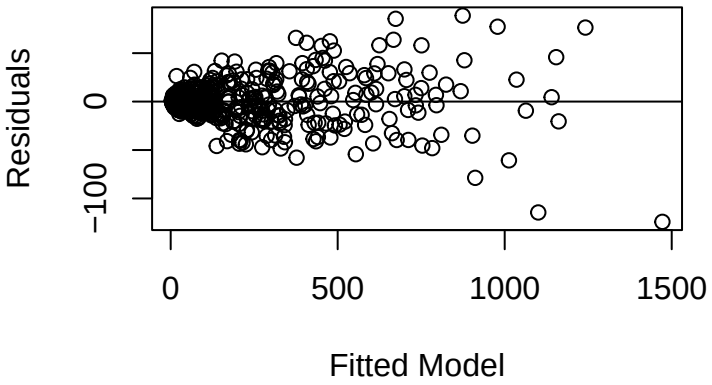
Conclusion

As we rejected H_0 in favor for H_α, we can determine that the reduced model does not model the data well enough to justify the reduction in predictors. As such, we decided to use model 1, the full model, as our computational model.

Statistical Model

We used a stepwise search to create the best model for our data. For a size of 4 predictors the variables home runs, singles, walks, and stolen bases create a well fit model.

Residual Plot and Summary table




```
##
## Call:
## lm(formula = RUNS ~ HOME_RUNS + SINGLES + WALKS + STOLEN_BASES,
##     data = batting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.140   -8.110   -0.528    6.956   88.759
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.781728   1.280636   -0.61    0.542
## HOME_RUNS     0.977261   0.031560   30.96   <2e-16 ***
## SINGLES       0.400909   0.007306   54.87   <2e-16 ***
## WALKS         0.268561   0.013396   20.05   <2e-16 ***
## STOLEN_BASES  0.490476   0.028537   17.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.6 on 495 degrees of freedom
## Multiple R-squared:  0.9925, Adjusted R-squared:  0.9925
## F-statistic: 1.642e+04 on 4 and 495 DF,  p-value: < 2.2e-16
```

Final model selection

Between the two models we created, the statistical model and computational model, we selected the Statistical model. The reason behind this selection is that the statistical model has a larger R^2_{adj} value and we want to explain as much of the variance as possible in our model.

Analysis of the Final Model:

Coefficient Interpretations

β_1: Every additional home run a player hits is associated with an increase of about 0.977261 mean runs, after accounting for singles, walks and stolen bases.

β_2: Every additional single a player hits is associated with an increase of about 0.400909 mean runs, after accounting for home runs, walks and stolen bases.

β_3: Every additional walk a player earns is associated with an increase of about 0.268561 mean runs, after accounting for home runs, singles and stolen bases.

β_4: Every additional stolen base a player earns is associated with an increase of about 0.268561 mean runs, after accounting for home runs, singles and walks.

Significance Tests

Looking at the p-values for each of our coefficients in the summary table, we can see that all of our coefficients are significant at the 0.05 level.

Analysis of residuals and influence points

Interpretation of Unusual Observations

Looking at the plots of leverage, externally studentized residuals, and Cook's Distance/Influence; we can see that there are a few outliers. However, when removing these observations and retraining the model, there was a negligable effect on the fit of the data. As a result, we elected to leave these values in to retain the information they hold.

Interpretation of Model

The big Skrunkly

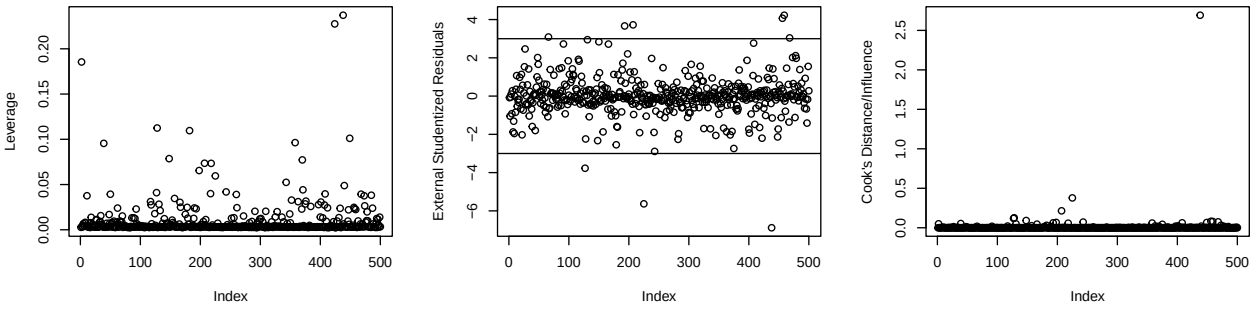


Figure 1: Plot of Leverage, Externally Studentized Residuals, and Influence

Confidence Interval

```
##                   2.5 %    97.5 %
## (Intercept)  -3.2978808 1.7344253
## HOME_RUNS     0.9152519 1.0392696
## SINGLES       0.3865542 0.4152631
## WALKS         0.2422416 0.2948808
## STOLEN_BASES  0.4344080 0.5465445
```

```
##   HOME_RUNS SINGLES   WALKS STOLEN_BASES
## 1    48.678  270.59 145.128       25.198
```

```
##       fit      lwr      upr
## 1 206.606 204.7084 208.5036
```

With 95% confidence, the mean predicted value is estimated to be between 204.71 and 208.50.

Prediction Interval

```
##   HOME_RUNS SINGLES WALKS STOLEN_BASES
## 1         2     477   112           17
```

```
##        fit      lwr      upr
## 1 230.8232 188.0602 273.5861
```

With 95% confidence, the predicted value for a combination of 2 home runs, 5 singles, 10 walks, and 7 stolen bases is between 188.06 and 273.59.

Summary

We sampled our data randomly from years 2000-2015 to get an accurate representation of the population and represent the changing baseball strategy. We analyzed multiple quantitative variables along with BMI and the batting hand of players in relation to runs. While checking the assumptions for linear regression, we found that our data was not normally distributed but were able to continue with our analysis due to the central limit theorem. Through hypothesis testing, we found that doubles is a significant predictor of how many runs the player scores and the residual plot showed that our model was a good fit. In order to create computational models, we chose variables that had low correlation, which consisted of intentional

walks singles, triples, stole bases, and home runs. Our other computation model was a reduced version. To create a statistical model, we used stepwise search with four predictors, resulting in a model with home runs, singles, walks, and stolen bases. Then we tested the fit of the model and created confidence and prediction intervals.