126 Data Project, Step 2

Sam Ream, Valeria Lopez, Skyler Yee

Introduction

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

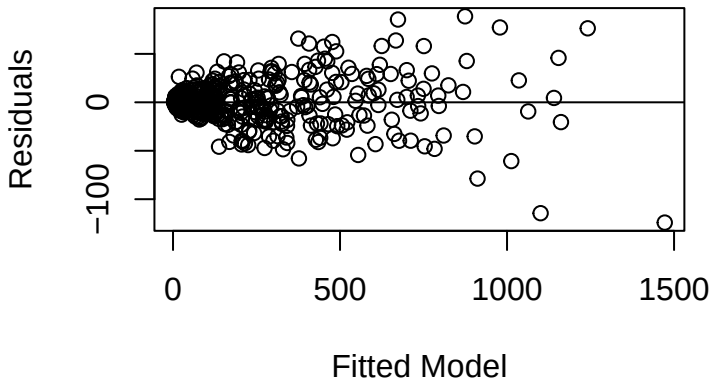
As we rejected H_0 in favor for H_α, we can determine that the reduced model does not model the data well enough to justify the reduction in predictors. As such, we decided to use model 1, the full model, as our computational model.

Statistical Models

```
## Subset selection object
## Call: regsubsets.formula(RUNS ~ HOME_RUNS + TRIPLE + DOUBLE + SINGLES +
##     WALKS + INT_WALKS + STOLEN_BASES + HIT_BY_PITCH, data = batting,
##     method = "seqrep", nbest = 1, nvmax = 4)
## 8 Variables  (and intercept)
##              Forced in Forced out
## HOME_RUNS        FALSE      FALSE
## TRIPLE           FALSE      FALSE
## DOUBLE           FALSE      FALSE
## SINGLES          FALSE      FALSE
## WALKS            FALSE      FALSE
## INT_WALKS        FALSE      FALSE
## STOLEN_BASES     FALSE      FALSE
## HIT_BY_PITCH     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: 'sequential replacement'
##          HOME_RUNS TRIPLE DOUBLE SINGLES WALKS INT_WALKS STOLEN_BASES
## 1  ( 1 ) " "       " "    "*"    " "     " "   " "       " "
## 2  ( 1 ) "*"       " "    " "    "*"     " "   " "       " "
## 3  ( 1 ) "*"       " "    " "    "*"     "*"   " "       " "
## 4  ( 1 ) "*"       " "    " "    "*"     "*"   " "       "*"
##          HIT_BY_PITCH
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
```

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.

Stat model residual plot




```
##
## 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: STAT-MODEL BECAUSE: AdjR^2 is larger and we want to have as little variance as possible.

Model Selection and Analysis

Between the computational and Statistical models, we selected the [TEMP] because [TEMP]. For this model:

- Interpret β_is and intercept. Are they significant? (SAM)

- Report R^2 and adj R^2/interpret/discuss (SAM)
- Complete analysis of residuals and influence points. Use plots Consider refitting the data with points that have large leverage and residuals (KOSYS)
- interpret the model in a way that makes sense. Why do you think some variables dropped out? (KOSYS)
- Give CIs for a mean predicted value and the PIs of a future predicted value for at lease one combination of X's (VALERIA)

Summary

- Summarize (VALERIA)