126 Data Project, Step 2

Sam Ream, Valeria Lopez, Skyler Yee

Sam

Valeria

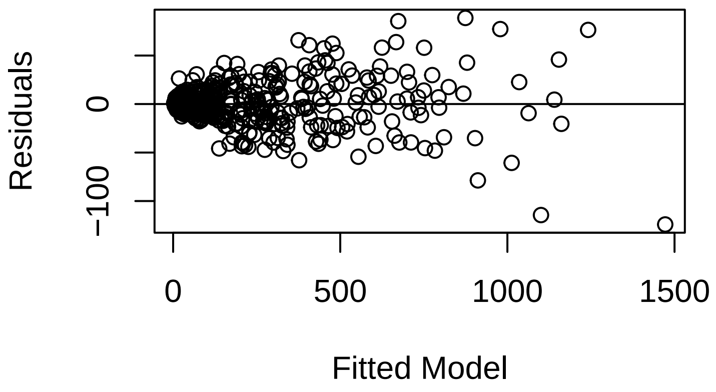
```
## # A tibble: 6 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     0.151   1.68       0.0902 9.28e-  1
## 2 INT_WALKS      -0.125   0.100     -1.25   2.13e-  1
## 3 SINGLES         0.435   0.00983   44.3    5.41e-174
## 4 TRIPLE          1.38    0.223      6.19   1.27e-  9
## 5 STOLEN_BASES    0.389   0.0441     8.81   2.08e- 17
## 6 HOME_RUNS       1.43    0.0350    40.9    1.32e-160
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -1.84    1.77        -1.04 3.00e-  1
## 2 SINGLES        0.458   0.00978     46.8  2.90e-184
## 3 TRIPLE         2.48    0.200       12.4  6.21e- 31
## 4 HOME_RUNS      1.34    0.0323      41.5  3.01e-163
```

```
## 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. Our adjusted R^2 is .9925.

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
```

CI

```
##                   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.

PI

```
##   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.

Skyler