Predictors on the Productivity of Hits in 2017 Major League Baseball



Table of Contents

Chapter 1: *A Multiple Regression Analysis of Batting Average and Hits*

Topic …………………………………………………………………………… 3

Data Source …………………………………………………………………… 3

Variables …………………………………………………………………… 3

Data View …………………………………………………………………… 4

Chapter 2: *A Simple Regression Model*

Scatterplots …………………………………………………………………… 5

Analysis of Scatterplot ………………………………………………………….. 5

The Linear Regression Model …………………………………………………… 5

SAS Output for the Fitted Model ………………………………………………… 6

Analysis of Output ………………………………………………………………. 7

The t-tests ………………………………………………………………... 7

The -equation ………………………………………………………….. 8

Chapter 3: *Matrix Methods*

Simple Linear Regression in Matrix Terms ……………………………………… 9

Multiple Linear Regression in Matrix Terms …………………………………..... 11

Chapter 4: *Model Selection*

Best Subsets Model Selection …………………………………........................... 14

Forward Stepwise Model Selection …………………………………………….. 18

Variance Inflation ………………………………………………………………. 18

Cook’s D ……………………………………………………………………….. 19

Chapter 5: *Cross Validation*

Model Selection with GLMSelect ……………………………………………… 20

Quality of the Fitted Model: Steps for Cross Validation……………………….. 20

Step 1: Splitting the Data ……………………………………………….. 20

Step 2: Fit the Regression Model ……………………………………….. 22

Step 3: Out of Sample Forecasting ……………………………………… 22

Step 4: Compare MSE Values …………………………………………… 23

Source of Cover Photo:

<https://www.freep.com/story/sports/mlb/tigers/2017/10/03/baseball-playoffs/106258104/>

(Photo: Mark J. Terrill, AP)

Chapter 1: *A Multiple Regression Analysis of Batting Average and Hits*

1. Topic

The topic of this subject matter is on the starting lineup for all teams in the 2017 MLB season. I’ve been a baseball fan, specifically of the Yankees, since around 10 years old. I would watch the games with my family and we would cause so much commotion that my mother would think something bad had happened. I became such a giant fan of the Yankees throughout the years that I even got season tickets. Admittingly, I started to grow tired of baseball after that year, maybe it was all the travelling to the Bronx that I had to do, but in recent years my interest has started to slowly grow back. Therefore, for this analysis I’m interested in seeing how specific variables affect hits.

1. Data Source

I was able to pull up the statistic for the starting lineup for 2017 of all Major League Baseball Teams on the following website:

<https://www.baseball-reference.com/leagues/MLB/2017.shtml>

1. Variables

There are a total of 246 observations in this dataset. For this analysis the dependent variable is the total number of hits for each player and the independent variables are batting average, age, strikeouts, and at bats. Definitions for each variable are provided below.

Hits: When the batter safely reaches first base after hitting the ball in fair territory

Batting Average: Number of hits divided by at bats

Age: Age of the batter

Strikeouts: Occurs when the batter racks up 3 strikes for failure to hit the ball into fair territory

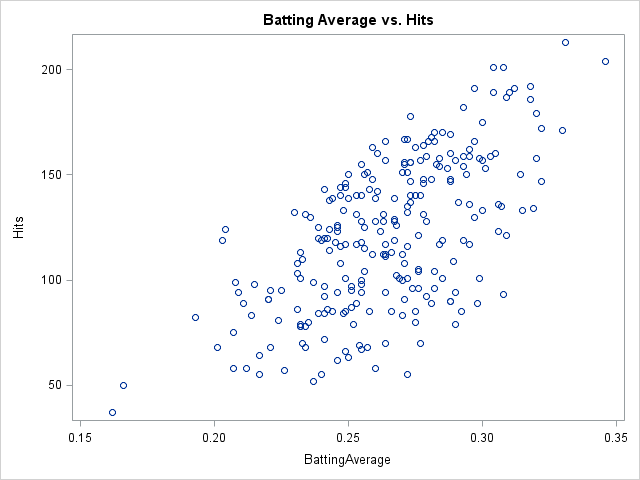
Walks: occurs when a batter receives four pitches outside the strike zone and is in turn awarded first base

1. Data View

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Team** | **Pos** | **Name** | **Age** | **Hits** | **Walks** | **BA** | **SO** |
| Diamondbacks | **C** | [Chris Iannetta](https://www.baseball-reference.com/players/i/iannech01.shtml) | 34 | 69 | 37 | 0.254 | 87 |
| Diamondbacks | **1B** | [Paul Goldschmidt](https://www.baseball-reference.com/players/g/goldspa01.shtml) | 29 | 166 | 94 | 0.297 | 147 |
| Diamondbacks | **2B** | [Brandon Drury](https://www.baseball-reference.com/players/d/drurybr01.shtml) | 24 | 119 | 28 | 0.267 | 103 |
| Diamondbacks | **SS** | [Ketel Marte#](https://www.baseball-reference.com/players/m/marteke01.shtml) | 23 | 58 | 29 | 0.26 | 37 |
| Diamondbacks | **3B** | [Jake Lamb\*](https://www.baseball-reference.com/players/l/lambja01.shtml) | 26 | 133 | 87 | 0.248 | 152 |
| Diamondbacks | **CF** | [A.J. Pollock](https://www.baseball-reference.com/players/p/polloaj01.shtml) | 29 | 113 | 35 | 0.266 | 71 |
| Diamondbacks | **RF** | [David Peralta\*](https://www.baseball-reference.com/players/p/peralda01.shtml) | 29 | 154 | 43 | 0.293 | 94 |
| Braves | **C** | [Tyler Flowers](https://www.baseball-reference.com/players/f/flowety01.shtml) | 31 | 89 | 31 | 0.281 | 82 |
| Braves | **1B** | [Freddie Freeman\*](https://www.baseball-reference.com/players/f/freemfr01.shtml) | 27 | 135 | 65 | 0.307 | 95 |
| Braves | **2B** | [Brandon Phillips](https://www.baseball-reference.com/players/p/phillbr01.shtml) | 36 | 137 | 19 | 0.291 | 57 |
| Braves | **SS** | [Dansby Swanson](https://www.baseball-reference.com/players/s/swansda01.shtml) | 23 | 113 | 59 | 0.232 | 120 |
| Braves | **LF** | [Matt Kemp](https://www.baseball-reference.com/players/k/kempma01.shtml) | 32 | 121 | 27 | 0.276 | 99 |
| Braves | **CF** | [Ender Inciarte\*](https://www.baseball-reference.com/players/i/inciaen01.shtml) | 26 | 201 | 49 | 0.304 | 94 |
| Braves | **RF** | [Nick Markakis\*](https://www.baseball-reference.com/players/m/markani01.shtml) | 33 | 163 | 68 | 0.275 | 110 |
| Orioles | **C** | [Welington Castillo](https://www.baseball-reference.com/players/c/castiwe01.shtml) | 30 | 96 | 22 | 0.282 | 97 |
| Orioles | **1B** | [Chris Davis\*](https://www.baseball-reference.com/players/d/davisch02.shtml) | 31 | 98 | 61 | 0.215 | 195 |
| Orioles | **2B** | [Jonathan Schoop](https://www.baseball-reference.com/players/s/schoojo01.shtml) | 25 | 182 | 35 | 0.293 | 142 |
| Orioles | **SS** | [J.J. Hardy](https://www.baseball-reference.com/players/h/hardyjj01.shtml) | 34 | 55 | 12 | 0.217 | 48 |
| Orioles | **3B** | [Manny Machado](https://www.baseball-reference.com/players/m/machama01.shtml) | 24 | 163 | 50 | 0.259 | 115 |
| Orioles | **LF** | [Trey Mancini](https://www.baseball-reference.com/players/m/mancitr01.shtml) | 25 | 159 | 33 | 0.293 | 139 |
| Orioles | **CF** | [Adam Jones](https://www.baseball-reference.com/players/j/jonesad01.shtml) | 31 | 170 | 27 | 0.285 | 113 |
| Orioles | **RF** | [Seth Smith\*](https://www.baseball-reference.com/players/s/smithse01.shtml) | 34 | 85 | 36 | 0.258 | 79 |
| Orioles | **DH** | [Mark Trumbo](https://www.baseball-reference.com/players/t/trumbma01.shtml) | 31 | 131 | 42 | 0.234 | 149 |
| Red Sox | **C** | [Christian Vazquez](https://www.baseball-reference.com/players/v/vazquch01.shtml) | 26 | 94 | 17 | 0.29 | 64 |

Chapter 2: *A Simple Regression Model*

1. Scatterplots



1. Analysis of Scatterplot

I don’t see obvious curvature in my data. We can see a bit of heteroscedasticity towards the bottom of the scatter plot. The pink line illustrates the vertical cut in the data. It could be that these players didn’t play a full season due to an injury or they might have just had a very bad year. No real leverage or outliers exist in this scatterplot. The two points that are separated by the pink line fall in line with the rest of the data and aren’t dramatically far away from the rest of the data points.

1. The Linear Regression Model

Yx= βo + β1x + ε

1. Yx = subpopulation average, which is normally distributed, of hits that share the same batting average.
2. E(Yx) = the expected value of the random variable, Hits, conditional on knowing the batting average. For example, it is the expected value of the subpopulation average of hits given all the players with 0.250 batting average.
3. V(Yx) = the variance of the random variable hits given the players with a batting average of 0.250.
4. SAS Output for the Fitted Model

The REG Procedure

Dependent Variable: Hits

|  |  |
| --- | --- |
| **Number of Observations Read** | 246 |
| **Number of Observations Used** | 246 |

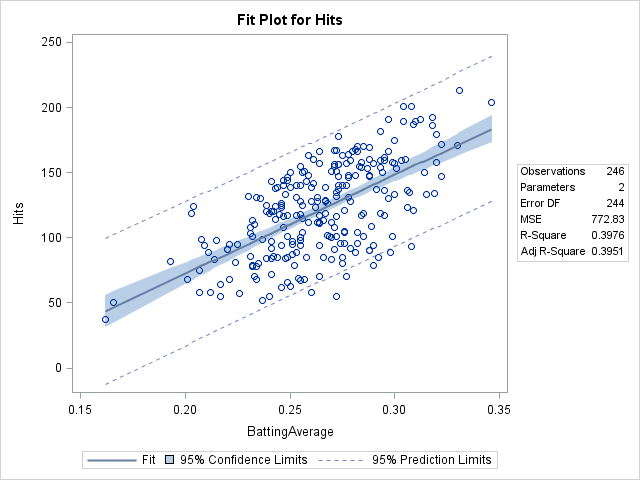
| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 124471 | 124471 | 161.06 | <.0001 |
| **Error** | 244 | 188570 | 772.82946 |  |  |
| **Corrected Total** | 245 | 313042 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 27.79981 | **R-Square** | 0.3976 |
| **Dependent Mean** | 121.21545 | **Adj R-Sq** | 0.3951 |
| **Coeff Var** | 22.93421 |  |  |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | **1** | -79.72118 | 15.93202 | -5.00 | <.0001 |
| **BattingAverage** | BattingAverage | **1** | 760.15209 | 59.89739 | 12.69 | <.0001 |

The REG Procedure

Dependent Variable: Hits



At a given count of x, say in my case is a batting average of 0.250, we can picture a vertical line down that count, illustrated in green. That vertical line intersects the 2 dotted lines. The y coordinates of those intersections, illustrated by the purple dots, define the endpoints for the 95% prediction interval for the number of hits of the 247th (the next observation) player. As the vertical line passes through the shaded band, given by the pink slanted lines, it defines the 95% confidence interval for the subpopulation expected value of hits given a batting average of 0.250. Where the vertical line intersects the line, the red dot, it gives us the point prediction of the number of hits of the 247th player with a batting average of 0.250.

1. Analysis of Output
   1. The t-tests
      1. We are testing if the population slope for hits and batting average is equal to zero.

H0: β1 =0

H1: β1 0

* + 1. Test statistic: t-stat = =

Rejection Region: |12.69| > 1.970, α=0.05

t-critical value with 2 d.f. = 1.970

Conclusion: null hypothesis is rejected because the |t-stat| = |53.19| greater than the t-critical value of 1.970

* + 1. For this hypothesis test, if the null hypothesis were true, then the observed t-statistic of 12.69 is a random value taken from t-distribution with 244 degrees of freedom. If we then computed the t-statistic for this same test for each of the 1200246 samples, we would end up with 1200246 t-statistics. If the null hypothesis were true, 5% percent of those 1200246 t-statistics would lie in the rejection region, which is below the t-critical value of

-1.970 or above the t-critical value of 1.970.

* 1. The equation
     1. The equation for the fitted model is as follows:

= -79.721 + 760.152x

* + 1. The true regression line is μY|x = β0 + β1x. The expected value, E(YX), is also equal to E(Yx) = μY|x. This denotes the subpopulation of hits given all batter with the same batting average, let’s say that is 0.250. To get the point estimate of E(Yx)= β0 + β1x, we use the equation, where = b0 + b1x. The parent population average of hits consists of all Major League Baseball players with a batting average of 0.250, let’s say that is 1200. This parent population has its own μY|x, its own β0 and its own β1. Since we are not able to get all of the averages, slopes and intercepts from the original 1200 population of players who share the same batting average, we must estimate it by using small sample sizes. In this sample of 246 players, we were able to compute a sample slope of 760.152 and a sample intercept of -79.721. But this is just ONE of the many possible samples we could have chosen. Suppose we chose a completely different set of players from the total parent population of 1,200. We most likely would get a completely different sample slope and intercept. In fact, the total number of sample slopes and intercepts of 246 players, sampling without replacement is 1200246 (an enormous number)! This is known as the daughter population of slopes, and intercepts. We can then sum up all the sample slopes and all of the sample intercepts in the daughter population and then divide that sum by 1200246. We denote the population average of the population of sample slopes as and the population average of the population of sample intercepts as μb0. If is equal to , then we say that the sample slope is an unbiased estimator. If μb0 is equal to β0, then we say that the sample slope and intercept is an unbiased estimator. Since we can estimate and β0, we use the equation to give us the point estimate of μY|x. Once we have estimated the , we can now estimate E(Yx) because E(Yx) = μY|x.

Chapter 3: *Matrix Methods*

1. Simple Linear Regression in Matrix Terms
   1. The Matrix below consists of 10 observations: 8 Yankee batters and 2 Met batters. The X vector represents the batting average for the 10 observations
   2. The y-vector represents total hits for the 10 observations.

|  |  |  |
| --- | --- | --- |
| X |  | y |
| 1 | 0.25 | 63 |
| 1 | 0.256 | 104 |
| 1 | 0.278 | 131 |
| 1 | 0.3 | 133 |
| 1 | 0.287 | 153 |
| 1 | 0.273 | 140 |
| 1 | 0.264 | 157 |
| 1 | 0.264 | 94 |
| 1 | 0.284 | 154 |
| 1 | 0.231 | 86 |

|  |  |
| --- | --- |
| X'X |  |
| 10 | 2.687 |
| 2.687 | 0.725627 |
|  |  |
| (X'X)^-1 |  |
| 19.98917 | -74.019999 |
| -74.02 | 275.474505 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X' |  |  |  |  |  |  |  |  |  |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.25 | 0.256 | 0.278 | 0.3 | 0.287 | 0.273 | 0.264 | 0.264 | 0.284 | 0.231 |

* 1. B-Vector, Hat Matrix, vector

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| H |  |  |  |  |  |  |  |  |  |
| 0.1963 | 0.165 | 0.052092 | -0.1 | 0.0057 | 0.078 | 0.124 | 0.1242 | 0.0212 | 0.2942 |
| 0.1654 | 0.144 | 0.067464 | -0 | 0.036 | 0.085 | 0.116 | 0.1164 | 0.0465 | 0.2319 |
| 0.0521 | 0.067 | 0.123826 | 0.18 | 0.1469 | 0.111 | 0.088 | 0.088 | 0.1392 | 0.0034 |
| -0.061 | -0.01 | 0.180188 | 0.37 | 0.2578 | 0.137 | 0.059 | 0.0595 | 0.2319 | -0.2251 |
| 0.0057 | 0.036 | 0.146883 | 0.26 | 0.1923 | 0.122 | 0.076 | 0.0763 | 0.1771 | -0.0901 |
| 0.0778 | 0.085 | 0.111016 | 0.14 | 0.1217 | 0.105 | 0.094 | 0.0944 | 0.1181 | 0.0553 |
| 0.1242 | 0.116 | 0.087959 | 0.06 | 0.0763 | 0.094 | 0.106 | 0.1061 | 0.0802 | 0.1488 |
| 0.1242 | 0.116 | 0.087959 | 0.06 | 0.0763 | 0.094 | 0.106 | 0.1061 | 0.0802 | 0.1488 |
| 0.0212 | 0.046 | 0.139197 | 0.23 | 0.1771 | 0.118 | 0.08 | 0.0802 | 0.1645 | -0.0589 |
| 0.2942 | 0.232 | 0.003416 | -0.2 | -0.09 | 0.055 | 0.149 | 0.1488 | -0.059 | 0.4915 |
|  |  |  |  |  |  |  |  |  |  |
| yhat =Hy |  | b-vector |  | y-hat = Xb |  |  |  |  |  |
| 99.769 |  | -190.753 |  | 99.769 |  |  |  |  |  |
| 106.74 |  | 1162.089 |  | 106.74 |  |  |  |  |  |
| 132.31 |  |  |  | 132.31 |  |  |  |  |  |
| 157.87 |  |  |  | 157.87 |  |  |  |  |  |
| 142.77 |  |  |  | 142.77 |  |  |  |  |  |
| 126.5 |  |  |  | 126.5 |  |  |  |  |  |
| 116.04 |  |  |  | 116.04 |  |  |  |  |  |
| 116.04 |  |  |  | 116.04 |  |  |  |  |  |
| 139.28 |  |  |  | 139.28 |  |  |  |  |  |
| 77.689 |  |  |  | 77.689 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

* 1. Regression of the 10 observations of Hits on Batting Average, which agree with the values given in the above b-vector.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| The SAS System | | |  |  |  |
| The REG Procedure | | |  |  |  |
| Dependent Variable: BattingAverage | | | | |  |
| **Number of Observations Read** | 10 |  |  |  |  |
| **Number of Observations Used** | 10 |  |  |  |  |
|  |  |  |  |  |  |
| **Analysis of Variance** | | |  |  |  |
| **Source** | **DF** | **Sum of** | **Mean** | **F Value** | **Pr > F** |
|  |  | **Squares** | **Square** | |  |
| **Model** | 1 | 4902.273 | 4902 | 8.32 | 0.02 |
| **Error** | 8 | 4716.227 | 590 |  |  |
| **Corrected Total** | 9 | 9618.5 |  |  |  |
|  |  |  |  |  |  |
| **Root MSE** | 24.28 | **R-Square** | 0.51 |  |  |
| **Dependent Mean** | 121.5 | **Adj R-Sq** | 0.45 |  |  |
| **Coeff Var** | 19.98 |  |  |  |  |
|  |  |  |  |  |  |
| **Parameter Estimates** | | |  |  |  |
| **Variable** | **DF** | **Parameter** | **Standard** | **t Value** | **Pr > |t|** |
|  |  | **Estimate** | **Error** | |  |
| **Intercept** | **1** | -190.753 | 109 | -1.76 | 0.117 |
| **Hits** | **1** | 1162.089 | 403 | 2.88 | 0.02 |

* 1. Computing the value of h2,3 and checking that it matches with the hat matrix for the second row third column.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **xbar=** | 0.2687 |  | x | xbar | (x-xbar)^2 |
| **SSx=** | 0.0036301 |  | 0.25 | 0.2687 | 0.00034969 |
| **x2=** | 0.256 |  | 0.256 | 0.2687 | 0.00016129 |
| **x3=** | 0.278 |  | 0.278 | 0.2687 | 8.649E-05 |
|  |  |  | 0.3 | 0.2687 | 0.00097969 |
|  |  |  | 0.287 | 0.2687 | 0.00033489 |
|  |  |  | 0.273 | 0.2687 | 1.849E-05 |
|  |  |  | 0.264 | 0.2687 | 2.209E-05 |
|  |  |  | 0.264 | 0.2687 | 2.209E-05 |
|  |  |  | 0.284 | 0.2687 | 0.00023409 |
|  |  |  | 0.231 | 0.2687 | 0.00142129 |
|  |  |  |  | SSx= | 0.0036301 |
|  |  |  |  |  |  |

h(2,3) = =

|  |  |
| --- | --- |
|  | 0.06746371 |

1. Multiple Linear Regression in Matrix Terms
   1. I added a second regressor, age, to the X matrix

|  |  |  |
| --- | --- | --- |
| The SAS System | | |
|  |  |  |
| **X** |  |  |
| 1 | 0.25 | 28 |
| 1 | 0.256 | 30 |
| 1 | 0.278 | 24 |
| 1 | 0.3 | 27 |
| 1 | 0.287 | 27 |
| 1 | 0.273 | 33 |
| 1 | 0.264 | 33 |
| 1 | 0.264 | 33 |
| 1 | 0.284 | 25 |
| 1 | 0.231 | 37 |

* 1. Using SAS Proc IML, I computed the X matrix and b-vector.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **mat1** |  |  |  |  |  |  |  |  |  |
| 10 | 2.687 | 297 |  |  |  |  |  |  |  |
| 2.687 | 0.726 | 79.281 |  |  |  |  |  |  |  |
| 297 | 79.28 | 8979 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **mat2** |  |  |  |  |  |  |  |  |  |
| 76.51653 | -193 | -0.82636 |  |  |  |  |  |  |  |
| -193.054 | 526.1 | 1.740132 |  |  |  |  |  |  |  |
| -0.82636 | 1.74 | 0.01208 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **mat3** |  |  |  |  |  |  |  |  |  |
| 5.114923 | 2.304 | 3.014864 | -3.7 | -1.202 | -3.457 | -1.72 | -1.72 | 1.0302 | 1.3457 |
| -12.7969 | -6.16 | -5.02572 | 11.8 | 4.9299 | 8.005 | 3.27 | 3.2696 | -0.1288 | -7.1322 |
| -0.05308 | -0.02 | -0.05268 | 0.02 | -8E-04 | 0.047 | 0.032 | 0.0317 | -0.0302 | 0.0226 |

|  |
| --- |
| **b** |
| -361.94 |
| 1522.569 |
| 2.502538 |

* 1. Using SAS Proc Reg, I checked the values in the Proc IML b-vector

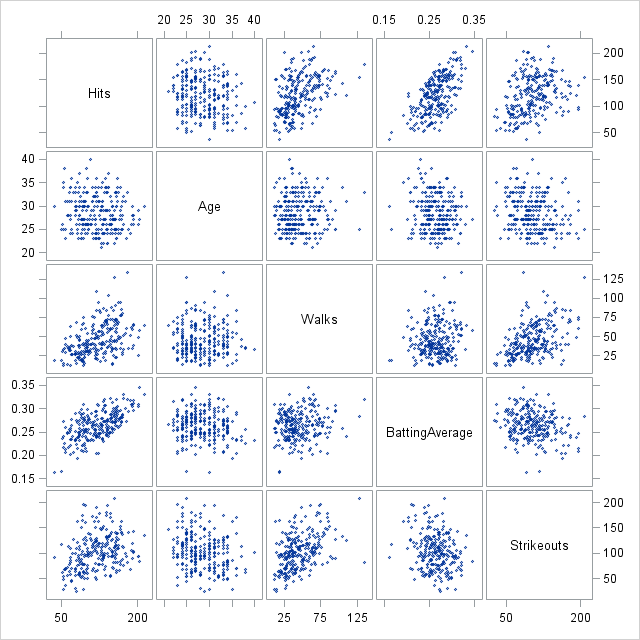
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| The SAS System |  |  |  |  |  |
| The REG Procedure | |  |  |  |  |
| Dependent Variable: Hits | |  |  |  |  |
| **Number of Observations Read** | 10 |  |  |  |  |
| **Number of Observations Used** | 10 |  |  |  |  |
|  |  |  |  |  |  |
| **Analysis of Variance** | |  |  |  |  |
| **Source** | **DF** | **Sum of** | **Mean** | **F Value** | **Pr > F** |
|  |  | **Squares** | **Square** | |  |
| **Model** | 2 | 5420.69 | 2710 | 4.52 | 0.055 |
| **Error** | 7 | 4197.81 | 600 |  |  |
| **Corrected Total** | 9 | 9618.5 |  |  |  |
|  |  |  |  |  |  |
| **Root MSE** | 24.49 | **R-Square** | 0.56 |  |  |
| **Dependent Mean** | 121.5 | **Adj R-Sq** | 0.44 |  |  |
| **Coeff Var** | 20.16 |  |  |  |  |
|  |  |  |  |  |  |
| **Parameter Estimates** | |  |  |  |  |
| **Variable** | **DF** | **Parameter** | **Standard** | **t Value** | **Pr > |t|** |
|  |  | **Estimate** | **Error** | |  |
| **Intercept** | **1** | -361.94 | 214 | -1.69 | 0.135 |
| **BattingAverage** | **1** | 1522.569 | 562 | 2.71 | 0.03 |
| **Age** | **1** | 2.50254 | 2.69 | 0.93 | 0.383 |

* 1. Commands used to compute vectors.

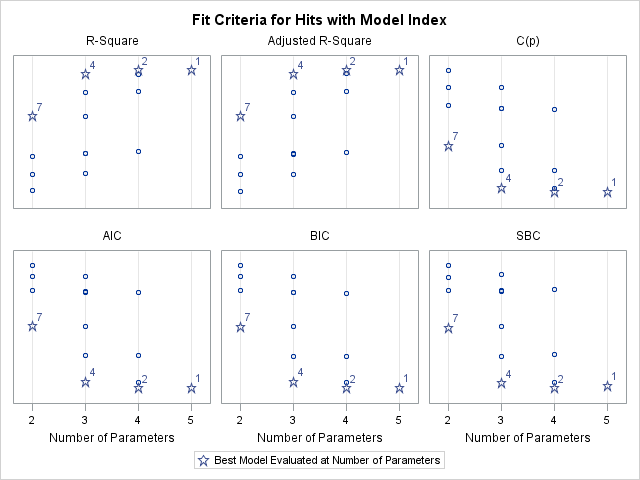
|  |  |  |
| --- | --- | --- |
| **data** bbdata; |  |  |
| input BattingAverage Age Hits; | | |
| datalines; |  |  |
| 0.250 28 063 |  |  |
| 0.256 30 104 |  |  |
| 0.278 24 131 |  |  |
| 0.300 27 133 |  |  |
| 0.287 27 153 |  |  |
| 0.273 33 140 |  |  |
| 0.264 33 157 |  |  |
| 0.264 33 094 |  |  |
| 0.284 25 154 |  |  |
| 0.231 37 086 |  |  |
| ; |  |  |
| **Proc** **IML**; |  |  |
| use bbdata; |  |  |
| read all; |  |  |
| one\_vec = j(**10**,**1**); | |  |
| X = one\_vec || BattingAverage || Age; | | |
| print X; |  |  |
| Y = Hits; |  |  |
| mat1 = t(X) \* X; | |  |
| mat2=  inv(t(X) \* X); | |  |
| mat3 = inv(t(X) \* X) \* t(X); | | |
| print mat1, mat2, mat3; | |  |
| b = inv(t(X) \* X) \* (t(X)\*Y); | | |
| print b; |  |  |
| **proc** **reg** data=bbdata; | |  |
| model Hits = BattingAverage Age; | | |
| **run**; |  |  |

Chapter 4 *Model Selection*

1. Best Subsets Model Selection
   1. Matrix Scatterplot of the Full Data Set

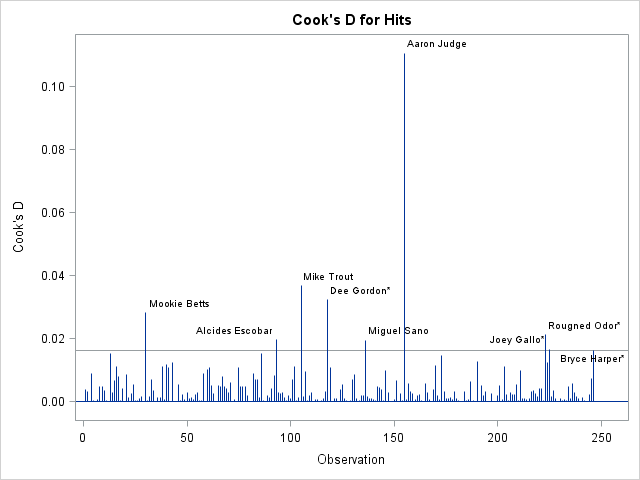


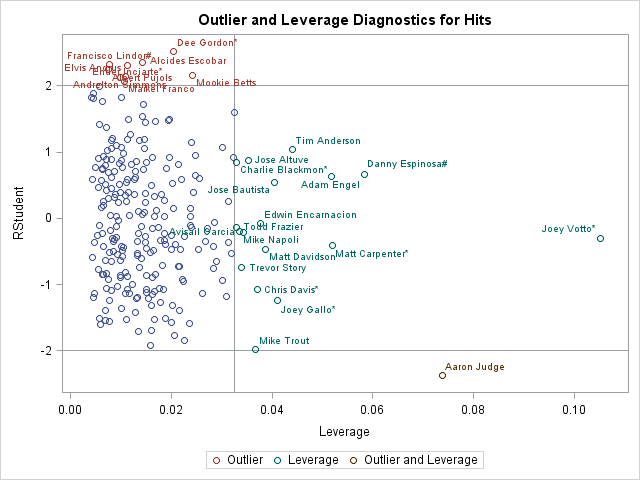
* 1. I tried taking log and square root transformations, but I didn’t notice any obvious improvements in the data. In the scatterplot showing Hits and Age, I don’t see an obvious relationship. It seems to be negatively correlated but it isn’t a very strong one. The scatterplot showing Hits vs. Walks there are a couple of leverage points which I have circled in purple. They lie a bit far in the horizontal direction, away from the rest of the data.
  2. The Criteria Plot and the Summary Table for all possible models are shown below.



| **Model Index** | **Number in Model** | **Adjusted R-Square** | **R-Square** | **C(p)** | **AIC** | **SBC** | **Variables in Model** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | **4** | **0.6335** | 0.6395 | 5.0000 | 1517.5917 | 1535.11837 | Age Walks BattingAverage Strikeouts |
| **2** | **3** | **0.6332** | 0.6377 | 4.2511 | 1516.8655 | 1530.88683 | Walks BattingAverage Strikeouts |
| **3** | **3** | **0.6156** | 0.6203 | 15.8711 | 1528.3910 | 1542.41236 | Age BattingAverage Strikeouts |
| **4** | **2** | **0.6120** | 0.6152 | 17.2726 | 1529.6652 | 1540.18121 | BattingAverage Strikeouts |
| **5** | **3** | **0.5233** | 0.5292 | 76.7919 | 1581.3019 | 1595.32326 | Age Walks BattingAverage |
| **6** | **2** | **0.5216** | 0.5255 | 77.2403 | 1581.2079 | 1591.72387 | Walks BattingAverage |
| **7** | **1** | **0.3951** | 0.3976 | 160.7316 | 1637.9062 | 1644.91686 | BattingAverage |
| **8** | **2** | **0.3946** | 0.3996 | 161.4194 | 1639.1034 | 1649.61935 | Age BattingAverage |
| **9** | **3** | **0.2110** | 0.2207 | 283.0428 | 1705.2664 | 1719.28770 | Age Walks Strikeouts |
| **10** | **2** | **0.2040** | 0.2105 | 287.8255 | 1706.4480 | 1716.96404 | Walks Strikeouts |
| **11** | **2** | **0.2023** | 0.2088 | 288.9921 | 1706.9911 | 1717.50712 | Age Walks |
| **12** | **1** | **0.1881** | 0.1914 | 298.6146 | 1710.3375 | 1717.34815 | Walks |
| **13** | **1** | **0.0968** | 0.1005 | 359.3555 | 1736.5315 | 1743.54213 | Strikeouts |
| **14** | **2** | **0.0961** | 0.1034 | 359.4086 | 1737.7338 | 1748.24976 | Age Strikeouts |
| **15** | **1** | **0.0102** | 0.0142 | 417.0640 | 1759.0736 | 1766.08422 | Age |

* 1. The model I prefer includes Walks, Batting Average and Strikeouts. C(p), AIC, BIC, and SBC, all agree that a model with 4 parameters is best, which means the model has 3 regressors. Next to the star aligned with the 4 parameters, a 2 is shown. Therefore, looking at the Summary Table above Model number 2 shows that the best model includes, Walks, Batting Average, and Strikeouts.
  2. Below are the diagnostic plots for the model selected above. The plots show the flagged players according to Cook’s D and Outlier and Leverage Plot. Aaron Judge seems to be the most obvious influential point according to both diagnostic plots.





1. Forward Stepwise Model Selection
   1. According to the summary of the Stepwise Model Selection, the first variable selected was Batting Average. Then it added Strikeouts and did not remove Batting Average. Next, it added Walks and did not remove any of the previous variables. Finally, the selection method stopped and did not add the last variable which is Age.

| **Summary of Stepwise Selection** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Entered** | **Variable Removed** | **Label** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | BattingAverage |  | BattingAverage | 1 | 0.3976 | 0.3976 | 160.732 | 161.06 | <.0001 |
| **2** | Strikeouts |  | Strikeouts | 2 | 0.2176 | 0.6152 | 17.2726 | 137.39 | <.0001 |
| **3** | Walks |  | Walks | 3 | 0.0225 | 0.6377 | 4.2511 | 15.01 | 0.0001 |
|  |  |  |  |  |  |  |  |  |  |

* 1. The Stepwise Selection method choose the same model as the best subsets method.

1. Variance Inflation
   1. Variance Inflation, VIFk = (1 – R2k)-1, shows us if multicollinearity exists in our analysis. If the VIF is high, the variance of the standard errors of the slope is increased because of collinearity. It helps us identify if we are regressing one variable on another almost identical to itself.

| **Parameter Estimates** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | Intercept | **1** | -170.83548 | 20.56859 | -8.31 | <.0001 | 0 |
| **Walks** | Walks | **1** | 0.27410 | 0.07640 | 3.59 | 0.0004 | 1.38322 |
| **BattingAverage** | BattingAverage | **1** | 848.75620 | 50.71901 | 16.73 | <.0001 | 1.18345 |
| **Strikeouts** | Strikeouts | **1** | 0.41565 | 0.04839 | 8.59 | <.0001 | 1.51444 |
| **Age** | Age | **1** | 0.42872 | 0.38328 | 1.12 | 0.2644 | 1.11984 |
|  |  |  |  |  |  |  |  |

* 1. All variables are close to 1. Therefore, there are no outrageous values.

1. Cook’s D
   1. Cook’s D measures if a data point is an outlier and a leverage (influential) point that may affect the accuracy of the regression analysis we are running. It is computed as follows: for i=1, compute  for all n cases, but without the first case in the fitting. Take the sum of the squares of the resulting residuals and divide by the MSE times p. Do this for i values. The formula is provided below.

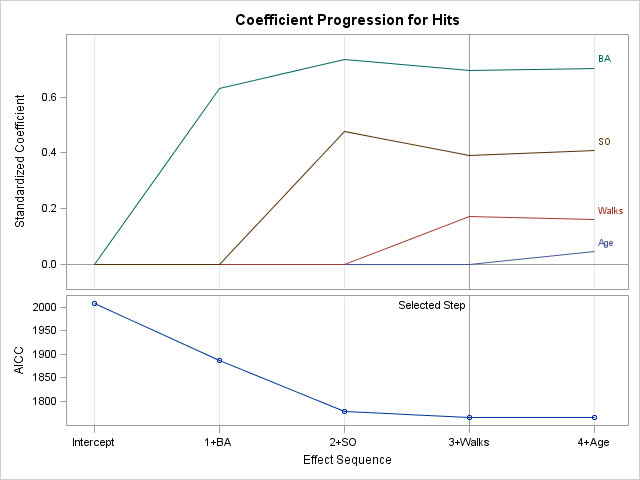
Di = 

* 1. The 4/246 cut off for my data set equals 0.016. I have a few data points, labeled by player, that have been flagged for influence. Those players include; Mookie Betts, Alcides Escobar, Mike Trout, Dee Gordon, Miguel Sano, Aaron Judge, Joey Gallo, Rougned Odor, and Bryce Harper.

Chapter 5***Cross-Validation***

1. Model Selection with GLMSelect

I used Proc GLMSelect in SAS, with AICC as a selection criterion. The first variable selected was Batting Average, then it chose Strikeouts, next it chose Walks. The bold line at 3 indicates that those three variables are the best model for Hits.



1. Quality of the Fitted Model: Steps for Cross Validation

I used Cross Validation to judge the quality of the fitted model shown above.

* 1. Step 1: Splitting the Data

The first step I took was to split the existing data into two groups. I did this by first randomizing the order of the 246 batters in my data, and then I split them into two groups. The first group, known as the training group, consists of 100 batters and the second group, known as the validation group, consists on 146 batters.

* + 1. Below, I am showing the first 13 observations. As you can see, they have not been randomized yet, but the random numbers have been attached in order for them to be set up for randomizing.

| **Obs** | **Team** | **Name** | **Age** | **Hits** | **Walks** | **BA** | **SO** | **RandNum** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Diamondbacks | Chris Iannetta | 34 | 69 | 37 | 0.254 | 87 | 0.73902 |
| **2** | Diamondbacks | Paul Goldschmidt | 29 | 166 | 94 | 0.297 | 147 | 0.27248 |
| **3** | Diamondbacks | Brandon Drury | 24 | 119 | 28 | 0.267 | 103 | 0.70953 |
| **4** | Diamondbacks | Ketel Marte# | 23 | 58 | 29 | 0.260 | 37 | 0.31916 |
| **5** | Diamondbacks | Jake Lamb\* | 26 | 133 | 87 | 0.248 | 152 | 0.36785 |
| **6** | Diamondbacks | A.J. Pollock | 29 | 113 | 35 | 0.266 | 71 | 0.10449 |
| **7** | Diamondbacks | David Peralta\* | 29 | 154 | 43 | 0.293 | 94 | 0.03680 |
| **8** | Braves | Tyler Flowers | 31 | 89 | 31 | 0.281 | 82 | 0.53333 |
| **9** | Braves | Freddie Freeman\* | 27 | 135 | 65 | 0.307 | 95 | 0.37130 |
| **10** | Braves | Brandon Phillips | 36 | 137 | 19 | 0.291 | 57 | 0.04019 |
| **11** | Braves | Dansby Swanson | 23 | 113 | 59 | 0.232 | 120 | 0.79345 |
| **12** | Braves | Matt Kemp | 32 | 121 | 27 | 0.276 | 99 | 0.95910 |
| **13** | Braves | Ender Inciarte\* | 26 | 201 | 49 | 0.304 | 94 | 0.64804 |
|  |  |  |  |  |  |  |  |  |

* + 1. Now I sorted the data by the order given by the random number.

| **Obs** | **Team** | **Name** | **Age** | **Hits** | **Walks** | **BA** | **SO** | **RandNum** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Mets | Travis d'Arnaud | 28 | 85 | 23 | 0.244 | 59 | 0.00325 |
| **2** | Blue Jays | Russell Martin | 34 | 68 | 50 | 0.221 | 83 | 0.01891 |
| **3** | Blue Jays | Steve Pearce | 34 | 79 | 27 | 0.252 | 68 | 0.02734 |
| **4** | Phillies | Tommy Joseph | 25 | 119 | 33 | 0.240 | 129 | 0.02845 |
| **5** | Reds | Scott Schebler\* | 26 | 110 | 39 | 0.233 | 125 | 0.03258 |
| **6** | Cardinals | Kolten Wong\* | 26 | 101 | 41 | 0.285 | 60 | 0.03532 |
| **7** | Diamondbacks | David Peralta\* | 29 | 154 | 43 | 0.293 | 94 | 0.03680 |
| **8** | Nationals | Daniel Murphy\* | 32 | 172 | 52 | 0.322 | 77 | 0.03759 |
| **9** | Braves | Brandon Phillips | 36 | 137 | 19 | 0.291 | 57 | 0.04019 |
| **10** | Rockies | Gerardo Parra\* | 30 | 121 | 20 | 0.309 | 67 | 0.04056 |
| **11** | Indians | Carlos Santana# | 31 | 148 | 88 | 0.259 | 94 | 0.04127 |
| **12** | Rockies | Charlie Blackmon\* | 30 | 213 | 65 | 0.331 | 135 | 0.04220 |
| **13** | Red Sox | Mitch Moreland\* | 31 | 125 | 57 | 0.246 | 120 | 0.04444 |

* + 1. Next, I split the data into two groups, the training sample (of size 100) and the validation sample (of size 146). Below, I have shown the last 10 observations of each group.

**Training Sample**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **90** | Braves | Nick Markakis\* | 33 | 163 | 68 | 0.275 | 110 | 0.33911 | 90 |
| **91** | Angels | Andrelton Simmons | 27 | 164 | 47 | 0.278 | 67 | 0.34854 | 91 |
| **92** | Marlins | Derek Dietrich\* | 27 | 101 | 36 | 0.249 | 98 | 0.34880 | 92 |
| **93** | Rangers | Elvis Andrus | 28 | 191 | 38 | 0.297 | 101 | 0.35749 | 93 |
| **94** | White Sox | Tim Anderson | 24 | 151 | 13 | 0.257 | 162 | 0.36064 | 94 |
| **95** | Brewers | Travis Shaw\* | 27 | 147 | 60 | 0.273 | 138 | 0.36090 | 95 |
| **96** | Pirates | David Freese | 34 | 112 | 58 | 0.263 | 116 | 0.36556 | 96 |
| **97** | Diamondbacks | Jake Lamb\* | 26 | 133 | 87 | 0.248 | 152 | 0.36785 | 97 |
| **98** | Mariners | Robinson Cano\* | 34 | 166 | 49 | 0.280 | 85 | 0.37035 | 98 |
| **99** | Braves | Freddie Freeman\* | 27 | 135 | 65 | 0.307 | 95 | 0.37130 | 99 |
| **100** | Mets | Lucas Duda\* | 31 | 62 | 37 | 0.246 | 73 | 0.37164 | 100 |

**Validation Sample**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **136** | Twins | Byron Buxton | 23 | 117 | 38 | 0.253 | 150 | 0.96634 | 236 |
| **137** | Tigers | Mikie Mahtook | 27 | 96 | 23 | 0.276 | 79 | 0.96645 | 237 |
| **138** | Royals | Jorge Bonifacio | 24 | 98 | 35 | 0.255 | 118 | 0.96738 | 238 |
| **139** | Twins | Joe Mauer\* | 34 | 160 | 66 | 0.305 | 83 | 0.96920 | 239 |
| **140** | Indians | Bradley Zimmer\* | 24 | 72 | 26 | 0.241 | 99 | 0.97225 | 240 |
| **141** | Dodgers | Joc Pederson\* | 25 | 58 | 39 | 0.212 | 68 | 0.97797 | 241 |
| **142** | Padres | Manuel Margot | 22 | 128 | 35 | 0.263 | 106 | 0.98274 | 242 |
| **143** | Yankees | Aaron Judge | 25 | 154 | 127 | 0.284 | 208 | 0.98738 | 243 |
| **144** | Mariners | Danny Valencia | 32 | 115 | 40 | 0.256 | 122 | 0.99184 | 244 |
| **145** | Angels | Kole Calhoun\* | 29 | 139 | 71 | 0.244 | 134 | 0.99899 | 245 |
| **146** | Rays | Steven Souza Jr. | 28 | 125 | 84 | 0.239 | 179 | 0.99966 | 246 |

* 1. Step 2: Fit the Regression Model

Now I fit the regression model on the training sample.

| **Obs** | **b0** | **b1** | **b2** | **b3** | **IntMSE** |
| --- | --- | --- | --- | --- | --- |
| **1** | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 |
|  |  |  |  |  |  |

* 1. Step 3: Out of Sample Forecasting

Next, I used the resulting equation to forecast the number of hits for the Validation Sample. Below, I have shown the last 10 observations of the equation Validation Sample.

| **Obs** | **Team** | **Name** | **Age** | **Hits** | **Walks** | **BA** | **SO** | **RandNum** | **NewInd** | **b0** | **b1** | **b2** | **b3** | **IntMSE** | **i** | **yhat** | **ExtRes** | **SqExtRes** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **136** | Twins | Byron Buxton | 23 | 117 | 38 | 0.253 | 150 | 0.96634 | 236 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 236 | 129.416 | -12.4155 | 154.15 |
| **137** | Tigers | Mikie Mahtook | 27 | 96 | 23 | 0.276 | 79 | 0.96645 | 237 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 237 | 111.785 | -15.7845 | 249.15 |
| **138** | Royals | Jorge Bonifacio | 24 | 98 | 35 | 0.255 | 118 | 0.96738 | 238 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 238 | 115.308 | -17.3082 | 299.58 |
| **139** | Twins | Joe Mauer\* | 34 | 160 | 66 | 0.305 | 83 | 0.96920 | 239 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 239 | 154.511 | 5.4894 | 30.13 |
| **140** | Indians | Bradley Zimmer\* | 24 | 72 | 26 | 0.241 | 99 | 0.97225 | 240 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 240 | 90.857 | -18.8565 | 355.57 |
| **141** | Dodgers | Joc Pederson\* | 25 | 58 | 39 | 0.212 | 68 | 0.97797 | 241 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 241 | 55.104 | 2.8960 | 8.39 |
| **142** | Padres | Manuel Margot | 22 | 128 | 35 | 0.263 | 106 | 0.98274 | 242 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 242 | 116.884 | 11.1164 | 123.57 |
| **143** | Yankees | Aaron Judge | 25 | 154 | 127 | 0.284 | 208 | 0.98738 | 243 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 243 | 215.004 | -61.0036 | 3721.44 |
| **144** | Mariners | Danny Valencia | 32 | 115 | 40 | 0.256 | 122 | 0.99184 | 244 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 244 | 119.798 | -4.7985 | 23.03 |
| **145** | Angels | Kole Calhoun\* | 29 | 139 | 71 | 0.244 | 134 | 0.99899 | 245 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 245 | 125.447 | 13.5530 | 183.68 |
| **146** | Rays | Steven Souza Jr. | 28 | 125 | 84 | 0.239 | 179 | 0.99966 | 246 | -179.334 | 892.943 | 0.46401 | 0.34825 | 389.806 | 246 | 146.390 | -21.3901 | 457.54 |

* 1. Step 4: Compare MSE values

In order to measure the quality of fit, we need to compare the Internal MSE (MSE for the training sample) with the External MSE (MSE for the validation sample). If the training MSE and the validation MSE are similar in value, then we can say that this is evidence for validating the chosen model.

| **Obs** | **\_TYPE\_** | **\_FREQ\_** | **IntMSE** | **ExtMSE** |
| --- | --- | --- | --- | --- |
| **1** | 0 | 146 | 389.806 | 539.500 |

Since, the slope estimates are computed in such a fashion as to minimize the Internal MSE, we must expect that the External MSE will be larger than the Internal MSE. We can use the following equation to account for how much larger the External MSE will be from the internal MSE:

Now, I can compare that with the Internal and External MSE:

They seem to be close in value, therefore the model has been validated.

**Appendix**

**#check contents of data**

**proc** **contents** data=bbdata;

**run**;

**#input variables and print data**

**data** bbdata;

input Team $**1**-**12** Pos $**1**-**2** Name $**1**-**19** Age Hits Walks BattingAverage SO

run;

**proc** **print** data=bbdata;

**run**;

**#Scatterplot**

**proc** **sgplot** data=bbdata;

scatter x=BattingAverage y=Hits;

**run**;

**#Regression and b-vector**

|  |  |
| --- | --- |
| **data** bbdata; |  |
| input BattingAverage Age Hits; | | |
| datalines; |  |
| 0.250 28 063 |  |
| 0.256 30 104 |  |
| 0.278 24 131 |  |
| 0.300 27 133 |  |
| 0.287 27 153 |  |
| 0.273 33 140 |  |
| 0.264 33 157 |  |
| 0.264 33 094 |  |
| 0.284 25 154 |  |
| 0.231 37 086 |  |
| ; |  |
| **Proc** **IML**; |  |
| use bbdata; |  |
| read all; |  |
| one\_vec = j(**10**,**1**); | |
| X = one\_vec || BattingAverage || Age; | | |
| print X; |  |
| Y = Hits; |  |
| mat1 = t(X) \* X; | |
| mat2=  inv(t(X) \* X); | |
| mat3 = inv(t(X) \* X) \* t(X); | | |
| print mat1, mat2, mat3; | |
| b = inv(t(X) \* X) \* (t(X)\*Y); | | |
| print b; |  |
| **proc** **reg** data=bbdata; | |
| model Hits = BattingAverage Age; | | |
| **run**; |  |

**#Scatterplot Matrix**

**proc** **sgscatter** data=bbdata;

matrix Hits Age Walks BattingAverage Strikeouts;

**run**;

**#Model Selection**

**proc** **reg** plots(label) = criteria;

model Hits = Age Walks BattingAverage Strikeouts /

selection = adjrsq cp aic sbc;

**run**;

**proc** **reg** data=bbdata;

model Hits = Age Walks BattingAverage Strikeouts /

selection = stepwise details=summary;

**run**;

**#VIF**

**proc** **reg** data=bbdata;

model Hits = Age Walks BattingAverage Strikeouts / VIF;

**run**;

**#CooksD, Outlier, and Leverage**

**proc** **reg** plots(label)=(CooksD RStudentByLeverage);

id Name;

model Hits = Age Walks BattingAverage Strikeouts;

**run**;

**#GLM Select and Cross Validation**

**proc** **glm** data=bbdata plots=(coefficient criteria);

model Hits = Age Walks BA SO

selection = forward(stop=none choose=AICC);

**run**;

dm "out; clear; log; clear";

%Let DataLen = 246;

%Let Split = 100;

**Data** MultData;

Set bbdata;

RandNum = RanUni(**123456**); \*Adds a random # to each obs.;

**proc** **print**;

**Proc** **Sort**;

By RandNum;

**Proc** **Print**;

**Run**;

**Data** MultData;

Set MultData;

NewInd+**1**;

**Proc** **Print**; **Run**;

**Data** Train;

Set MultData;

if NewInd < (&Split+**1**);

**proc** **print**; **run**;

**Data** Test;

Set MultData;

if NewInd > &Split;

**proc** **print**; **run**;

**Proc** **Reg** Data=Train outest=OutA tableout;

model NewHits = BA SO Walks;

**run**;

**proc** **print** Data=OutA;

**Data** Slopes;

Set OutA;

where \_type\_ = "PARMS";

IntMSE = \_RMSE\_\*\***2**;

Keep IntMSE b0 b1 b2 b3;

**run**;

**proc** **print**; **run**;

**Data** SlopesM;

Set Slopes;

do i = (&Split+**1**) to &DataLen;

output;

end;

**proc** **print**; **run**;

**Data** Test;

Set Test;

Merge SlopesM;

yhat = b0 + b1\*BA + b2\*SO +b3\*Walks;

ExtRes = NewHits - yhat;

SqExtRes= ExtRes\*\***2**;

**proc** **print**; **run**;

**proc** **means** Data=Test mean noprint;

var IntMSE SqExtRes;

output out=summary

Mean = IntMSE ExtMSE;

**proc** **print**; **run**; quit;