Sport Analytics: A case study of baseball based on Hitters dataset

Regression Analysis 5020
Proposal

By: Group 20 (Iman, Mostafa, Olusegun, Virtus)

1. Definition of the problem

As Baseball is loved by most people in the United States, it has a very high turnover. A lot of people, especially sports team owners, are curious to estimate what influences the salary and placement of players in American (A) and National (N) leagues. The Hitters is one the most trusted dataset in this field. In this work, we investigate the dataset using Regression Analysis approaches to predict the salary of baseball players and identify the most significant variables in predicting the salary of the baseball players. Several algorithms have been used including multiple linear regression, polynomial regression, stepwise regression. To validate the calculations, the 10-fold cross-validation technique is used. For regression models, all independent variables are used to estimate the response variable. Obviously, the New League dependent variable is dropped in regression models and 17 other independent variables are used to estimate the response variable New League.

2. The project and our specialization

In a ranking done by Rulesofsport.com in 2021, the baseball job market ranks second as the highest-paid sport in the world, running closely behind Basketball. We hope that such projects will prepare us for this job market.

3. Hitters dataset

The Hitters is a dataset from the ISLR package with 20 variables and 322 observations. Of these 20 variables, we would use 18 of them as predictors and two as dependent. One the dependent is

continuous (Salary), and the other one is categorical (New League). Of those 18 independent variables, 16 are continuous, and two are categorical (League and Division).

Hitters {ISLR} R Documentation Baseball Data
Description: Major League Baseball Data from the 1986 and 1987 seasons.

Variable	Description
AtBat	Number of times at bat in 1986
Hits	Number of hits in 1986
HmRun	Number of home runs in 1986
Runs	Number of runs in 1986
RBI	Number of runs batted in in 1986
Walks	Number of walks in 1986
Years	Number of years in the major leagues
CAtBat	Number of times at bat during his career
CHits	Number of hits during his career
CHmRun	Number of home runs during his career
CRuns	Number of runs during his career
CRBI	Number of runs batted in during his career
CWalks	Number of walks during his career
League	A factor with levels A and N indicating player's league at the end of 1986
Division	A factor with levels E and W indicating player's division at the end of 1986
PutOuts	Number of put outs in 1986
Assists	Number of assists in 1986
Errors	Number of errors in 1986
Salary	1987 annual salary on opening day in thousands of dollars
NewLeague	A factor with levels A and N indicating player's league at the beginning of 1987

'data.frame': **322 obs. of 20 variables**:
$ AtBat : int 293 315 479 496 321 594 185 298 323 401 ...
$ Hits : int 66 81 130 141 87 169 37 73 81 92 ...
$ HmRun : int 1 7 18 20 10 4 1 0 6 17 ...
$ Runs : int 30 24 66 65 39 74 23 24 26 49 ...
$ RBI : int 29 38 72 78 42 51 8 24 32 66 ...
$ Walks : int 14 39 76 37 30 35 21 7 8 65 ...
$ Years : int 1 14 3 11 2 11 2 3 2 13 ...
$ CAtBat : int 293 3449 1624 5628 396 4408 214 509 341 5206 ...
$ CHits : int 66 835 457 1575 101 1133 42 108 86 1332 ...
$ CHmRun : int 1 69 63 225 12 19 1 0 6 253 ...
$ CRuns : int 30 321 224 828 48 501 30 41 32 784 ...
$ CRBI : int 29 414 266 838 46 336 9 37 34 890 ...
$ CWalks : int 14 375 263 354 33 194 24 12 8 866 ...
$ League : Factor w/ 2 levels "A","N": 1 2 1 2 2 1 2 1 2 1 ...
$ Division : Factor w/ 2 levels "E","W": 1 2 2 1 1 2 1 2 2 1 ...
$ PutOuts : int 446 632 880 200 805 282 76 121 143 0 ...
$ Assists : int 33 43 82 11 40 421 127 283 290 0 ...
$ Errors : int 20 10 14 3 4 25 7 9 19 0 ...
$ **Salary** : num NA 475 480 500 91.5 750 70 100 75 1100 ...
$ **NewLeague**: Factor w/ 2 levels "A","N": 1 2 1 2 2 1 1 1 2 1 …