Model Selection in R for MLB Wins as a Function of Team Batting Statistics

Thomas Maguire

4/17/2020

**Abstract:**

The purpose of this analysis is to use four regression model selection techniques to analyze potential models for how to forecast wins in Major League Baseball. Each model will be comprised of Team Batting Statistics and will aim to project wins. Model selection will be performed on 4 seasons (1989,1999,2009,2019) to analyze the differences between the models in each year, and to make a larger statement about the changes in Major League Baseball across the decades.

**Introduction:**

Major League Baseball is a multi-billion-dollar industry comprised of 30 Major League teams, and hundreds of associated Minor League teams. Each one of these teams competes for the chance to win a World Series Championship, which can only be obtained by winning baseball games during the regular season. Due to the difficulty of making the Postseason, Major League teams must put heavy emphasis on winning as many games as possible, and thus need to construct their teams in ways that will maximize their wins.

In this analysis, Model Selection techniques will be used to find models to forecast Wins as a function of teams’s batting statistics. These forecasts will use statistics from four seperate Major League seasons (1989,1999,2009,2019). The “All Possible Regression” technique will be used to select a model for the 2019 season. The “Best Subset” model selection technique will be used to select a model for the 2009 season. The “Stepwise Forward Regression” model selection technique will be used to select the model for 1999. And finally, the “Backwards Selection Regression” model selection technique will be used to select the model for 1989.

The variables in each of these four models can then be compared to analyze the changes in the ways winning teams construct their rosters as time has passed. Based on the cited literature, as well as examining the statistics, I believe there will be several differences in the models from each season. Most noticeably, I believe there will be a large difference in the variables from the 1999 model and the 2009 model, as baseball changed “eras” and shifted team building strategies.

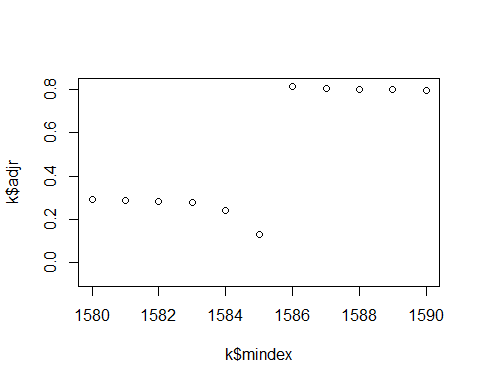
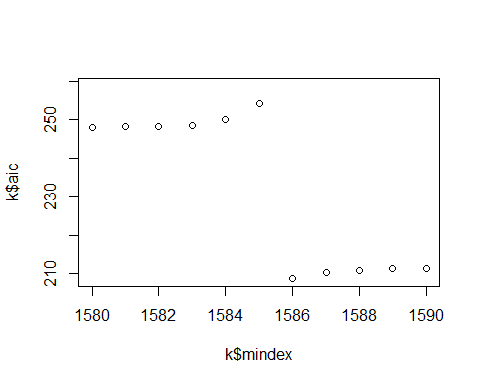
**Materials and Methods:**

The first model selection technique used is the “All Possible Regression” Technique used on the 2019 MLB Batting Statistics to forecast wins.

lmod19 <- lm(Wins~ AB+ H + X1B + X2B + X3B+ HR + R + RBI + BB + SO + SB + AVG, data=MLB2019 )  
k <- ols\_step\_all\_possible(lmod19, print\_plot=TRUE)

Below is shown the Model number, variables, and Selection Criteria of the best model for the 2019 season as selected using the “All Possible Regression Technique.”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model # | # of Variables | Variables | R-Square | Adj R-Square |
| 1586 | 6 | AB H X1B X3B BB AVG | 0.8516829561 | 0.812991553 |



These two graphs above show the Adjusted R-Square Values as well as the AIC Criteria values of a small number of the models generated by the selection process. As shown in the graphs, Model #1568 has both the highest R-Square Value and the lowest AIC, so it is the best model for the 2019 data.

The second model selection technique used is the “Best Subset” Model Selection Technique used on the 2009 MLB Batting Statistics to forecast wins.

lmod09 <-lm(Wins~ AB+ H + X1B + X2B + X3B+ HR + R + RBI + BB + SO + SB + AVG, data=MLB2009 )  
j <- ols\_step\_best\_subset(lmod09,print\_plot = TRUE)

## Best Subsets Regression   
## -----------------------------------------------------  
## Model Index Predictors  
## -----------------------------------------------------  
## 1 R   
## 2 X2B R   
## 3 X2B R RBI   
## 4 X2B HR R RBI   
## 5 AB X1B HR R RBI   
## 6 AB X1B HR R RBI AVG   
## 7 AB H X1B HR R RBI AVG   
## 8 AB H X1B HR R RBI BB AVG   
## 9 AB H X1B HR R RBI BB SB AVG   
## 10 AB H X1B HR R RBI BB SO SB AVG   
## 11 AB X1B X2B X3B HR R RBI BB SO SB AVG   
## 12 AB H X1B X2B X3B HR R RBI BB SO SB AVG   
## -----------------------------------------------------  
## Subsets Regression Summary   
## ---------------------------------------------------------------------------------------------------------------------------------  
## Adj. Pred   
## Model R-Square R-Square R-Square C(p) AIC SBIC SBC MSEP FPE HSP APC   
## ---------------------------------------------------------------------------------------------------------------------------------  
## 1 0.3737 0.3513 0.2931 2.0766 222.2842 NA 226.4878 2545.1692 90.4800 3.1417 0.7158   
## 2 0.4154 0.3721 0.3064 2.2053 222.2150 NA 227.8198 2466.9017 90.3141 3.1578 0.7145   
## 3 0.4639 0.4021 0.2929 2.0299 221.6152 NA 228.6212 2352.6045 88.6084 3.1274 0.7010   
## 4 0.5199 0.4431 0.3394 1.5206 220.3064 NA 228.7136 2194.7184 84.9568 3.0342 0.6721   
## 5 0.5542 0.4613 0.3444 1.9826 220.0820 NA 229.8903 2126.4714 84.5199 3.0623 0.6687   
## 6 0.5649 0.4514 0.3153 3.5038 221.3545 NA 232.5641 2169.8701 88.4730 3.2607 0.6999   
## 7 0.5796 0.4458 0.2853 4.8459 222.3251 NA 234.9359 2196.5184 91.7900 3.4508 0.7262   
## 8 0.5901 0.4339 0.2506 6.3763 223.5680 NA 237.5800 2248.8668 96.2320 3.7012 0.7613   
## 9 0.5974 0.4162 0.2187 8.0492 225.0291 NA 240.4423 2325.0882 101.7896 4.0180 0.8053   
## 10 0.5982 0.3867 0.1603 10.0124 226.9680 NA 243.7823 2449.2615 109.6019 4.4554 0.8671   
## 11 0.5984 0.3531 0.0694 12.0000 228.9473 NA 247.1629 2591.5507 118.4311 4.9761 0.9370   
## 12 0.5984 0.3531 0.0694 12.0000 230.9473 NA 250.5641 2591.5507 118.4311 4.9761 0.9370   
## ---------------------------------------------------------------------------------------------------------------------------------

As shown in the summary above, Model #5 has both the highest Adjusted R-Square value as well as the lowest AIC criteria value, making it the best choice for modeling the 2009 data.

The third model selection technique used is the “Stepwise Forward” Regression Model Selection Technique used on the 1999 MLB Batting Statistics to forecast wins.

lmod99 <-lm(Wins~ AB+ H + X1B + X2B + X3B+ HR + R + RBI + BB + SO + SB + AVG, data=MLB1999 )  
m <- ols\_step\_forward\_p(lmod99,details = TRUE)

## Forward Selection Method   
## We are selecting variables based on p value...  
##   
## Forward Selection: Step 1   
##   
## - RBI   
##   
## Model Summary   
## --------------------------------------------------------------  
## R 0.706 RMSE 9.020   
## R-Squared 0.499 Coef. Var 11.149   
## Adj. R-Squared 0.481 MSE 81.353   
## Pred R-Squared 0.446 MAE 6.620   
## --------------------------------------------------------------  
## Forward Selection: Step 2   
##   
## - BB   
##   
## Model Summary   
## --------------------------------------------------------------  
## R 0.764 RMSE 8.368   
## R-Squared 0.584 Coef. Var 10.343   
## Adj. R-Squared 0.553 MSE 70.020   
## Pred R-Squared 0.499 MAE 6.370   
## --------------------------------------------------------------  
##

## Forward Selection: Step 3   
##   
## - SB   
##   
## Model Summary   
## --------------------------------------------------------------  
## R 0.796 RMSE 8.008   
## R-Squared 0.633 Coef. Var 9.898   
## Adj. R-Squared 0.591 MSE 64.126   
## Pred R-Squared 0.526 MAE 6.066   
## --------------------------------------------------------------  
##   
## Forward Selection: Step 4   
##   
## - X2B   
##   
## Model Summary   
## --------------------------------------------------------------  
## R 0.828 RMSE 7.550   
## R-Squared 0.686 Coef. Var 9.332   
## Adj. R-Squared 0.636 MSE 56.996   
## Pred R-Squared 0.565 MAE 5.429   
## --------------------------------------------------------------  
##   
## Forward Selection: Step 5   
##   
## - X1B   
##   
## Model Summary   
## --------------------------------------------------------------  
## R 0.843 RMSE 7.398   
## R-Squared 0.711 Coef. Var 9.144   
## Adj. R-Squared 0.651 MSE 54.728   
## Pred R-Squared 0.562 MAE 5.665   
## --------------------------------------------------------------  
##   
## Forward Selection: Step 6   
##   
## - SO   
##   
## Model Summary   
## --------------------------------------------------------------  
## R 0.852 RMSE 7.350   
## R-Squared 0.727 Coef. Var 9.085   
## Adj. R-Squared 0.655 MSE 54.016   
## Pred R-Squared 0.545 MAE 5.668   
## --------------------------------------------------------------  
## Variables Entered:   
##   
## + RBI   
## + BB   
## + SB   
## + X2B   
## + X1B   
## + SO   
##   
##   
## Final Model Output   
## ------------------  
##   
## Model Summary   
## --------------------------------------------------------------  
## R 0.852 RMSE 7.350   
## R-Squared 0.727 Coef. Var 9.085   
## Adj. R-Squared 0.655 MSE 54.016   
## Pred R-Squared 0.545 MAE 5.668   
## --------------------------------------------------------------  
##   
## ANOVA   
## --------------------------------------------------------------------  
## Sum of   
## Squares DF Mean Square F Sig.   
## --------------------------------------------------------------------  
## Regression 3300.330 6 550.055 10.183 0.0000   
## Residual 1242.370 23 54.016   
## Total 4542.700 29   
## --------------------------------------------------------------------  
## Parameter Estimates   
## ------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## ------------------------------------------------------------------------------------------  
## (Intercept) 31.681 48.907 0.648 0.524 -69.491 132.854   
## RBI 0.065 0.030 0.397 2.174 0.040 0.003 0.127   
## BB 0.051 0.023 0.354 2.217 0.037 0.003 0.099   
## SB 0.109 0.044 0.293 2.485 0.021 0.018 0.200   
## X2B 0.120 0.079 0.218 1.512 0.144 -0.044 0.284   
## X1B -0.056 0.031 -0.262 -1.822 0.082 -0.120 0.008   
## SO -0.022 0.019 -0.180 -1.147 0.263 -0.062 0.018   
## ------------------------------------------------------------------------------------------

##   
## Selection Summary   
## -------------------------------------------------------------------------  
## Variable Adj.   
## Step Entered R-Square R-Square C(p) AIC RMSE   
## -------------------------------------------------------------------------  
## 1 RBI 0.4986 0.4807 16.8823 221.0305 9.0196   
## 2 BB 0.5838 0.5530 11.5905 217.4391 8.3678   
## 3 SB 0.6330 0.5906 9.3871 215.6686 8.0079   
## 4 X2B 0.6863 0.6361 6.8247 212.9564 7.5496   
## 5 X1B 0.7109 0.6506 6.7269 212.5135 7.3979   
## 6 SO 0.7265 0.6552 7.3882 212.8437 7.3496   
## -------------------------------------------------------------------------

Based on the results of this technique, it appears as though the best model for forecasting wins with the 1999 season data uses 5 variables, as shown above.

The fourth model selection technique used is the “Backward Elimination” Model Selection Technique used on the 1989 MLB Batting Statistics to forecast wins.

lmod89 <-lm(Wins~ AB+ H + X1B + X2B + X3B+ HR + R + RBI + BB + SO + SB + AVG, data=MLB1989 )  
n <- ols\_step\_backward\_p(lmod89,details = TRUE)

## Backward Elimination Method   
## We are eliminating variables based on p value...

## Note: model has aliased coefficients  
## sums of squares computed by model comparison

## - AVG   
##   
## Backward Elimination: Step 1   
##   
## Variable AVG Removed   
##   
## Model Summary   
## ---------------------------------------------------------------  
## R 0.743 RMSE 8.628   
## R-Squared 0.552 Coef. Var 10.667   
## Adj. R-Squared 0.254 MSE 74.439   
## Pred R-Squared -0.503 MAE 5.206   
## ---------------------------------------------------------------

## - AB   
##   
## Backward Elimination: Step 2   
##   
## Variable AB Removed   
##   
## Model Summary   
## ---------------------------------------------------------------  
## R 0.743 RMSE 8.362   
## R-Squared 0.552 Coef. Var 10.338   
## Adj. R-Squared 0.299 MSE 69.926   
## Pred R-Squared -0.357 MAE 5.170   
## ---------------------------------------------------------------  
## - BB   
##   
## Backward Elimination: Step 3   
##   
## Variable BB Removed   
##   
## Model Summary   
## ---------------------------------------------------------------  
## R 0.726 RMSE 8.331   
## R-Squared 0.527 Coef. Var 10.300   
## Adj. R-Squared 0.304 MSE 69.412   
## Pred R-Squared -0.239 MAE 5.343   
## ---------------------------------------------------------------

## No more variables satisfy the condition of p value = 0.3  
##   
## Variables Removed:   
##   
## - AVG   
## - AB   
## - BB   
##   
## Final Model Output   
## ------------------

## Model Summary   
## ---------------------------------------------------------------  
## R 0.726 RMSE 8.331   
## R-Squared 0.527 Coef. Var 10.300   
## Adj. R-Squared 0.304 MSE 69.412   
## Pred R-Squared -0.239 MAE 5.343   
## ---------------------------------------------------------------  
## Parameter Estimates   
## --------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## --------------------------------------------------------------------------------------------  
## (Intercept) -107.296 72.607 -1.478 0.158 -260.483 45.890   
## H -0.102 0.191 -0.737 -0.537 0.598 -0.504 0.300   
## X1B 0.139 0.165 0.827 0.842 0.411 -0.209 0.487   
## X2B 0.091 0.153 0.260 0.596 0.559 -0.232 0.415   
## X3B 0.030 0.209 0.029 0.142 0.889 -0.412 0.471   
## HR NA 0.310 0.944 1.619 0.124 NA NA   
## R 0.502 0.333 -2.008 -1.178 0.255 -0.152 1.156   
## RBI -0.392 0.036 0.349 1.995 0.062 -1.095 0.310   
## SO 0.071 0.059 0.901 1.850 0.082 -0.004 0.146   
## SB 0.109 NA -0.331 NA NA -0.015 0.234   
## --------------------------------------------------------------------------------------------

## Elimination Summary   
## -------------------------------------------------------------------------  
## Variable Adj.   
## Step Removed R-Square R-Square C(p) AIC RMSE   
## -------------------------------------------------------------------------  
## 1 AVG 0.5524 0.254 10.0081 197.5429 8.6278   
## 2 AB 0.5515 0.2992 8.0362 195.5949 8.3622   
## 3 BB 0.527 0.3044 6.8039 194.9794 8.3314   
## -------------------------------------------------------------------------

Based on the results from using the Backward Elimination technique, only 3 of the 12 variables were removed, leaving the remaining 9 variables as the best model for forecasting wins using the 1989 season data.

**Results:**

After using each of the four model selection techniques on the four data sets some conclusions can be made about forecasting wins in Major League Baseball. When conducting the All Possible Regression technique to select a model for the 2019 season, it was determined that the combination of the variables: AB, H, X1B, X3B, BB, and AVG were the best predictors of wins. For the 2009 season data, the Best Subset model selection technique was used and determined that the subset of: AB, X1B, HR, R, and RBI was best for forecasting. The Stepwise Forward Regression model selection technique was then used for the 1999 season data. By using this technique, it was determined that the variables: RBI, BB, SB, X2B, X1B, and SO were the best variables to forecast wins. Finally, the Backward Elimination selection technique was used to forecast the 1989 season’s wins, for this model, the variables: H, X1B, X2B, X3B, HR, R, RBI, SO, and SB created the best forecasts. After examining each of the years’ models, it appears as though the model selected by for the 2019 season by the All Possible Regressions technique resulted in the model with the highest R-Square value, yet the model has one of the lowest amount of variables in the model. What is extremely interesting however, is that each model has a drastically different combination of variables for predicting wins.

**Discussion:**

In this analysis, four models were selected using four different techniques for four different seasons. Each one of these models were selected as the best possible combination of the variables available, yet this combination, and how it changes across time, is very interesting, and potentially important. For example, the variable AVG, which represents batting average, is only prevalent in one of the four models. Historically, when analyzing the value of a batter, batting average is considered an extremely important statistic. However, as these models show, teams’ batting averages may not affect their number of wins as previously thought. This brings up the question of why players are being paid for their abilities to produce high batting averages if those numbers do not directly lead to wins for their teams. This may be a cause for concern for team owners, and fans, alike.

While AVG is barely represented, the variable, X1B, which represents singles, is in all four of the selected models. And more interestingly, the variable, runs, as denoted by R, is only in two of the models selected. In order to win a baseball game, teams need to score more runs than their opponents, so by Runs not appearing in each model, this may lead some to believe that scoring more runs may not be important. However, the lack of Runs prevalent may be indicative of things that affect baseball games outside of the performances of batters. For example, Runs were not selected to be used in the model for 2019, however the 2019 baseball season saw historic numbers of runs being scored. So, the lack of runs in the model may have less to do with the number of runs being scored, but rather the importance of the runs being scored. In baseball winning by one run or winning by 10 runs makes no difference in the number of wins a team has, so it is possible that teams are winning by less runs overall more recently. This may be an interesting topic for future research as it could potentially show a shift in the way baseball games are being played.

**Literature Cited:**

*Batting Leaders: Team Stats, 1989-2019* (2020, April). Fan Graphs. Retrieved from <https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=0&type=0&season=1989&month=0&season1=1989&ind=0&team=0,ts&rost=0&age=0&filter=&players=0>

Soto Valero, C. (2016). Predicting Win-Loss outcomes in MLB regular season games – A comparative study using data mining methods, International Journal of Computer Science in Sport, 15(2), 91-112. doi: <https://doi.org/10.1515/ijcss-2016-0007>

Sports Reference LLC. "1989-2019 MLB Standings." Baseball-Reference.com - Major League Statistics and Information. https://www.baseball-reference.com/.

W. D. Kaigh (1995) Forecasting Baseball Games, CHANCE, 8:2, 33-37, DOI: [10.1080/09332480.1995.10542458](https://doi.org/10.1080/09332480.1995.10542458)