DSC 205 Final Project Report

Title: How Do Game Statistics Affect Player Salary?

Team Members: Rachel Matherly and Princess Allotey

Background

Many young athletes aspire to play professionally in the future. Not only because they get to continue playing the sport they love, but also because there is the assumption that great players (i.e. players that have good statistics) are paid more. However, is this assumption really true? A recent news article from the *Sports Illustrated* magazine features the top 10 highest contracts made for the 2020 season. It also discusses each player's high game statistics. In the same vein, one might wonder which player statistics are most significant in determining their salary. To investigate this, we obtained two datasets with game statistics on 672 baseball players from the year 2010. We obtained the *2010MLBSalary.csv* dataset from the OpenIntro website and the *MLBSTATS2010.csv* dataset from the Rotowire website. We then merged these datasets, based on player name, to create a usable dataset called *fulldata.csv*. The following are the variables contained within our merged dataset.

- Salary: yearly salary in thousands of dollars
- Team: out of 30 Major League Baseball (MLB) teams
- Pos: field position (C,P, 1B, 2B, 3B, SS, OF, DH)
- Age
- G: number of games played
- AB: at-bats
- R: runs
- H: hits
- X2B: doubles
- X3B: triples
- HR: homeruns
- RBI: runs-batted-in
- SB: stolen bases
- CS: caught stealing
- SH: sacrifice hits
- SF: sacrifice flyouts
- HBP: hit-by-pitch
- AVG: batting average
- OBP: on-base percentage
- SLG: slugging percentage
- OPS: the sum of on-base percentage and slugging average
- SO: Strikeouts
- BB: Walks

Questions

To explore the dataset, we focused on the questions below:

1. Can we create a model to determine the salary of baseball players based on their statistics?
2. How effective are each of our predictor variables at determining salary?
3. What predictor variables are most significant in determining salary?
4. How can we change our model to decrease error?

Methods

Ultimately, we wanted to find the best linear model to predict the salary of a Major League Baseball (MLB) player. To achieve this goal, we created different linear models: incorporating concepts on first-order models, interactions, and transformation. We then used F-tests for model utility, t-tests, stepwise regressions, and nested F-test. We also examined our

variables for multicollinearity, and plotted each of our predictor variables against salary. This approach assisted with deciding whether to transform any of our variables non-linearly in order to find ways to improve our model.

Conclusions

We created a first-order linear model, using all of the statistics listed above, to calculate a player's salary. We performed an F-test for model utility on this model with our null hypothesis being all of our βi=0 and alternative hypothesis being at least one βi is non-zero. The p-value for this test was < 2.2e-16 meaning we could reject the null hypothesis and conclude that our model was useful in predicting MLB player salary at the .05 significance level. Following this, we performed t-tests on each of our βi's to test whether or not each of our predictor variables were useful in predicting salary. Our hypotheses for these tests were Ho: βi=0 and Ha: βi≠0. However, we were able to reject the null hypothesis and conclude that βi was significant in predicting salary at the .05 significance level in only nine out of our fifty-four βi cases .

Therefore, we decided to use stepwise regression to determine which predictor variables were most significant. We used all three types of stepwise regression (forward selection, backward selection, and forward and backward selection), and by looking at the Akaike Information Criterion (AIC) of each variable, were able to determine that the variables below were most significant in determining salary in our forward and backward selection and forward selection. Our backward selection included two other predictor variables (SB and CS), but because these variables had rather high AICs, we decided to only use the significant variables determined previously.

- RBI
- Age
- G
- BB
- AB
- Pos
- SO
- SH
- H
- AVG

With this new linear model, we decided to look at the correlations between the predictor variables to see if there were any signs of multicollinearity. Through this, we detected that games and RBIs were highly correlated with many of the other predictor variables. We decided to remove games and RBIs from our model. Then, we used a nested F-test to determine if this removal further improved our model. Our hypotheses for this test were Ho: $\beta_{games} = \beta_{RBI} = 0$ and Ha: Either β_{games} or β_{RBI} are non-zero. The F-statistic for this test was 22.481 and our p-value 5.359e-10 meaning we could conclude that the removal of those two predictor variables improved our model at the .05 significance level.

With this new model, we performed another forward and backward selection stepwise regression to see if we could further narrow our variables down to improve our model. In doing so, we found that we should remove AB and AVG. We called this model our simple model.

Next, we wanted to see if there were any interactions between our simple model predictor variables that were significant, so we created a model where that included all two-way interactions of each variable with one other. Using t-tests on each of our βi's, we tested to see which of our predictor variable interactions were significant in predicting salary. Our hypotheses for these tests were Ho: βi=0 and Ha: βi≠0. We found two interactions in which the p-value was

lower than the significance level of .05, meaning we could reject the null hypothesis and conclude that those interactions were significant. Those interactions were age and SH as well as age and hits. We then added those two interactions to our simple model and performed another nested F-test to see if our model was improved by those interactions with hypotheses Ho:β_{Age*SH}= β_{Age*H}= 0 and Ha: Either β_{Age*SH} or β_{Age*H} are non-zero. The F-statistic for this test was 10.822 and the p-value was 2.355e-08. Therefore, we could reject the null hypothesis and conclude that these two interactions improved our model at the .05 significance level.

After that, we then plotted all of our predictor variables to see if any transformations could be performed on those variables in order to improve our model like in Figures 1-3 below. Because we did not see any clear trends that were non-linear between salary and our predictor variables, we decided not to make a new linear model with transformations.

Figure 1: Hits vs. Salary

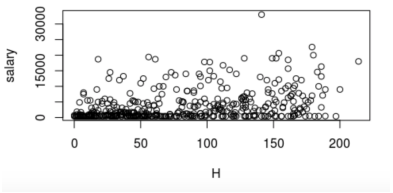


Figure 2: SO vs. Salary

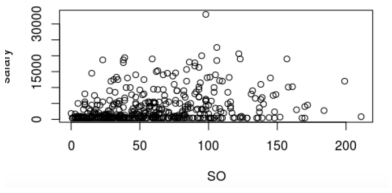
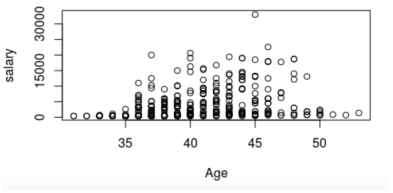


Figure 3: Age vs. Salary



So, we used the improved model with the two interaction terms to create prediction intervals for the salaries of some popular player in the MLB. Table 1 below shows the results of these intervals.

Table 1: Predicting Salary using improved linear model with interactions

Player Name	Fit	Lower Prediction Interval	Upper Prediction Interval	Actual
Derek Jeter	14708.66	7206.272	22211.04	22600
Drew Stubbs	1780.55	-5598.187	9159.288	400
Hunter Pence	3641.753	-3703.747	10987.25	3500
Ichiro Suzuki	14060.83	6513.436	21608.21	18000

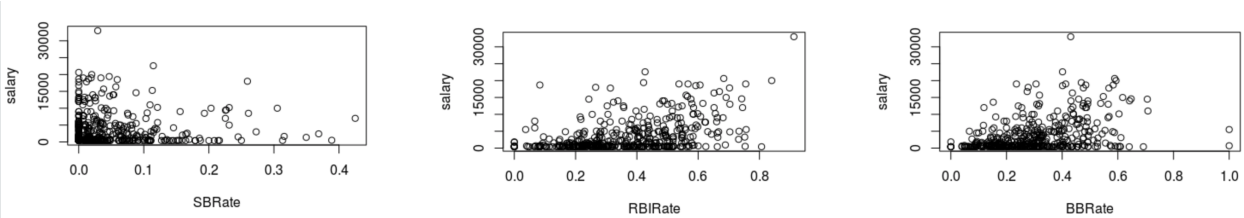
*All values are represented in thousands of dollars

As you can see, these intervals did a good job predicting some salaries, but others did not do so. Because of this, we decided we needed to modify our model. We concluded that it would be more fair for the variables to be represented as rates per game. So, instead of using hits, we used hits per game and so on with each variable. This was necessary because if a player does not play in as many games, he does not have the chance to build up his statistics like a player who plays in every game. So, this transforms levels all players to the same playing field.

As we did before, we created a linear model using the rate variables and did a forward and backward selection stepwise regression to see which variables were most significant in determining salary. When looking at the AIC, we determined that RBI rate, Age, BB rate, position, hit rate, OPS, and SB rate were significant.

After that, we looked at the correlation between each predictor variable and plotted them against salary as in Figures 4-6 below to determine if we needed to remove or transform any of our predictor variables to further improve our model. We did not detect any multicollinearity or find that we needed to transform any of our variables, so our model remained the same.

Figure 4: SB Rate vs. Salary **Figure 5: RBI Rate vs. Salary** **Figure 6: BB Rate vs. Salary**



However, we performed t-tests to determine if each of our βi's were significant in determining salary with hypotheses Ho: βi=0 and Ha: βi≠0. We found that in the case of five of our position dummy variables that our conclusion was to fail to reject the null hypothesis meaning we did not have enough evidence to conclude that those βi's were significant in determining salary. Because of this, we decided to look at our model, with position removed, and conduct a nested F-test. Our hypotheses were Ho: All $\beta_{position}=0$ and Ha: At least one $\beta_{position}$ is non-zero. Our F-statistic for this test was 4.0235 and our p-value was 0.0006232. So, we rejected the null hypothesis and concluded that the removal of position improved our model at the .05 significance level.

One last time, we did a forward and back regression to see if any of our remaining variables needed to be removed. This allowed us to remove SB rate from our model which would make sense because not all batters get the opportunity to sacrifice bunt since the coach or manager decides when players should do so. Unfortunately, though, when we performed a nested F-test to see if this removal improved our model, the p-value was 0.3257, meaning we had to fail to reject the null hypothesis and concluded that we did not have enough evidence to suggest that this removal improved our model at the .05 significance level.

Finally, with our rate model including SB rate, we created prediction intervals for the same popular players as before. Our new predictions are shown in Table 2 below. These findings were not as successful. Before, 3 out of 4 of our players' actual salary was contained in the interval we found, and the fourth one was only 400 thousand off from the interval. In our rate model, however, only 2 of the players' actual salary was contained in the interval, and the other two were more than 4 million away from the interval.

Table 2: Predicting Salary using rate model including SB rate

Player Name	Fit	Lower Confidence Interval	Upper Confidence Interval	Actual

Derek Jeter	8276.075	642.4682	15909.68	22600
Drew Stubbs	3253.329	-4400.862	10907.52	400
Hunter Pence	4537.788	-3079.254	12154.83	3500
Ichiro Suzuki	6291.166	-1453.367	14035.7	18000

*All values are represented in thousands of dollars

Therefore, our final conclusion was that our linear model that we used in Table 1 was our best model in predicting salary for MLB players in 2010, which can be proven by looking at the correlation coefficients (R^2) for each of these models. R^2 for the model used in Table 2 were 0.3726 meaning only 37.26% of the variability of salary is accounted for when using this, but R^2 for the model used in Table 1 was 0.4337 which is an improvement.

Limitations

Despite our significant findings, our study had some potential limitations. Even though our dataset focused on 672 players from 2010, the salaries stated for each player might have represented amounts from contracts signed a number of years before 2010. Therefore, these amounts were estimates of the actual salaries baseball players obtained in 2010.

In addition, our final dataset had 432 total entries since we limited our exploration to only batters. Therefore, our model might not adequately predict the salary of non-batters.

However, we found this model to be useful in predicting salaries of MLB players in 2010 and it might provide some indication of salaries in future years, but we can not say it will do as well at predicting salaries because financial circumstances of the league and individual teams change from year to year.

References

1. OpenIntro. (2011). *Salary data for Major League Baseball (2010)*. Retrieved from https://www.openintro.org/data/index.php?data=mlb
2. Rotowire. (2020). *2010 MLB Player Stats*. Retrieved from https://www.rotowire.com/baseball/stats.php?season=2010
3. Sports Illustrated (2020). *Richest MLB Contracts: Mookie Betts Trails Mike Trout, Passes Bryce Harper*. Retrieved from https://www.si.com/mlb/2020/07/22/richest-mlb-contracts-mookie-betts-bryce-harper-mike-trout