Development and Analysis of a Model for Predicting Salaries in Major League Baseball

Zach Cairney, Molly Creagar, Collin Dougherty, Jesse Osnes
University of Nebraska-Lincoln

Abstract: In this paper, MLB player statistics from the 1986 season and career statistics for those players' were analyzed to determine which statistics and facts were most helpful in predicting salaries in the 1987 season. We also analyzed the statistics to understand which statistic a player should have attempted to improve if he wanted to raise his salary for the 1987 season. Lastly, we investigated if a player's current season statistics or career statistics were more important in determining the salary for the upcoming season. We used All Possible Subsets variable selection and identified the five variables that comprise the best model. Career batting average is found to have the largest estimated effect on salary, and career statistics in general are more important in efforts to model the effect of statistics on salary. We conclude with a discussion on the findings and possible steps to take in future studies.

Keywords: multiple linear regression, baseball, model analysis

1. Introduction

Major League Baseball (MLB) is a multi-billion dollar industry, and salaries of MLB players account for a non-trivial proportion of that amount. To that end, there is great interest and importance in understanding what a player is worth based on his performance. Teams interested in garnering good players want to give reasonable offers, and players and their agents want to make sure they know what they should be asking for.

Statistics are gathered on every MLB player for every aspect of the game. Based on these statistics, we can develop some notion of how skilled a player is, and thus, we can start to consider a player's relative worth based on his performance in previous seasons. Additionally, the career statistics of a player may be of interest as players with strong career statistics may have larger fan bases who would want to attend games in which their favorite player may make an appearance. There are also other factors, such as which league and division a player is in, that may affect the salary based on the economic situation of the particular league or division in a given year.

In this paper, we used MLB player statistics from the 1986 season and career statistics for those players' to determine what statistic was most predictive of the salary, as well as whether career or current season statistics were more important. We introduce the methods used in this analysis in Section 2 and provide the results in Section 3. Finally, we conclude this report with a discussion on the impacts of the findings and what future work should be done in Section 4.

1.1. *Research Questions*

Research Question 1: Which variable has the largest estimated effect on the salary? Is this also the most statistically significant?

Research Question 2: What variables are most important in predicting salaries?

Research Question 3: Are career statistics or previous year statistics more important in determining

a baseball player's salary?

2. Methods

The data was obtained from [1]. There were 322 observations; however, only 263 of those observations were complete. The 59 incomplete cases were removed from the data set before analysis took place. The data contained statistics for current season and career at bats, hits, runs, runs batted in (RBIs), home runs, and walks, as well as the number of years the player has been in the MLB. The response variable in question is the player's salary in 1987 on opening day in thousand of dollars. There is also information about the player's league in 1986 (American or National), the player's current division (East or West), and the player's league in 1987.

Data Visualization

We analyzed the model for predicting salary using all 19 predictor variables. Figure 1 shows the relationships between which league the players are part of in 1986, in 1987, and the division plotted against the respective salaries. There does not seem to be much distinction between the groups, but there are quite a few salary outliers present in all cases.

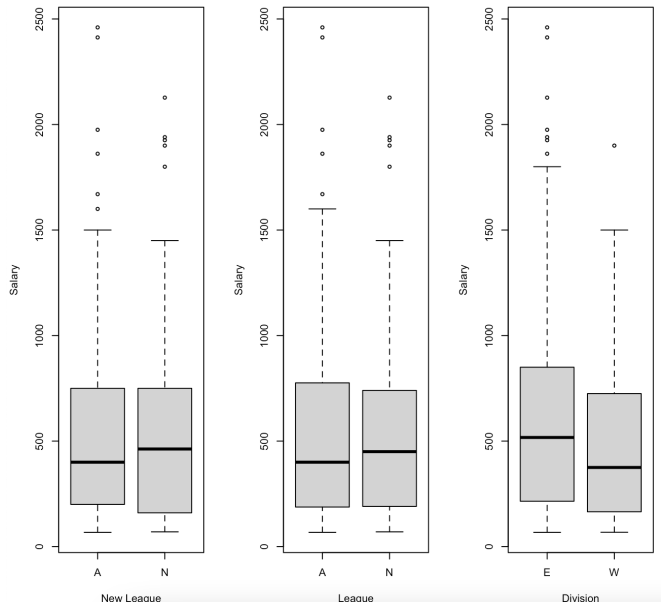

Fig. 1: Boxplots showing the players' salaries vs. a) new leagues in 1987 (Left), b) current leagues in 1986 (Center), and c) division in 1986 (Right).

Figure 2 contains visual depictions of the relationships between the predictor variables and the salary. From these figures, it is clear that the full model is not optimal. In the top left plot, there is clustering of the data points towards the lower salary values, which may indicate transformations (specifically, a log transformation of the salary) are necessary. We also have clear non-normality present in the Quantile-Quantile plot in the bottom right graph.

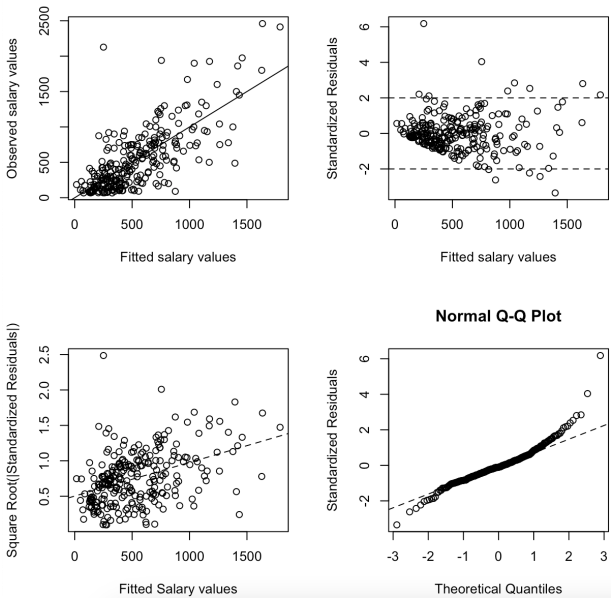

Fig. 2: Top left: Observed vs. Fitted values for the full model; Top right: Standardized residuals for the fitted values; Bottom left: Square Root of standardized residuals for fitted values; Bottom Right: Normal Quantile-Quantile Plot for the Residual Values

Appendix A contains pairwise plots of each quantitative variable. From these plots, we see some obvious multicollinearity and correlation between the predictor variables. Because most baseball statistics are reported as averages or ratios, we created scaled variables from the raw data to use in our analysis; the benefits are two-fold. First, we obtain a model containing predictor variables that are more widely reported than the raw values. Second, instead of trying to construct a model from a set of variables with obvious dependencies (e.g., since the number of hits obtained and the number of times at bat are inherently highly correlated), we constructed a model from a set of variables with some of the dependencies accounted for (e.g., by dividing the number of times at bat by the hits earned, which would give the player's batting average). This is the premise under which the rest of the analysis was conducted. Figure 3 contains the pairwise plots of the 11 scaled variables; the names and descriptions of these 11 variables can be found in Appendix A (Table A.1).

Model Selection

We performed model selection via all possible subsets on our 11 scaled variables and the three indicator variables for player division, league in 1986, and league in 1987. After diagsnostic analysis and experimentation, it was clear that a log transformation should be applied to the salary response variable (see *Model Diagnostics*). Table 1 contains the criteria scores for the subsets of variables (up to size 10) obtained through all possible subsets selection. [Note that this model selection was performed with approximately half of the data points; the other half are used in model validation.] All subsets with greater than 10 variables had worse criteria scores than the scores for 10 variables.

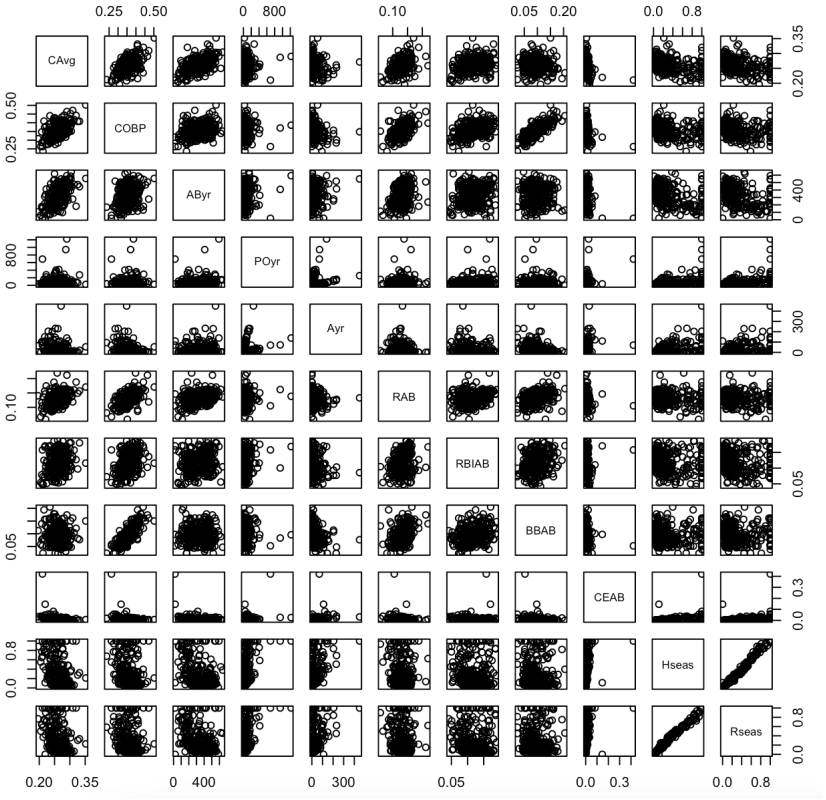

Fig. 3: Pairwise plots of the 11 auxiliary variables.

Table 1.

size	R^2_{adj}	AIC	AIC_c	BIC
1	0.557	-158.77	-156.56	-98.05
2	0.636	-180.51	-178.30	-119.78
3	0.708	-205.44	-203.23	-144.71
4	0.738	-215.94	-213.73	-155.21
5	0.751	**-218.83**	**-216.62**	**-158.1**
6	0.754	-216.35	-214.14	-155.62
7	0.755	-213.35	-211.14	-152.62
8	**0.755**	-209.63	-207.42	-148.9
9	0.754	-205.34	-203.13	-144.61
10	0.755	-201.74	-199.53	-141.01

Based on the output above, we chose to use five variables in our model since three of the criteria chose five variables as optimal, and the adjusted R^2 value for the five variable model is still quite high (0.751). These five variables are **CAvg** (career batting average), **AByr** (times at bat in the career per year played), **CEAB** (errors per times at bat in the career), **RBIAB** (career runs batted in per at bat), and **Rseas** (proportion of career runs scored that occurred in the 1986 season). Thus, the chosen

model is

$$\log(\text{Salary}) = 3.44 + 7.76(\text{CAvg}) + 7.27(\text{CEAB}) + 0.0016(\text{AByr}) + 4.23(\text{RBIAB}) - 2.025(\text{Rseas}) + \epsilon. \quad (1)$$

Model Diagnostics

We performed model diagnostics on the model in the previous section. First, it is necessary to note that following the results of a Box-Cox transformation of the variables, we used a log transformation on the Salary response variable (incorporated into model above). A plot of the fitted values of the salary in the model prior to transformation is available in Appendix A (Figure 8). Diagnostic plots for $\log(\text{Salary})$ and the five predictor variables are in Figure 4. The marginal model plot and added variable plots for the model are in Figure 5. Table 2 contains the variance inflation factors.

From Figure 4, we can see there is likely some non-normality still present. However, the observed versus fitted values in the top left plot is relatively accurate. Additionally, there is only one value outside of the (-3,3) interval, so there are not many points we visually classify as outliers. In Figure 5, we see the marginal model plot is quite accurate which indicates the model fits the data exceptionally well. The variance inflation factors in Table 2 show no serious multicollinearity as all are below 2 (and thus, below our cutoff of 5).

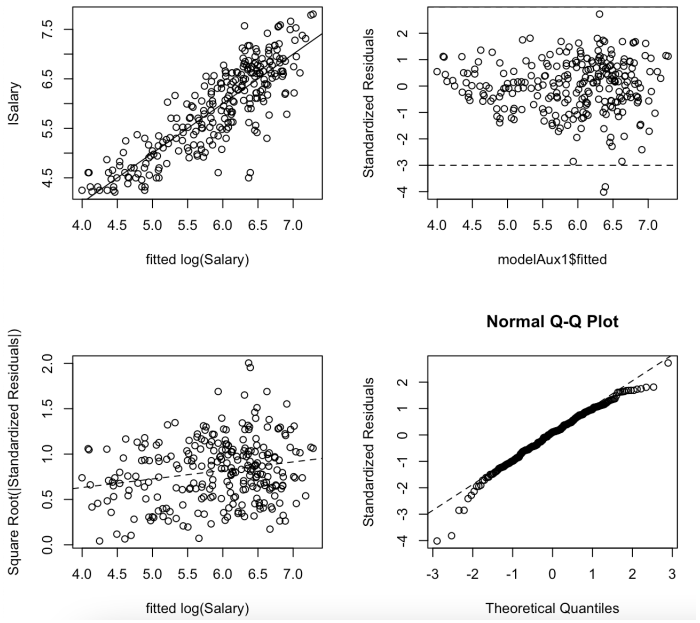

Fig. 4: Top left: Observed vs. Fitted values for the chosen model; Top right: Standardized residuals for the fitted values; Bottom left: Square Root of standardized residuals for fitted values; Bottom Right: Normal Quantile-Quantile Plot for the Residual Values

Table 2: Variance Inflation Factors

Variable	CAvg	CEAB	AByr	RBIAB	Rseas
VIF	1.6003	1.2259	1.5869	1.0962	1.2810

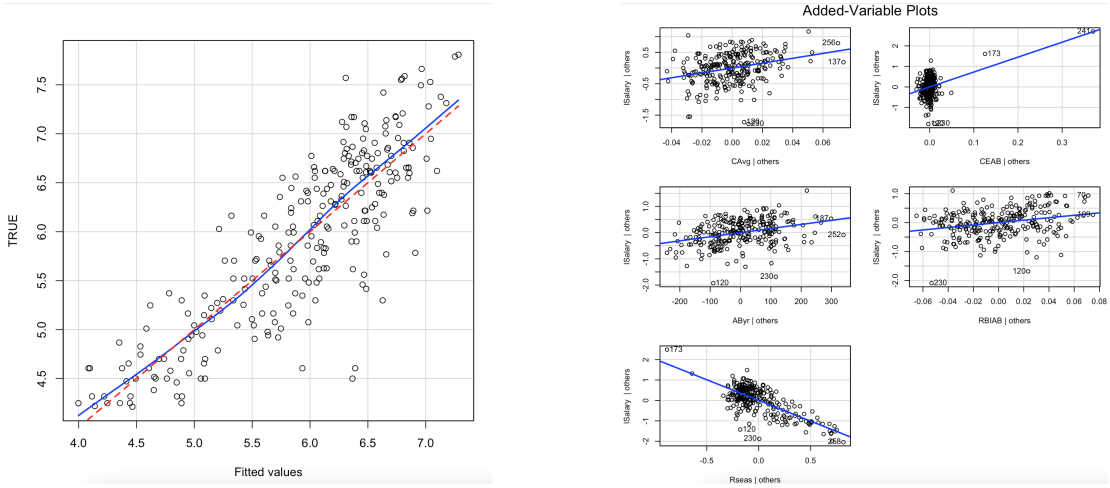

Fig. 5: (Left) Marginal model plot for the chosen model; (Right) Added variable plots for the chosen model

Model Validation

To validate the model, we predicted the values of the log transformation of the players' salaries using the coefficients for the model in (1) and the data from our test set (approximately half of the data). Figure 6 contains the plot of predicted values against observed values using the test data, which shows good adherence to the $y = x$ line.

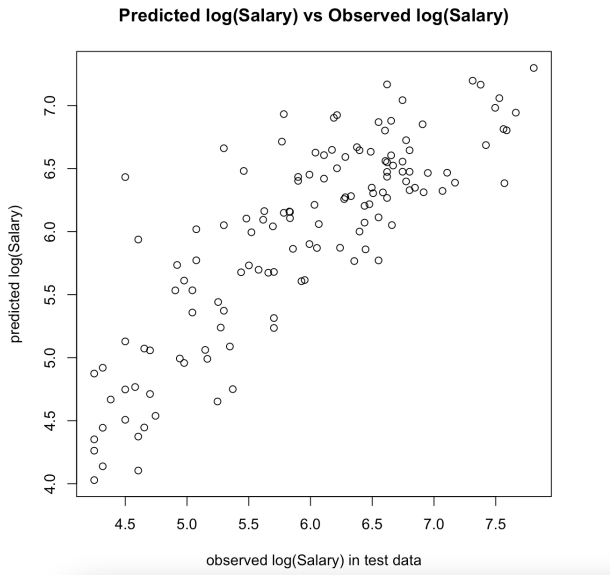

Fig. 6: Predicted vs. Observed values of log(Salary) using test data

Additionally, we calculated the $MSPR$ of the model in (1) to be $MSPR = 0.2551985$, and we calculated the MSE of the model (using training data) to be $MSE = 0.1792994$. Since the MSE of the model using training data and the $MSPR$ using test data are both small and somewhat similar, we can be relatively confident that our model has good predictive ability.

3. Results

We answer RQ1 and RQ2 in the same section as they are quite related. RQ3 is answered separately and continued in Appendix B with further details.

RQ1: Which variable has the largest estimated effect on the salary? Is this also the most statistically significant?

RQ2: What variables are most important in predicting salaries?

When evaluating these questions, we looked at the estimated regression coefficients since those are evaluated by holding all other variables constant to determine the relationship between the predictor and the dependent variable. To determine if a variable is the most statistically significant, we compare the p-values of the predictors and the lowest one is the most statistically significant.

When looking at all of the variables in respect to log(Salary), the largest estimated coefficient of 8.162 was associated with CEAB, or errors per career at bats. This was found to be statistically significant, and was included in the final model, but within the final model, the most influential predictor switched.

The variable with the largest estimated effect on log(Salary) in the final model is the career batting average random variable, with an estimated coefficient of 7.76, with p-value of $p = 1.01 \times 10^{-6}$, which is highly significant. However, the batting average is not the most significant variable in the model. **Rseas** has the smallest p-value, with $p < 2 \times 10^{-16}$.

It is reasonable that career batting average is very important when considering a player's worth because the batting average is a well-known and important statistic, and it is a common method in comparing players' values. Our model estimates that if a player increases his batting average from 0.200 (which is considered poor) to 0.400 (which is considered excellent), he can expect a 1.552% increase in his salary on average. [a]

RQ3: Are career statistics or previous year statistics more important in determining a baseball player's salary?

In order to determine whether career statistics or previous year statistics are more important in predicting a baseball player's salary, we consider our model discussed in the methods section. One thing that is immediately apparent is that 3 of the 5 predictor variables are career statistics rather than statistics from the previous season. This jumps out as an immediate pointer that career statistics are more predictive than previous season statistics.

Another way that we see the importance of career statistics over previous year statistics is through examination of the size of the coefficients comparatively. As we can see from Figure 7, the career based statistics are the 1st, 2nd and 5th largest coefficients in magnitude, indicating the strength and weight of their predictive power; that is, Figure 7 gives a visual indication to the strength and variability of the coefficients in the model. What is readily apparent is though three of the career statistics (CAvg, CEAB, AByr) all have larger magnitude than our predictor from the previous season (Rseas, RBIAB), we are also less confident (i.e., they have a wider confidence interval) about the exact effect of these statistics on a player's salary.

Another reason we feel this model is valid comes from consideration of the structure of MLB

[a] This is due to the fact that a 1 unit in career batting average would correspond to a 7.76% increase in salary. However, batting average is reported as a decimal between 0 and 1 inclusive.

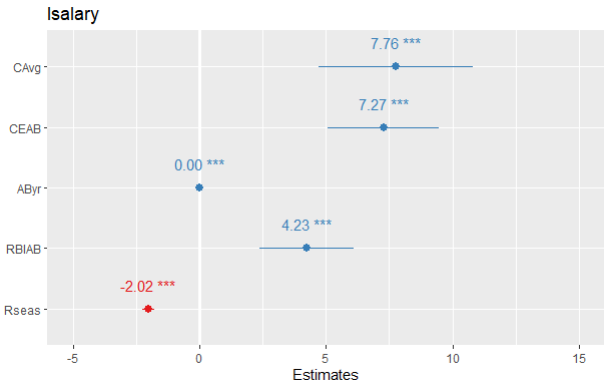

Fig. 7: Confidence intervals of coefficients in model.

salaries. As players are signed to multiple year contracts in a majority of cases, it is only reasonable that the statistics at the time of the contract signing are more predictive than those from the previous season. This is better reflected in the player's career statistics, which include their statistics from the time at which they signed their contract.

Lastly, we also considered an alternative method for answering this question, which we did by finding the best model based on previous year statistics and comparing it to the best model based on career statistics. We found that the career model performed better, providing another piece of evidence that career statistics have more predictive power for salaries. Our methodology in modeling is detailed further in Appendix B.

4. Discussion and Future Work

Further work would be valuable to the game of baseball and to the home offices of each team in the league. This is because having a tool that uses statistical modeling to model salaries can be helpful when managing a team's salary cap. If such a model could be refined, it could be used to optimize a team's ability to allocate more money for better players and less for worse players. This could potentially lead even to a better team as a whole, because if more cap space results by this, then the team has more money to potentially go get another star player from another team.

This is all optimistic conjecture at this point with the model we have, in reality. However, it is based on theoretically valid arguments; thus, further steps should be taken to refine the model and improve it. One of the things with most impact that can be done is to gather more data, and specifically, more data from recent years. Due to inflation and possible changes in how baseball is strategically played, there could be relatively major changes to our model if we wanted to use it for current salary predictions. Other statistics could be also be gathered for our model such as Slugging Percentage, Earned Run Average (or, ERA, for pitchers), and number of games played. These could yield a model that is a better predictor for salaries.

One other thing that could be improved in this model is analysis and removal of outlier data. We do not know if any of the players that present as outliers should be removed from this analysis for some reason (e.g., a pitcher who is brilliant at pitching and a valuable asset to teams but a very terrible hitter). The skewness of our final model (which can be seen in Figure 4) could be helped by removing these players that should not be considered in this analysis. This goes hand-in-hand with

the previous paragraph, as pitching data is absent from this data set. Incorporating more variables would improve our model.

Overall, there is further work in this area of baseball predictive statistics, but the work done so far is a good start to help aid further explorations.

References

(1) Carnegie Mellon University StatLib Library (2005). *Baseball*. Retrieved from Hitters data set (Version 1) by Mehmet A. at `https://www.kaggle.com/mathchi/hitters-baseball-data`.

Appendix A

Table A.1

Variable Name	Description
CAvg	Career batting average; career hits divided by career at bats.
COBP	Career on-base percentage; (career hits + career walks) divided by career at bats.
AByr	Average times at bat per year; career at bats divided by years in the MLB
POyr	Put Outs per year in the MLB
Ayr	Assists per year in the MLB
RAB	Career runs per at bat
RBIAB	Career runs batted in per at bat
BBAB	Career walks per at bat
CEAB	Errors per career at bat
Hseas	Proportion of career hits in the 1986 season; hits in 1986 season divided by career hits
Rseas	Proportion of career runs in the 1986 season; runs in 1986 season divided by career runs

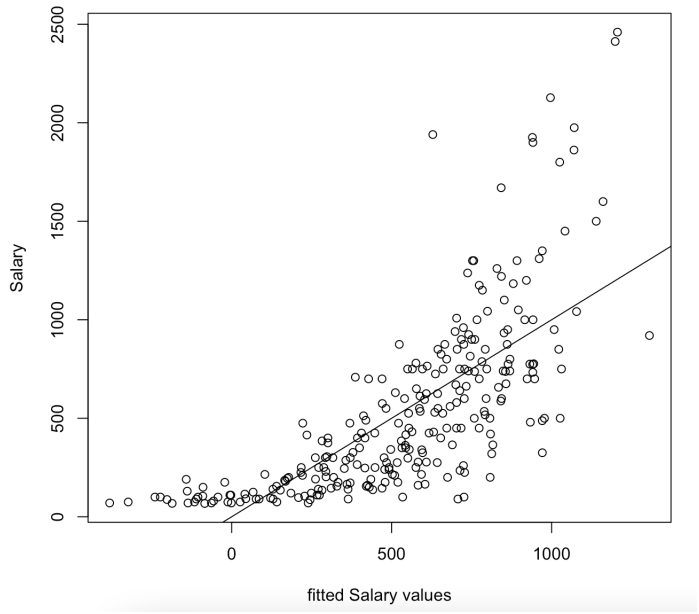

Fig. 8: Fitted values of Salary from the model in (1) prior to incorporating a log transformation of the salary.

Appendix B: More Modeling for Research Question 3

RQ3: Are career statistics or previous year statistics more important in determining a baseball player's salary?

In order to further test our research question, we will consider the best available model using only previous year statistics and the best available model using career statistics.

Previous Year Model: To find the best available model based on previous year statistics, we first began by considering all the available statistics in a linear model. The model was as follows:

$$\begin{aligned}\text{Salary} = \beta_0 &+ \beta_1(\text{AtBat}) + \beta_2(\text{Hits}) + \beta_3(\text{Runs}) \\ &+ \beta_4(\text{HmRuns}) + \beta_5(\text{RBI}) + \beta_6(\text{Walks}) + \beta_7(\text{League}) \\ &+ \beta_8(\text{Division}) + \beta_9(\text{NewLeague}) + \beta_{10}(\text{Avg}) + \beta_{11}(\text{OBP}) + \beta_{12}(\text{HRperAB}) \\ &+ \beta_{13}(\text{RAB}) + \beta_{14}(\text{RBIAB}) + \beta_{15}(\text{BBAB}) + \epsilon. \end{aligned} \tag{2}$$

Upon analysis of the results, we find very high multicollinearities present. This is to be expected as many of the stats measure similar things and are expected to be highly correlated. As such, we chose to remove several of the variables.

$$\begin{aligned}\text{Salary} = \beta_0 &+ \beta_1(\text{AtBat}) + \beta_2(\text{League}) + \beta_3(\text{Division}) + \beta_4(\text{NewLeague}) + \beta_5(\text{Avg}) + \beta_6(\text{OBP}) \\ &+ \beta_7(\text{HRperAB}) + \beta_8(\text{RAB}) + \beta_9(\text{RBIAB}) + \epsilon. \end{aligned} \tag{3}$$

After the removal of the highly collinear variables, we considered the best model through variable selection processes. According to adjusted R^2, AIC, AICc, and BIC, we select the variables AtBat, Division, and OBP for the model. With the use of the Box Cox method we obtain the following model:

$$Salary^{0.13} = 1.733 + .00062(\text{AtBat}) - .0593(Division^{0.16}) + 1.182(OBP^{1.67}) + \epsilon. \tag{4}$$

We see from the summary of the R output that each of the variables are statistically significant as well as is the model as a whole. The overall adjusted R^2 value is .2245, so the model is not very predictive overall. We will thus transition our attention to career stats and see if they are more predictive of player salaries.

Career Model: To find the best available model based on career statistics, we first began by considering all the available statistics in a linear model. The model was as follows:

$$\begin{aligned}\text{Salary} = \beta_0 &+ \beta_1(\text{CAtBat}) + \beta_2(\text{CHits}) + \beta_3(\text{Runs}) \\ &+ \beta_4(\text{CHmRuns}) + \beta_5(\text{PutOuts}) + \beta_6(\text{RBI}) + \beta_7(\text{Walks}) + \beta_8(\text{Assists}) \\ &+ \beta_9(\text{Errors}) + \beta_{10}(\text{Years}) + \beta_{11}(\text{CAvg}) + \beta_{12}(\text{COBP}) + \beta_{13}(\text{CHRperAB}) \\ &+ \beta_{14}(\text{CRAB}) + \beta_{15}(\text{CRBIAB}) + \beta_{16}(\text{CEAB}) + \beta_{17}(\text{Ayr}) + \beta_{18}(\text{AByr}) + \beta_{19}(\text{POyr}) + \epsilon. \end{aligned} \tag{5}$$

We then considered the best model through variable selection processes. According to adjusted R^2, AIC, AICc, and BIC, we select the 8 variables CAtBat, CRuns, CRBI, PutOuts, CAvg, AByr, POyr and CEAB for the model. With the use of the Box Cox method we obtain the following model:

$$Salary^{-0.04} = 1.04 + .103(CAtBat^{0.08}) - 0.174(CRuns^{0.08}) - 0.134(CRBI^{0.06}) - 0.008(PutOuts + 1)^{.2} + 0.007(CAvg^{-1}) + 0.002(AByr^{0.33}) - 0.007(CEAB^{-105.72}) + \epsilon. \quad (6)$$

The model has an adjusted R^2 of .59, greatly outperforming the previous year stats model.

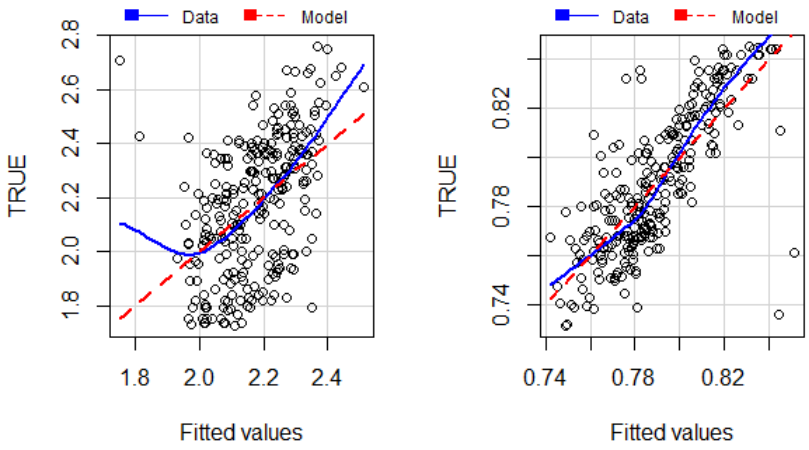

Fig. 9: MMP plots of PreviousYr Model on left and Career model on right

Based on MMP plots, we also see that the Career Stats model performs better in predicting salaries.

Based on our additional modeling that we have done, the conclusion that career statistics are more predictive than previous year statistics seems even more secure. We can thus confidently say that in order to improve your pay as a baseball player in the 1980's, it was significantly more important to have sustained success than short term success.