PREDICTING THE SALARIES OF MAJOR LEAGUE BASEBALL PLAYERS

Submitted by: Sangyun Kang

May 11, 2019

Predicting the salaries of major league baseball players

Abstract

This project analyses the salary of Major League Baseball (MLB) players and how much players are rewarded based on their performance information for the players' previous season and MLB careers. Each salary was examined on both the yearly performance and the overall career production of the player. Several different performance statistics were collected for the 1986-1987 MLB seasons. A random sample of players was selected from each season and separate models were created for position players. Significant performance attributes that were helpful in predicting salary were selected for each different model. These models were deemed to be good models having a predictive r-squared value of at least 0.85 for each of the different models. After the regression models were found, the models were tested for accuracy by predicting the salaries of a random sample of players from the 1986-1987 MLB season.

The goal is to predict the salary of MLB players. It is needed to explore the variables which may affect the salaries and suggests a progressive approach to predict the salaries. There are some missing values and outliers so it is needed to explore the data and handle these kinds of before building the prediction models. Minimizing and selecting the optimal features are used to maximize the prediction accuracy. And, normality test and multicollinearity are performed to in order to fit the model properly. Because the target is the continuous variable, linear regression is implemented to predict the dependent variable. The feature selection and correlation analysis between the two variables are performed to uncover the most important variables. Choosing the optimum features are used to maximize the prediction accuracy.

TABLE OF CONTENTS

1. Introduction

There is a hard salary cap in Major League Baseball (MLB). So, MLB teams are allowed to spend as much as they prefer on salaries. However, MLB tries to discourage overspending by the implementation of the Luxury Tax. Luxury Tax means that teams with a salary above the limitation have to pay a fine. Because of Luxury Tax and managerial reasons, MLB teams try to manage the players' payroll efficiently. In this project, our objective is to increase the prediction accuracy by finding the main factors that predict the salary of MLB players. The baseball dataset is used to create a regression model to predict the value of a baseball player's salary. Our first step before building our model, we viewed all the variables given to us and determined what the different variables mean and we check for any missing values or other ways to clean the dataset. Data preparations, namely Exploratory Data Analysis include handling the missing data, imbalanced data, outliers, selecting the important features. Also, the multicollinearity has to be detected and handled properly. After preprocessing the data, we build models based on our dataset and validate their predictive performance by using R-squared and RMSE.

2. The Objective of Analysis

The goal of this study is to examine the significant factors that determine the average yearly salary of the MLB player. Different statistics need to be evaluated for hitters. In this study, there are many different statistics examined for production instead of only using only one production statistic. By predicting possible salaries, MLB teams will be better able to manage its players by providing them with reasonable salaries and manage a roster on a limited budget. The following are the main goals of this project.

1. Visualize the data to understand the categories of each attribute and their influence on the dependent variable.
2. Data preprocessing (Handling the missing values and outliers in the dataset, feature selection, etc.).
3. Detect the multicollinearity by using the correlation matrix, variable inflation factor (V.I.F), conditional indices and resolve this issue through variable elimination, model re-specification, centering of regressors, Ridge Regression, Principal Component Regression.
4. Build the regression models (Forward, Backward, Stepwise, LASSO, etc.) and find the best model.
5. Compare and evaluate the R-squared and RMSE for all models.
6. Explaining the influence of each independent variable for the target variable.

3. Understanding the Data (Data descriptive)

Our data are related with 322 Major League Baseball (MLB) players who played at least one game in both the 1986 and 1987 seasons. The attributes are for only hitters, not pitchers. The salaries are for the 1987 season and the performance attributes are from 1986. There are 19 input variables and a target variable (Salary). It is needed to predict the salary. The main purpose of this step is data understanding. The data is uploaded to the SAS by using the DATA LOADING method. Then, it is performed to demonstrate if there will be null values and white space cells. To understand the unique category and its weight within each feature and how it would affect the output (the target), data visualization is performed. Surprisingly, it founds out that salary variable has 59 missing values.

The first five records of the data are represented in table 1. Also, some of the input variables are described in table 2.

The first five observations out of 322

Obs	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary	NewLeague
1	66	1	30	29	14	1	293	66	1	30	29	14	A	E	446	33	20	.	A
2	81	7	24	38	39	14	3449	835	69	321	414	375	N	W	632	43	10	475.0	N
3	130	18	66	72	76	3	1624	457	63	224	266	263	A	W	880	82	14	480.0	A
4	141	20	65	78	37	11	5628	1575	225	828	838	354	N	E	200	11	3	500.0	N
5	87	10	39	42	30	2	396	101	12	48	46	33	N	E	805	40	4	91.5	N

Table 1: The representation of the dataset in SAS

Variable	Type	Description
1.Hits	num	Number of hits in 1986
2.HmRun	num	Number of home runs in 1986
3.Runs	num	Number of runs in 1986
4.RBI	num	Number of runs batted in in 1986
5.Walks	num	Number of walks in 1986
6.Years	num	Number of years in the major leagues
7.CAtBat	num	Number of times at bat during his career
8.Chits	num	Number of hits during his career
9.CHmRun	num	Number of home runs during his career
10.CRuns	num	Number of runs during his career
11.CRBI	num	Number of runs batted in during his career
12.CWalks	num	Number of walks during his career
13.League	char	"A" or "N" indicating player's league at the end of 1986
14.Division	char	"E" and "W" indicating player's division at the end of 1986
15.PutOuts	num	Number of put outs in 1986
16.Assists	num	Number of assists in 1986
17.Errors	num	Number of errors in 1986
18.Salary	num	1987 annual salary on opening day in thousands of dollars
19.NewLeague	char	"A" and "N" indicating player's league at the beginning of 1987

Table 2: List of attributes of dataset

4. Exploratory Data Analysis

4.1 Dealing with missing data and imbalance data

The presence of a lot of missing data affects the predictive behavior of a model. It can be found that there are 59 observations in the first output table with at least one of the input variables with missing values; These missing values are not used as building the linear regression model. We analyzed the number of missing values under each variable category from the following table:

Number of Observations Read	322
Number of Observations Used	263
Number of Observations with Missing Values	59

Missing Values

Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	59	18.32	100.00

Table 3: The representation of missing data

Salary	Frequency
Missing	59
Not Missing	263

Table 4: Missing data in Salary

There are two approaches to solve the missing dataset. The first solution is omission (removing data). If many records are missing values on a small set of variables, we can drop those variables. The second one is imputation. We can replace missing values with reasonable substitutes. (Mean, median or mode). In this project, we omit the missing values. SAS program automatically delete the row which has missing values.

*** Imputation using median values**

This works by calculating the median of the non-missing values in salary variable and then replacing the missing values into the median. It works well with small numerical datasets. But, it is not very accurate because it decreases the R-squared.

```
PROC STDIZE DATA=mlb_hit OUT=Imputed
  oprefix=Orig_          /* prefix for original variables */
  REPONLY                /* only replace; do not standardize */
  METHOD=MEDIAN;         /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
  VAR salary;            /* you can list multiple variables to impute */
RUN;
```

Figure 1: Imputation method in SAS

First, we can check the distribution of target variable(salary) by using PROC UNIVARIATE function.

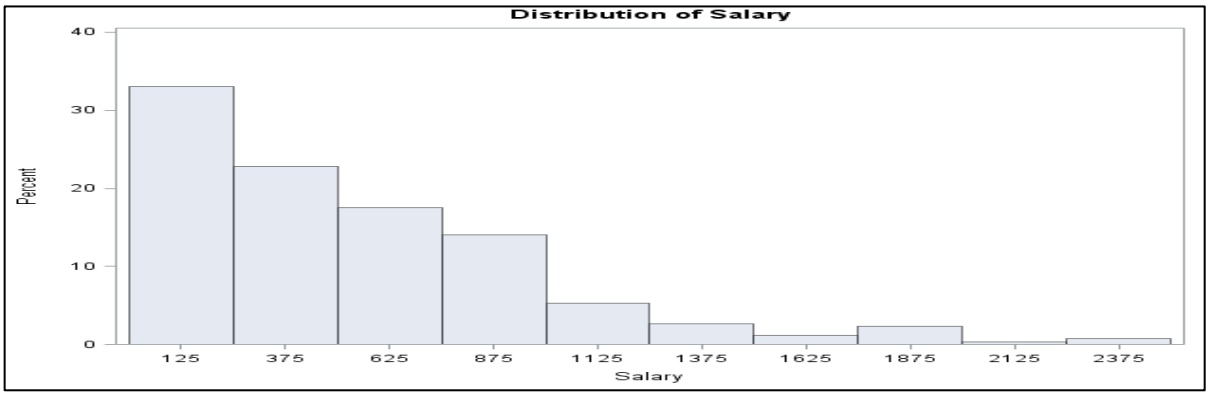


Figure 2: Distribution of Salary

4.2 Variable transformation

The variable transformation, which make response to be normal, can improve the model fitting. The imbalance of the data is not easy to understand and it is required more attention to all input variables within training set and test set. An additional approach is performed on the target to handle the imbalanced dataset. The majority of the points within target distributed around zero, therefore keeping the original data can be biased toward one class. So, we need to handle this condition first. The distribution transformation method can be utilized here. The implementation of this technique is required either x or y axis to improve the model fitting. Figure 3 on the left shows the distribution of the original values of the response. One can see that the distribution is positively skewed. To get a good prediction, it prefers the distribution to be equally distributed over the given values. So, in order to handle this skewed data, the log transformation can be performed as depicted in figure 3 on the right.

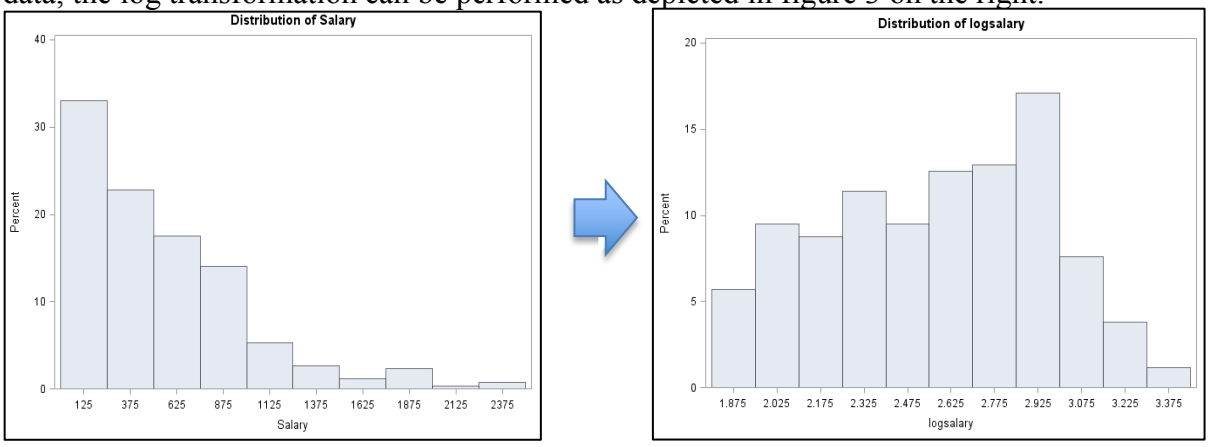


Figure 3: The distribution of the response variable and log-transformation of the data

In this project, it is better to predict the log salary value of a player instead of salary. Because the variation of salaries is much greater as the salaries are higher, it is proper to need a log transformation to the salaries before building the regression models. In order to adjust for the different variances in salary for a given set of production statistics and to make a better predictive model, the log of salary will be used for the dependent variable in all of the models.

- **Fit Full Model (include all the regressors)**

Dependent Variable: Salary

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	28147125.61	1563729.20	15.16	<.0001
Error	244	25171987.17	103163.88		
Corrected Total	262	53319112.79			

R-Square	Coeff Var	Root MSE	Salary Mean
0.527899	59.93205	321.1913	535.9259

Dependent Variable: logsalary

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	21.40768122	1.18931562	16.43	<.0001
Error	244	17.66393592	0.07239318		
Corrected Total	262	39.07161715			

R-Square	Coeff Var	Root MSE	logsalary Mean
0.547909	10.45234	0.269060	2.574160

Table 5: Salary output from PROC GLM / Table 6: logsalary output from PROC GLM

Before model selections, check the model including all the regressors from original dataset and model including all variables after log transformation. From table 5 on the left, this is first full model which do not transform anything and just use the original dataset. From table 5 on the right, this is full model after the log transformation. We can check that the prediction accuracy gets better (R-square: 0.527899 -> 0.547909, RMSE: 321.1913 -> 0.269060).

4.3 Detecting and handling the outliers

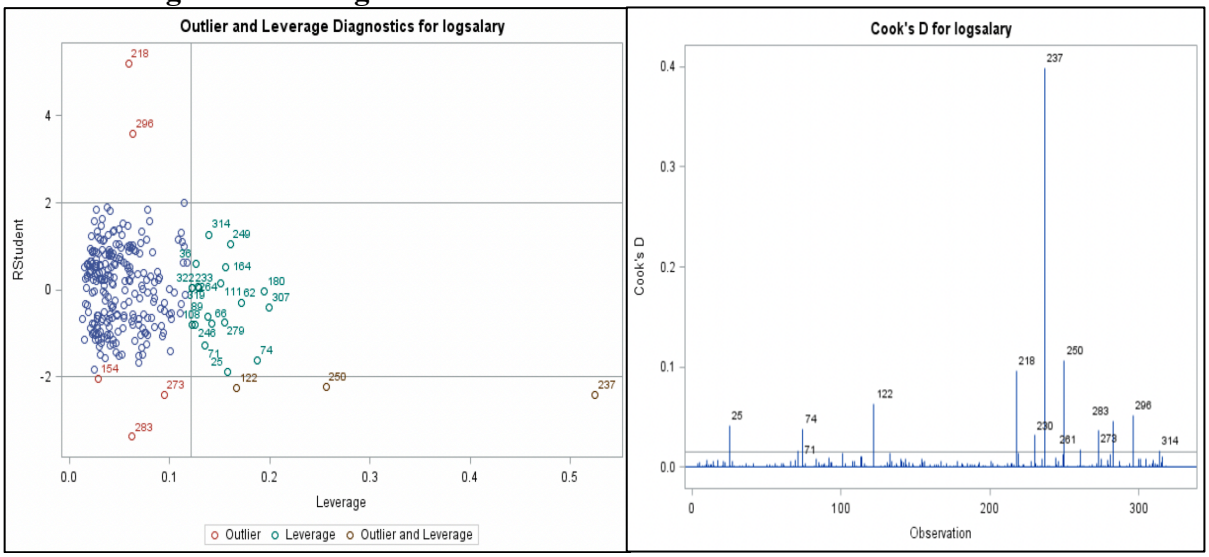


Figure 4: Outlier and Leverage Diagnostics and Cook's D for logsalary

When we see the figure 4, we can check that there are four to six individuals whose input variables may have excessive influence on fitting this model. Those points are needed to delete and if those are handled properly, the model can be improved. When player 237, 218, 250, and 296 are removed, R-squared is increased and RMSE is decreased.

Root MSE	0.24799	R-Square	0.6092
Dependent Mean	2.56814	Adj R-Sq	0.5851
Coeff Var	9.65637		

Table 7: Logsalary output from PROC GLM

4.4 Handling variables

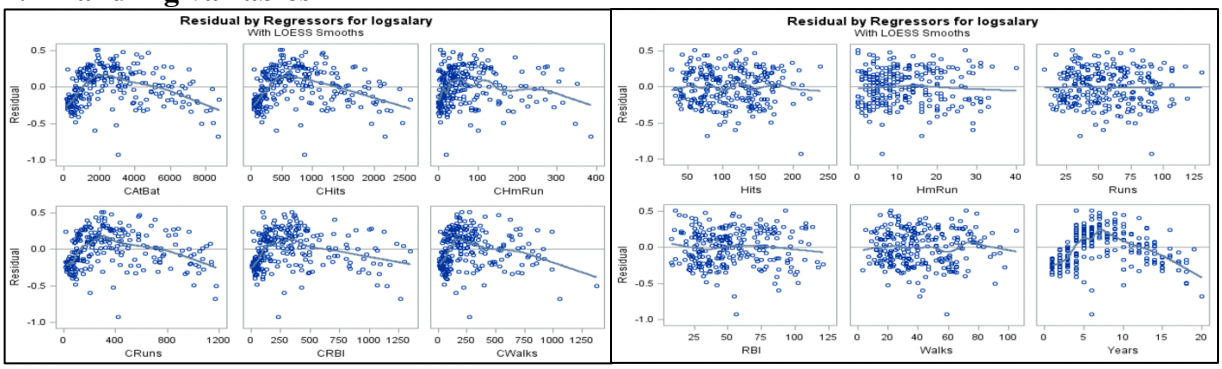


Figure 5: Residuals by Regressors

From figure 5, we can check that a loess fit is overlaid on each of these plots. There is the same clear pattern in the residual plots for Years, CAtBat, Chits, CHmRun, CRuns, CRBI, and CWalks. Players near the start of their careers and players near the end of their careers get paid less than the model predicts.

In order to address this lack of fit, it is needed to use polynomials of degree two for these variables, Years, CAtBat, Chits, CHmRun, CRuns, CRBI, and CWalks.

After removing some outliers and adjusting for the overlaid variable, the improved R-square value of 0.8147 can be checked.

Model: MODEL1
Dependent Variable: logsalary

Number of Observations Read	318
Number of Observations Used	259
Number of Observations with Missing Values	59

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	31.15369	1.41608	47.17	<.0001
Error	236	7.08522	0.03002		
Corrected Total	258	38.23891			

Root MSE	0.17327	R-Square	0.8147
Dependent Mean	2.56814	Adj R-Sq	0.7974
Coeff Var	6.74685		

Table 8: PROC REG output

4.5 Clinically significant regressors

In this part, we can create additional variables from the data by using the domain knowledge in baseball. Investigation on this site (http://en.wikipedia.org/wiki/Baseball_statistics) indicated other input variables that could be created from the original data. From this site, the following variables were created using a SAS Code because they were considered significant for predicting the baseball player's salary:

BA (Batting average) - CHits / CAtBat

HR/H (Home runs per hit) – HmRun / Hits

The R-square is increased into 0.822607, which means that adding new variables are helpful.

* If we have other variables, we can use the other important variables such as OPS (On-base plus slugging) – on-base percentage plus slugging average, SLG (slugging average) – TB (Total bases) / AB (At Bat), injury(health), etc.

4.6 Data Normalization

Different values in the dataset have a variety of ranges. Some variables range from 2~23 while other attributes range from 0~8474. Therefore, we normalize the data by using auto data preparation. In fact, most of the models request the data scaling. The input variables should be logic in terms of units and scaling.

*** Dropping the insignificant variables**

Too many variables can make the model overfitting. Overfitting can happen when data that over-presents the target event, and therefore it will produce a poor performing model when using real-world data as input.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CBA	1	11.18088759	11.18088759	380.76	<.0001
HR_H	1	0.85359240	0.85359240	29.07	<.0001
Years2	1	2.85950277	2.85950277	97.38	<.0001
CAtBat2	1	0.61306090	0.61306090	20.88	<.0001
CHits2	1	0.02270623	0.02270623	0.77	0.3801
CHmRun2	1	0.13553288	0.13553288	4.62	0.0327
CRuns2	1	0.42491918	0.42491918	14.47	0.0002
CRBI2	1	0.09643677	0.09643677	3.28	0.0713
CWalks2	1	0.03593579	0.03593579	1.22	0.2698
Hits	1	2.55748652	2.55748652	87.09	<.0001
HmRun	1	0.00052197	0.00052197	0.02	0.8941
Runs	1	0.03835304	0.03835304	1.31	0.2543
RBI	1	0.13111768	0.13111768	4.47	0.0357
Walks	1	1.25065928	1.25065928	42.59	<.0001
Years	1	9.77245543	9.77245543	332.79	<.0001
CAtBat	1	1.07632165	1.07632165	36.65	<.0001
CHits	1	0.03585804	0.03585804	1.22	0.2703
CHmRun	1	0.00049382	0.00049382	0.02	0.8969
CRuns	1	0.05664710	0.05664710	1.93	0.1662
CRBI	1	0.01161371	0.01161371	0.40	0.5300
CWalks	1	0.00333782	0.00333782	0.11	0.7363
League	1	0.04475552	0.04475552	1.52	0.2183
Division	1	0.07180192	0.07180192	2.45	0.1193
PutOuts	1	0.17640446	0.17640446	6.01	0.0150
Assists	1	0.00324059	0.00324059	0.11	0.7400
Errors	1	0.00128371	0.00128371	0.04	0.8346
NewLeague	1	0.00068331	0.00068331	0.02	0.8789

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CBA	1	11.18088759	11.18088759	385.29	<.0001
HR_H	1	0.85359240	0.85359240	29.41	<.0001
Years2	1	2.85950277	2.85950277	98.54	<.0001
CAtBat2	1	0.61306090	0.61306090	21.13	<.0001
CHits2	1	0.02270623	0.02270623	0.78	0.3773
CHmRun2	1	0.13553288	0.13553288	4.67	0.0317
CRuns2	1	0.42491918	0.42491918	14.64	0.0002
CRBI2	1	0.09643677	0.09643677	3.32	0.0696
CWalks2	1	0.03593579	0.03593579	1.24	0.2669
Hits	1	2.55748652	2.55748652	88.13	<.0001
Runs	1	0.03877821	0.03877821	1.34	0.2488
RBI	1	0.09035034	0.09035034	3.11	0.0789
Walks	1	1.29026144	1.29026144	44.46	<.0001
Years	1	9.66314337	9.66314337	332.99	<.0001
CAtBat	1	1.12642933	1.12642933	38.82	<.0001
CHits	1	0.03054233	0.03054233	1.05	0.3060
League	1	0.03957870	0.03957870	1.36	0.2440
Division	1	0.05767238	0.05767238	1.99	0.1599
PutOuts	1	0.18651022	0.18651022	6.43	0.0119

Table 9: The representation of p-value in each variable

Since several of the input values appear to have little predictive power on the target, we need to drop these variables, thereby reducing the need for that information to make a decent prediction. The p-value associated with each of the input variables provides the analyst with an insight into which variables have the biggest impact on helping to predict the target variable. In this case, the smaller the value, the higher the predictive value of the input variable. The dropping variables are: HmRun, CHmRun, CRBI, CWalks, Assists, Errors, and NewLeague. But, since it is not helpful for prediction accuracy, this method is not used.

5. Multicollinearity

From now, we finished data preprocessing but before building the regression model, multicollinearity is needed to resolve appropriately. If there are two highly correlated regressors in the models, they make the prediction accuracy worse. Here are methods for resolving the multicollinearity

- Detecting multicollinearity: CORR, VIF, Tol, and Collin
- Solving multicollinearity: (1) Drop a variable, (2) Use ridge regression (3) Utilize principal components regression.

First, we can check the correlation. In correlation matrix (Appendix), if any variable has a high correlation about 0.8 or higher with other variables. It has to be removed. If there are correlations between two attributes, it can cause problems when we fit the model. They are needed to delete the variables which are highly correlated with each other. Next, we can check multicollinearity through the Variance Inflation Factor and Tolerance:

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	515.60095	17.82022	28.93	<.0001	.	0
Hits	1	17.44726	73.28646	0.24	0.8120	0.05931	16.86050
HmRun	1	18.24271	51.20644	0.36	0.7219	0.12149	8.23137
Runs	1	-4.32583	73.05940	-0.06	0.9528	0.05968	16.75619
RBI	1	0.60911	64.13103	0.01	0.9924	0.07745	12.91099
Walks	1	79.25439	35.62066	2.22	0.0268	0.25106	3.98315
Years	1	-17.16948	54.29826	-0.32	0.7521	0.10805	9.25539
CAtBat	1	-812.13138	253.54977	-3.20	0.0015	0.00496	201.81314
CHits	1	583.79148	377.66042	1.55	0.1232	0.00223	447.74006
CHmRun	1	-44.76582	129.51237	-0.35	0.7298	0.01899	52.65571
CRuns	1	325.31785	218.96087	1.49	0.1384	0.00664	150.50677
CRBI	1	207.00893	218.05722	0.95	0.3432	0.00670	149.26705
CWalks	1	-86.01314	73.44418	-1.17	0.2425	0.05906	16.93315
PutOuts	1	54.93341	19.94962	2.75	0.0062	0.80040	1.24937
Assists	1	45.29064	28.86933	1.57	0.1177	0.38221	2.61635
Errors	1	-36.17392	26.52125	-1.36	0.1736	0.45289	2.20806

Table 10: Variance Inflation Factor and Tolerance

If tolerance fall below 0.1 and variance inflation is above the value of 10, it can be concluded that there is a threat of multicollinearity. And, we see the collinearity diagnostics for an eigensystem analysis of covariance comparison:

Collinearity Diagnostics											
			Proportion of Variation								
Number	Eigenvalue	Condition Index	Intercept	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun
1	7.08521	1.00000	0	0.00025710	0.00062668	0.00025517	0.00052128	0.00139	0.00144	0.00008499	0.00003819	0.00029113
2	3.55537	1.41167	0	0.00294	0.00309	0.00307	0.00308	0.00655	0.00179	0.00003613	0.00001426	0.00004219
3	1.64906	2.07280	0	0.00007250	0.00774	0.00009265	0.00114	0.00040226	0.00049123	0.00004240	0.00001815	0.00014586
4	1.00000	2.66181	1.00000	0	0	0	0	0	0	0	0	0
5	0.85182	2.88404	0	0.00081189	0.00227	0.00229	0.00075679	0.00044069	0.00011283	0.00000317	0.00000247	0.00000672
6	0.66536	3.26323	0	0.00217	0.04448	0.00428	0.00861	0.12777	0.00001164	0.00000358	0.00000417	0.00231
7	0.43364	4.04215	0	0.03084	0.01108	0.00610	0.00133	0.19081	0.00643	0.00029701	0.00024226	0.00432
8	0.25127	5.31018	0	0.00066633	0.00788	0.00440	0.00306	0.00247	0.00007271	0.00005607	0.00007889	0.00162
9	0.17606	6.34372	0	0.01732	0.10117	0.00689	0.01281	0.01905	0.14759	0.00024850	0.00004141	0.04167
10	0.12330	7.58058	0	0.00933	0.13301	0.07291	0.12354	0.17621	0.12106	0.00010904	0.00004120	0.00201
11	0.09362	8.69939	0	0.00036703	0.00989	0.10014	0.15376	0.00178	0.46646	0.00125	0.00172	0.01697
12	0.05932	10.92872	0	0.05957	0.09358	0.00363	0.09072	0.33746	0.02544	0.00447	0.00375	0.00094509
13	0.03506	14.21484	0	0.65814	0.37810	0.47655	0.41335	0.08367	0.00222	0.00008650	0.00028293	0.00040150
14	0.01426	22.28818	0	0.00616	0.03491	0.09716	0.07948	0.01999	0.05276	0.00616	0.00119	0.11185
15	0.00532	36.48600	0	0.01084	0.08215	0.01578	0.04500	0.00086175	0.15745	0.54261	0.00376	0.19853
16	0.00132	73.14623	0	0.20053	0.09004	0.20643	0.06284	0.03116	0.01668	0.44454	0.98881	0.61887

Table 11: Collinearity Diagnostics

If eigenvalues are getting small and the corresponding condition numbers become large, this is a clear indication of multicollinearity. Above the table, condition number is getting large as eigen value closes to zero so there is multicollinearity. So, we drop the variables: CAtBat2, Chits2, CABat, Chits, CRuns2, CRuns, CRBI2, CRBI, CWalks2, CWalks. In conclusion, multicollinearity has to be resolved. The easy way is to drop one of variables, which has correlation between two variables.

Because multicollinearity is solved after dropping the variables, ridge regression and PCR is not needed but we can also try these two methods. There are at least two alternative methods of resolving the issue of multicollinearity: ridge regression and principal component regression.

- **Ridge Regression**

Ridge regression is used when a multicollinearity is identified after standardizing regressors (centering and scaling) and dropping near-zero-coefficient. It is a variant to least squares regression. The traditional ordinary least squares (OLS) regression produces unbiased estimates for the regression coefficients, however, if there are the confounding issue of highly correlated explanatory variables, your resulting OLS parameter estimates end up with large variance. Thus, using ridge regression could be helpful to obtain a smaller variance in resulting parameter estimates.

$$\min\left(\left|\left|Y - X(\theta)\right|\right|_2^2 + \lambda\left|\left|\theta\right|\right|_2^2\right)$$

Equation 1: Ridge Regression

Logsalary - Ridge Regression Results

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	CBA	HR_H	Years2	CHmRun2	Hits	HmRun	Walks	Years	CHmRun
1	MODEL1	PARMS	logsalary	.	.	0.17949	2.54365	0.08070	-0.03624	-0.6728	-0.1002	0.03496	0.0137	0.05050	0.7259	0.2392
2	MODEL1	RIDGEVIF	logsalary	0.000	.	.	.	2.16359	7.86979	19.9395	15.4575	7.27742	12.7400	1.74444	21.8535	26.1388
3	MODEL1	RIDGE	logsalary	0.000	.	0.17949	2.54365	0.08070	-0.03624	-0.6728	-0.1002	0.03496	0.0137	0.05050	0.7259	0.2392
4	MODEL1	RIDGEVIF	logsalary	0.002	.	.	.	2.10458	7.25945	17.0213	12.9756	6.72580	11.4745	1.72193	18.1571	21.9045
5	MODEL1	RIDGE	logsalary	0.002	.	0.17980	2.54360	0.08031	-0.03480	-0.6252	-0.1099	0.03813	0.0109	0.05177	0.6802	0.2459
6	MODEL1	RIDGEVIF	logsalary	0.004	.	.	.	2.05338	6.72712	14.7763	11.1182	6.25297	10.4178	1.70072	15.3952	18.7099
7	MODEL1	RIDGE	logsalary	0.004	.	0.18057	2.54360	0.08014	-0.03363	-0.5841	-0.1162	0.04064	0.0090	0.05291	0.6413	0.2490
8	MODEL1	RIDGEVIF	logsalary	0.006	.	.	.	2.00796	6.25928	13.0002	9.6857	5.84238	9.5204	1.68063	13.2693	16.2275
9	MODEL1	RIDGE	logsalary	0.006	.	0.18163	2.54362	0.08012	-0.03265	-0.5481	-0.1201	0.04269	0.0079	0.05396	0.6078	0.2497

Table14: Ridge Regression Results

From table14, the appropriate ridge parameter can be taken through this analysis. In order to achieve this, as the ridge parameter of .002 increases the _RMSE_ only slightly from 0.17555 to 0.17655 and drops the VIF for each of our problem variables to below 10. Hence, we can choose the ridge parameter of 0.002. Finally, the multicollinearity issue controlled!

- **Principal Components Regression**

Another way to solve multicollinearity is through Principal Components Regression.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.08945851	1.55380385	0.3408	0.3408
2	2.53565465	0.78052448	0.2113	0.5521
3	1.75513018	0.58139359	0.1463	0.6984
4	1.17373658	0.36149288	0.0978	0.7962
5	0.81224371	0.22421401	0.0677	0.8639
6	0.58802969	0.15489421	0.0490	0.9129
7	0.43313548	0.16852918	0.0361	0.9489
8	0.26460630	0.02940601	0.0221	0.9710
9	0.23520029	0.18090476	0.0196	0.9906
10	0.05429553	0.01597233	0.0045	0.9951
11	0.03832320	0.01813734	0.0032	0.9983
12	0.02018586		0.0017	1.0000

Table 13: Eigenvalues of the correlation matrix

These methods are from the eigenvalues of the correlation matrix and the scree plot. Through checking these plots, it is needed to decide the number of factors to build our model. Any factor with an eigenvalue higher than 1.000 can remain in the model as it explains at least 1 variable's worth of information. Thus, our model includes 5 factors. And, from Appendix, we can check the output and the r-squared is 0.5636.

Ridge Regression and PCR is good for predicting the dependent variable but they can't explain the relationship between the dependent and independent variables.

If multicollinearity is not resolved properly, it can have a negative impact on the accuracy of your model. Through the above methods, it is important to detect and solve the issue of multicollinearity before fitting the regression model.

6. Residual Analysis (Model diagnostics)

Before log transform, after log transform, and after adding new variables, residual analysis is implemented to check the model adequacy. Now, this part is for the final model. It is important to examine influence and fit diagnostics to see whether the model might be unfittingly influenced by a few outliers and whether the data support the assumptions that underset the linear regression.

There are five steps:

(a) Detection of Outliers

(b.1) Detection of Heteroscedasticity

(b.2) Detection of Correlation

(b.3) Detection of serious violation against Normality

(c) Misspecification of the model

Except for step b.3, only four steps will be conducted because normality will be checked finally.

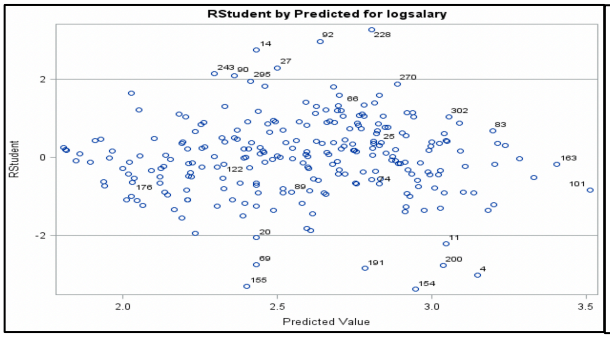


Figure 6: Rstudent versus fitted response

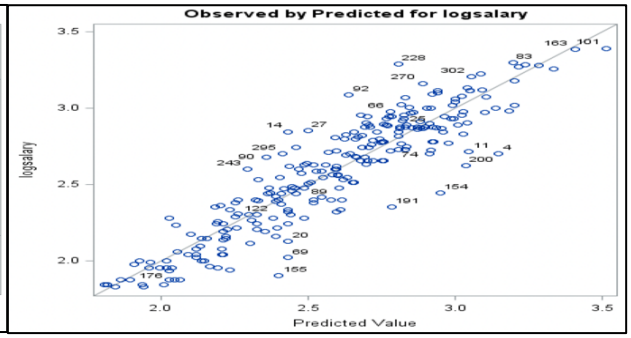


Figure 7: Observed by predicted for salary

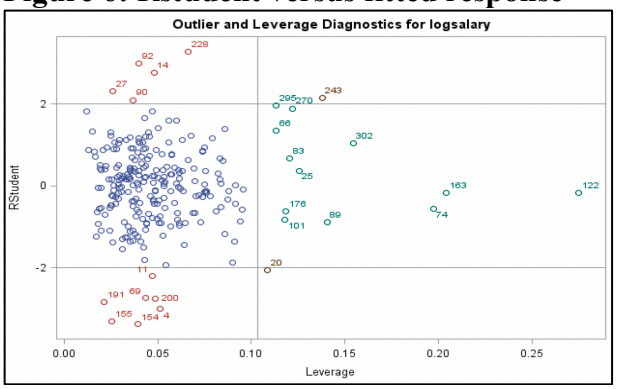


Figure 8: Rstudent residual-Leverage plot

So, we can see that Rstudent residual-leverage plot provides possible outlying information in both dimensions. As for (b.1), Detection of Heteroscedasticity, one can refers to the Rstudent residual-fitted response plot above. Excluding those possible outliers, there seems no serious violation against homoscedasticity.

As for (b.2), Detection of (auto-) Correlation, we will use Durbin-Watson test.

Durbin-Watson D	1.826
Pr < DW	0.0806
Pr > DW	0.9194
Number of Observations	251
1st Order Autocorrelation	0.086

Table 14: Durbin-Watson test

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

The note in the printout indicates how to interpret the result already. Remind yourself that the two p-values (Pr<DW and Pr>DW) refer to two different tests (H_0: no positive correlation; H_0: no negative correlation, respectively). In this example, there is neither significant positive correlation nor significant negative correlation.

As for (b.3), Detection of serious violation against Normality, one could use the histogram and QQ-plot first.

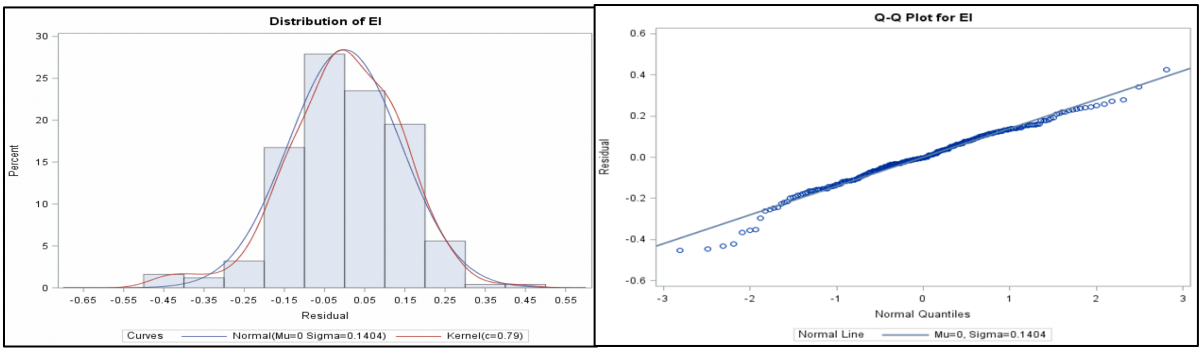


Figure 9: Distribution of Residuals / Figure 10: Q-Q plot of Residuals

We see that there is no concern about serious violation against normality. We will also perform several hypothesis tests to confirm this. A SAS macro has implemented, %NORMTEST(VAR, DATA), to facilitate this task. We will need to
(1) Define the macro in SAS
(2) Extract the targeted quantity whose normality is to be tested (Rstudent residual)
(3) Call the macro

Note that I have suppressed the printout and added an extra **OUTPUT** statement. **OUT**= specifies the name of dataset and **R**= specifies the name of raw residual; **RSTUDENT**= specifies the name of Rstudent residual residual; **STUDENT**= specifies the name of studentized residual; and **PRESS**= specifies the name of PRESS residual. (3) Call/Run the MACRO on EI in dataset fit.

%***NORMTEST***(EI, FIT)

We see the following testing results

NORMAL-TEST

Variable: EI (Residual)

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.983439	Pr < W	0.0051
Kolmogorov-Smirnov	D	0.0509	Pr > D	0.1118
Cramer-von Mises	W-Sq	0.089007	Pr > W-Sq	0.1607
Anderson-Darling	A-Sq	0.742761	Pr > A-Sq	0.0529

```
D'AGOSTINO TEST OF NORMALITY FOR VARIABLE EI, N=251
G1=-0.41671   SQRTB1=-0.41421   Z=-2.65576     P=0.0079
G2=0.82208    B2=3.78198        Z= 2.17285     P=0.0298
K**2=CHISQ(2 DF)=11.77434                      P=0.0028
```

Table 15: Normality test

Again, we see some p-values are larger than 0.05, H_0, data are normal, is hence rejected.

As for (c) misspecification of the model, we turn our attention to misspecification of the model. Before we discuss whether the currently deployed model works or not, one can look into partial regression plots (of regressors) to see their marginal potentials to contribute to explain response.

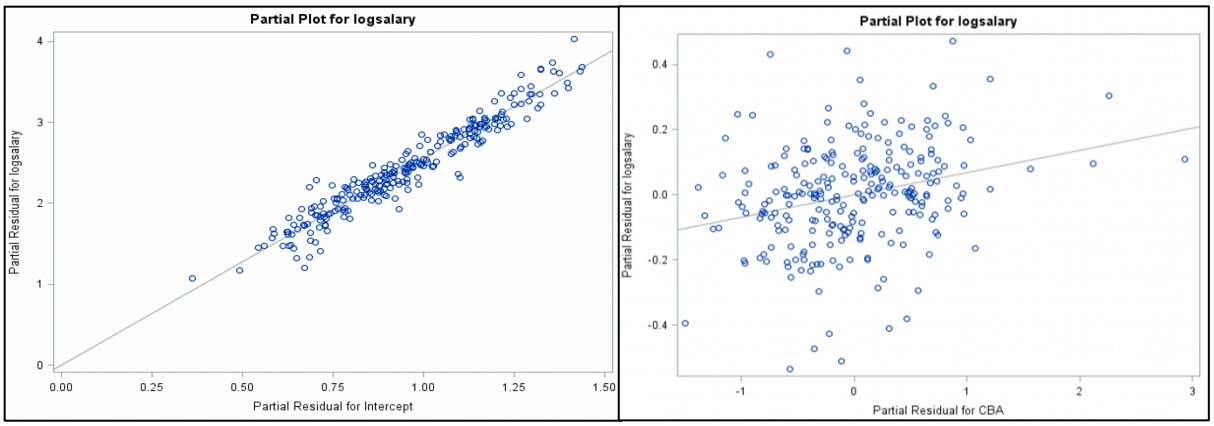


Figure 11: Partial plot for logsalary

Now, we look into whether the deployed model is working acceptably. We see that the overall F-test is significant.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	30.85220	2.57102	106.21	<.0001
Error	238	5.76127	0.02421		
Corrected Total	250	36.61346			

Table 15: Analysis of Variance

Its graphical alternative is the following plot, Residual-Fit Spread plot. The left panel is "centered fitted response" ($\hat{y}_i - \bar{y}$) versus "F_n ($\hat{y}_i - \bar{y}$)"; the right panel is "e_i" versus "(e_i)". If the deployed model works, the spread of residuals is expected to be SMALLER (shorter in the graph) than the spread of "centered fitted response". In this example, the line on the right is shorter than the line on the left, the deployed model seems working.

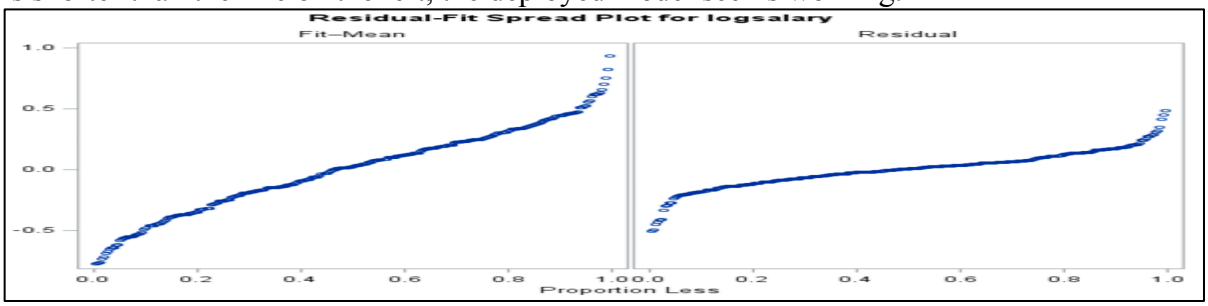


Figure 12: Residual-Fi Spread Plot for logsalary

Moreover, one might be wondering if the exponential orders of regressors are specified correctly. To see that, one could look into whether the leftover (residuals) is still related to each of the regressors. IF THERE IS STILL FUNCTIONAL PATTERN (not uniform banded) REVEALED, HIGHER ORDER TERMS MIGHT BE CONSIDERED. We use the following implementation to help us to see that.

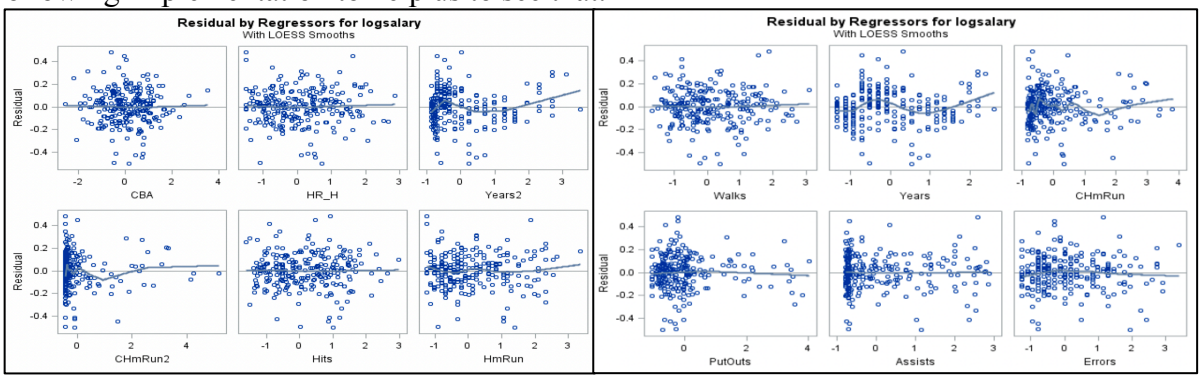


Figure 13: Residual plot for logsalary

Among all those three panels, there is no obvious pattern revealed. There seems no need to include higher order terms of regressors. If one is to improve the modeling, extra external regressors might be worth of investigation.

7. Model selection

Normally, we do regression analyses using only continuous variables. However, our dataset includes categorical predictors in a regression analysis. So, PROG REG is not appropriate in this case because there are categorical variables in this dataset. it is need to use PROC GLM which handle the numerical regressors as well as categorical ones.

There are large numbers of candidate predictor variables so Statistical model selection is needed to find out which ones are important to predict the results accurately. Model selection is to estimate the performance of different models in order to choose the approximate best model. Methods include familiar methods such as forward, backward, and stepwise selection. Also, new methods such as LAR, LASSO are also used.

7.1 Standard selection methods

There are traditional selection methods: Forward selection, backward, and stepwise.
- Forward Selection: Begins with just the intercept and at each step adds the effect that shows the largest contribution to the model.
- Backward Elimination: Begins with the full model and at each step deletes the effect that shows the smallest contribution to the model.
- Stepwise Selection: Modification of the forward selection technique that differs in that effects already in the model do not necessarily stay there.

7.2 Advanced selection methods

7.2.1 PROC GLMSELECT

PROC GLMSELECT provides several selection algorithms that you can modify by stating criteria for selecting effects, stopping the selection process, and choosing a model from the sequence of models at each step. In GLMSELECT, LAR, LASSO, ELASTICNET, STEPWISE are used.

- LAR (Least angle regression) -L1 norm

Least-angle regression (LARS) is a regression algorithm for high-dimensional data. This method is similar to forward selection. It starts with no effects in the model and adds effects. Firstly, all parameters are set to zero, and then parameters are added based on correlations with current residuals. If there are categorical variables in the model, then these variables are divided. When there are multiple variables which are correlated with one another, LARS that the variable selection appears to have problems with highly correlated variables.

- LASSO (At least absolute shrinkage and selection operator) – L2 norm

$$\min\left(||Y - X\theta||_2^2 + \lambda||\theta||_1\right)$$

Equation 2: LASSO Regression

This method adds and deletes parameters based on a version of ordinary least squares where the sum of the absolute regression coefficients is constrained. Like LAR, if there are categorical variables in the model, then these variables are divided.

In this Lasso function, Selection=LASSO (STEP=20, CHOOSE=AICC) is used.
This means that effects enter and leave successively until 20 steps. The CHOOSE=AICC (corrected Akaike's information criterion) identifies that the selected model be the model at step 18 that yields the optimal value of AICC statistic.

- ELASTICNET
ELASTICNET is a linear regression model trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple attributes which are correlated with one another. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both. It combines the penalties of both Ridge and LASSO, but with the option of unequal weights.

$$\min\left(\left|\left|Y - X\theta\right|\right|_2^2 + \lambda_1||\theta||_1 + \lambda_2||\theta||_2^2\right)$$

Equation 3: ELASTICNET Regression

- STEPWISE (SELECT=SL STOP=PRESS)
The stepwise selection process continues to the step where the stopping condition based on entry and stay significance levels is met. The PRESS statistic is used to choose among the models evaluated during the stepwise selection.

PROC GLMSELECT uses to determine the order in which effects enter and leave at each step. In LAR and LASSO, SELECT option can't be used. The SELECT= option are ADJRSQ, AIC, AICC, BIC, CP, CV, PRESS, RSQUARE, SBC, SL, and VALIDATE.

If STOP=n is specified, then PROC GLMSELECT stops selection at the first step for which the selected model has n effects.

7.2.2 PROC HPREG
The selection method in HPREG is a variant of the traditional stepwise selection where the decisions about what effects to add or drop at any stage and when to finish the selection are both based on the SBC (Schwarz Bayesian information criterion). The effect in the current model whose elimination makes the maximal decrease in the SBC statistic is dropped provided this lowers the SBC value. When no further decrease in the SBC value can be obtained by dropping an effect in the model, the effect whose addition to the model yields the lowest SBC statistic is added and the whole process is repeated. The method ends when dropping or adding any effect increases the SBC statistic.

7.2.3 PROC QUANTSELECT <- Adaptive LASSO
LASSO has non-ignorable bias when it estimates the nonzero coefficients. But, Adaptive LASSO produces unbiased estimated because it gives a somewhat higher penalization for zero coefficients and a lower penalty for nonzero coefficients.
When we use the code like SELECTION=LASSO (ADAPTIVE) we can use the adaptive LASSO method, which controls the effect selection process. The STOP=AIC option specifies that Akaike's information criterion (AIC) be used to determine the stopping condition. The CHOOSE=SBC option specifies that the Schwarz Bayesian information criterion (SBC) be used to determine the final selected model.

From Figure 15, Model Information shows that the effect selection settings. And, the default quantile type is single level, in which this effect selection is effective only for $\tau = 0.5$.

Model Information	
Data Set	WORK.MLB_HIT_MM
Dependent Variable	logsalary
Selection Method	Adaptive LASSO
Quantile Type	Single Level
Stop Criterion	AIC
Choose Criterion	SBC

Figure 15: Model Information

From Figure 16, the effect selection process starts with an intercept model at step 0. At step 1, the Years variable is added to the model that reduced the AIC value from -917.5703 to -1032.0176. You can see that step 11 has the minimum AIC and that step 9 has the minimum SBC. Normally, the SBC prefers a smaller model than the AIC.

Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	AIC	SBC
0	Intercept		1	-917.5703	-914.0449
1	Years		2	-1032.0176	-1024.9667
2	Years2		3	-1148.7379	-1138.1615
3	CHmRun		4	-1232.9181	-1218.8163
4	Hits		5	-1375.5581	-1357.9308
5	CWalks		6	-1380.8232	-1359.6705
6	CBA		7	-1395.8055	-1371.1274
7	PutOuts		8	-1412.6263	-1384.4226
8	CWalks2		9	-1431.0040	-1399.2749
9	Errors		10	-1434.7712	-1399.5167*
10	HR_H		11	-1434.9918	-1396.2118
11	League A		12	-1433.5415	-1391.2360
12		League A	11	-1434.9918*	-1396.2118
13	League A		12	-1433.5415	-1391.2360
14	CHmRun2		13	-1434.4605	-1388.6296
15	Walks		14	-1432.9422	-1383.5858
16	Division E		15	-1430.9971	-1378.1153
17	HmRun		16	-1429.1250	-1372.7177
18	Assists		17	-1427.1641	-1367.2314
* Optimal Value Of Criterion					

Figure 16: Selection Summary

Figure 17 demonstrates that the selection procedure stopped at step 10. All the AIC values for step 11 through step 17 are no less than the AIC at step 10.

Selection stopped at a local minimum of the AIC criterion.

Figure 17: Stop Reason

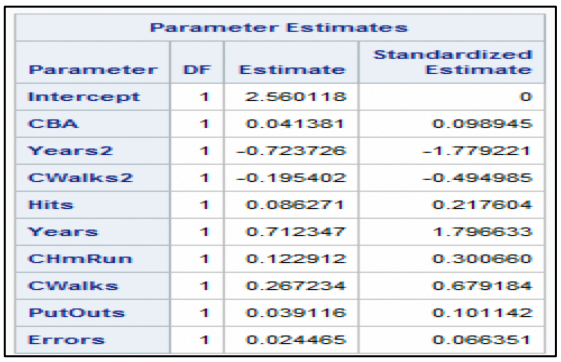
Figure 18 shows how the final selected model is created. The model at step 9 is chosen as the final selected model.

The model at step 9 is selected where SBC is -1399.52.

Figure 18: Selection Reason

Selected Effects:	Intercept CBA Years2 CWalks2 Hits Years CHmRun CWalks PutOuts Errors

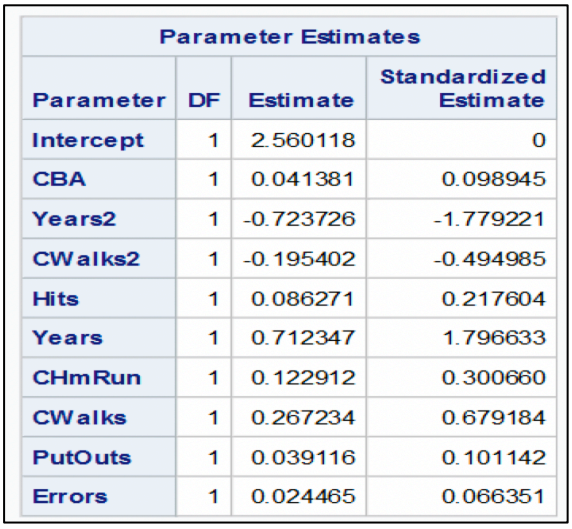
Figure 19: Selected Effects



Parameter Estimates			
Parameter	DF	Estimate	Standardized Estimate
Intercept	1	2.560118	0
CBA	1	0.041381	0.098945
Years2	1	-0.723726	-1.779221
CWalks2	1	-0.195402	-0.494985
Hits	1	0.086271	0.217604
Years	1	0.712347	1.796633
CHmRun	1	0.122912	0.300660
CWalks	1	0.267234	0.679184
PutOuts	1	0.039116	0.101142
Errors	1	0.024465	0.066351

Figure 20: Parameter Estimates

From Figure 20, the parameter estimates for the final selected model with $\tau = 0.5$. Models with $\tau = 0.1$ and $\tau = 0.9$, which stand for low-end salaries and high-end salaries respectively are also performed.



Parameter Estimates			
Parameter	DF	Estimate	Standardized Estimate
Intercept	1	2.560118	0
CBA	1	0.041381	0.098945
Years2	1	-0.723726	-1.779221
CWalks2	1	-0.195402	-0.494985
Hits	1	0.086271	0.217604
Years	1	0.712347	1.796633
CHmRun	1	0.122912	0.300660
CWalks	1	0.267234	0.679184
PutOuts	1	0.039116	0.101142
Errors	1	0.024465	0.066351

Figure 21: Parameter Estimates: $\tau = 0.1$

From figure 21, we can conclude that low-end salaries for MLB players depend mainly on CBA, Years, Walks, Hits, CHmRun, CWalks, Putouts, and Errors in 1986.

8. Conclusion (Evaluation)

The objective of regression analysis is 1) interpretation of regressors against response 2) prediction. Through this project, I attempt to check the relationship between the independent variable and dependent variable and increase the prediction accuracy. First, talk about doings to build the model properly. Because the dataset is small, extreme outlier can distort the prediction. So, deleting the outlier make the prediction accurate. Through domain knowledge, new variables are created and they become the significant variable for the target. And, variable transformation and resolving the multicollinearity also help fit the model.

When we check the "goodness-of-fit", R-squared, RMSE, and p-values are considered. The Root Mean Squared Error (RMSE) and R-square are statistics that typically inform the analyst how good the model is in predicting the target. The R-square is a measure of the fit of the model and ranges from 0 to 1.0 with higher values typically indicating a better model. The higher the R-squared values typically indicate a better performing model but sometimes conditions or the data used to train the model over-fit and don't represent the true value of the

prediction power of that particular model. Third, p-value measures how likely the coefficient has no effect on the outcome. If p-value is too high, the variable becomes insignificant.

- Comparison of the Model Selections

	R-Squared	Adjusted R-Squared	Root MSE
Forward	0.85233	0.84745	0.14947
Backward	0.85233	0.84745	0.14947
Stepwise	0.85233	0.84745	0.14947
LAR	0.0821	0.0784	0.36739
LASSO	0.8547	0.8474	0.14951
ELASTIC NET	0.8547	0.8474	0.14951
STEPWISE	**0.8551**	**0.8497**	**0.14836**
Adaptive LASSO	0.8523	0.8474	0.14947

Table 5: The comparison of results

"STEPWISE" has the highest R-Squared (0.8551), adjusted R-Squared (0.8497), and the lowest RMSE (0.14836). So, I have considered this model for prediction.

9. Recommendation (Actionable Insights)

The best model is stepwise (Select: SL, Stop: PRESS) with variables: CBA, Years2 CWalks2 Hits Years CHmRun CWalks League PutOuts, which has the most accurate prediction. In this part, the relationship between regressors and the response will be interpreted. Years (Number of years in the major leagues), Cwalks (Number of walks during his career), Hits (Number of hits in 1986) are important variables to predict the salaries. we also conclude the following insights after analyzing the dataset. If MLB players played for a long time, they had a lot of experience and they performed well so the MLB team needed to pay more on the experienced players. Secondly, if players have a lot of Cwalks, it means that they have higher on base average. This factor is important for hitters. Third factor, there is no need to say about Hits. If the player has the number of hits, they perform well so the team have to reward them. So, if the player played baseball for a long time, and has the high number of hits and walks, he has to pay higher.

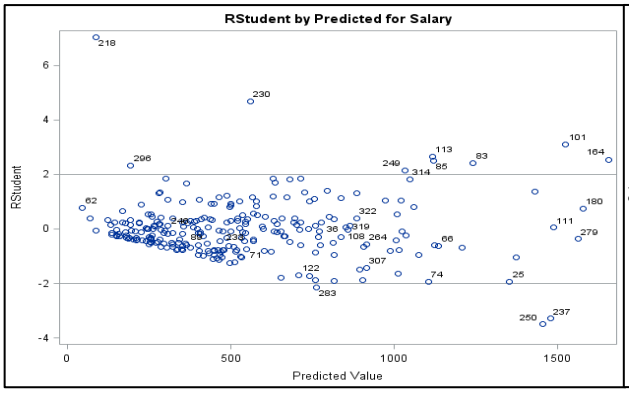
To improve the prediction models, deeper data exploration and better feature engineering are needed. The dataset is small so extreme outliers can distort the results easily. It is required to collect more data in order to predict the salary accurately. And, it is required to increase number of observations and try to get a better balance of the data set. If other factors in analysis such as All-Star game appearances, Most Valuable Player (MVP) awards, Gold Glove awards, and possibly the number of days on the disabled list in their career are included, the accuracy prediction is increasing more. And, if other advanced techniques such as decision tree, neural network are used, more accurate prediction can be accomplished.

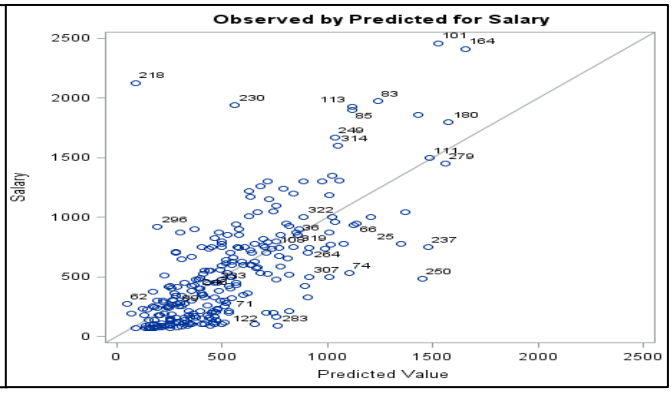
During the project, there are a lot of methods to predict the baseball players' salaries and according to changing the parameter (options) and model selection techniques, the output is changing a lot. That is, there is no perfect techniques. When facing with real world data, it is more challenging to solve the cases. Keep in mind that there is no such thing as a free lunch and try to choose the optimal model after many analyses.

REFERENCES

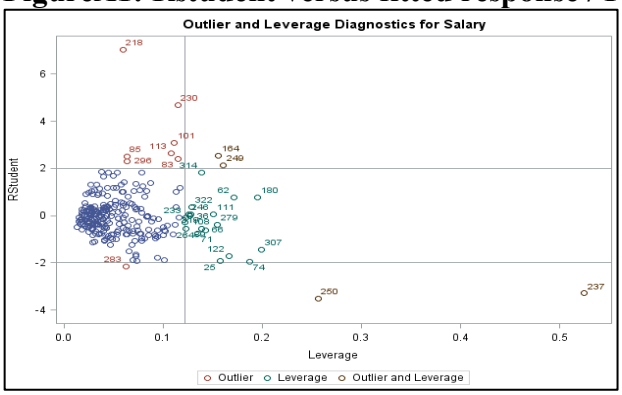
Reason for predicting salaries: https://www.forbes.com/sites/maurybrown/2018/05/22/mlb-wanted-a-salary-cap-for-decades-now-with-soaring-revenues-theyd-likely-reject-one/#6e884fce2676
Clinically significant regressors: https://en.wikipedia.org/wiki/Baseball_statistics
Handling missing values: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779
Multicollinearity: Concept: https://newonlinecourses.science.psu.edu/stat501/node/2/
Ridge Regression, Lasso, Elastic1: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/
Ridge Regression, Lasso, Elastic2: https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares-complexity
HPREG:
http://support.sas.com/documentation/cdl/en/stathpug/66860/HTML/default/viewer.htm#stathpug_hpreg_gettingstarted.htm
GLMSELECT:
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_glmselect_sect030.htm
QUANTSELECT1:
http://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/viewer.htm#statug_qrsel_gettingstarted.htm
QUANTSELECT2:
https://support.sas.com/documentation/onlinedoc/stat/143/qrsel.pdf
MODEL SELECTION:
https://support.sas.com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels.pdf
SAS CODE:
http://math.usu.edu/jrstevens/stat5100/

Appendix A: Residual analysis for salary





FigureA1: Rstudent versus fitted response / FigureA2: Observed by predicted for salary



FigureA3: Rstudent residual-Leverage plot

So, we can see that Rstudent residual-leverage plot provides possible outlying information in both dimensions.

As for (b.1), Detection of Heteroscedasticity, one can refers to the Rstudent residual-fitted response plot above. Excluding those possible outliers, there seems no serious violation against homoscedasticity. As for (b.2), Detection of (auto-) Correlation, we will use Durbin-Watson test.

Series of VIFi

The REG Procedure
Model: MODEL1
Dependent Variable: Salary

Durbin-Watson D	2.025
Pr < DW	0.5713
Pr > DW	0.4287
Number of Observations	263
1st Order Autocorrelation	-0.013

Table A1: Durbin-Watson Test

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

The note in the printout indicates how to interpret the result already. Remind yourself that the two p-values (Pr<DW and Pr>DW) refer to two different tests (H_0: no positive correlation;

H_0: no negative correlation, respectively). In this example, there is neither significant positive correlation nor significant negative correlation.
As for (b.3), Detection of serious violation against Normality, one could use the histogram and QQ-plot first.

Figure A4: Distribution of Residuals / Figure A5: Q-Q plot of Residuals

We see that there is no concern about serious violation against normality. We will also perform several hypothesis tests to confirm this. I have implemented a SAS macro, %NORMTEST(VAR, DATA), to facilitate this task. We will need to
(1) Define the macro in SAS
(2) Extract the targeted quantity whose normality is to be tested (Rstudent residual)
(3) Call the macro

Note that I have suppressed the printout and added an extra **OUTPUT** statement. **OUT**= specifies the name of dataset and **R**= specifies the name of raw residual; **RSTUDENT**= specifies the name of Rstudent residual residual; **STUDENT**= specifies the name of studentized residual; and **PRESS**= specifies the name of PRESS residual. (3) Call/Run the MACRO on ti in dataset fit.

%***NORMTEST***(TI,FIT)

We see the following testing results

NORMAL-TEST

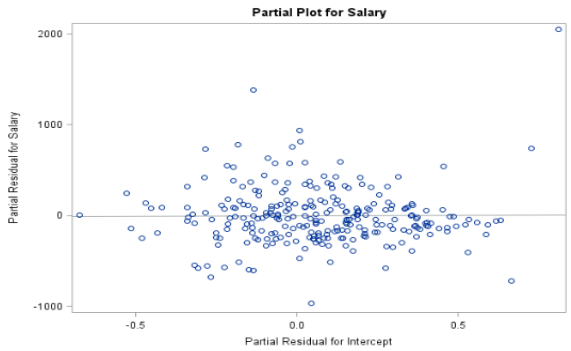
Variable: TI (Studentized Residual without Current Obs)

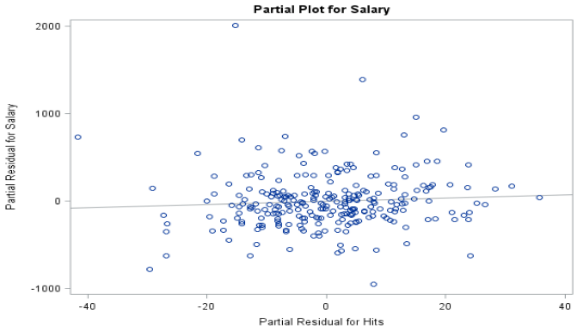
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.887365	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.132187	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.025537	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.749723	Pr > A-Sq	<0.0050

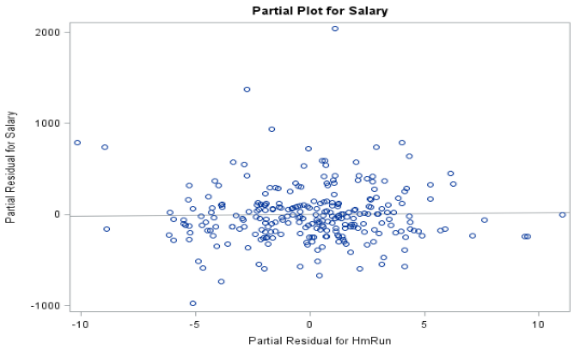
Table A2: Normal-Test

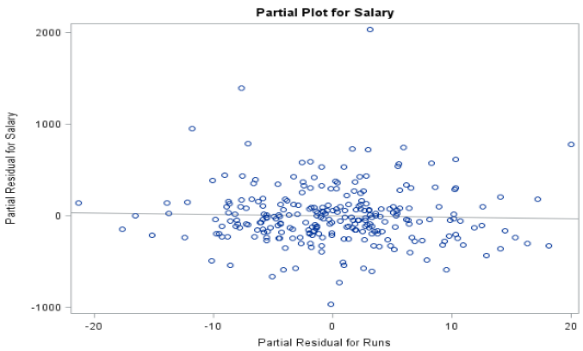
Again, we see some p-values are less than 0.05, H_0, data are normal, is hence rejected.
So, there can be a variable transformation to improve the model.

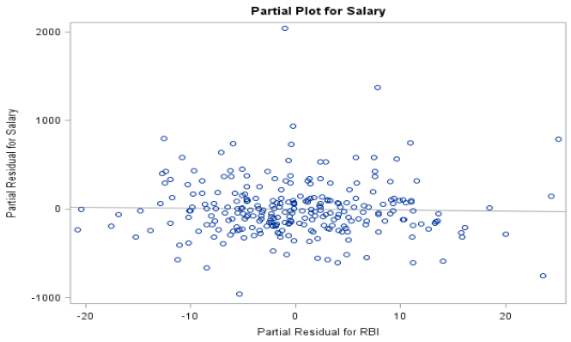
As for (c) misspecification of the model, we turn our attention to misspecification of the model. Before we discuss whether the currently deployed model works or not, one can look into partial regression plots (of regressors) to see their marginal potentials to contribute to explain response.

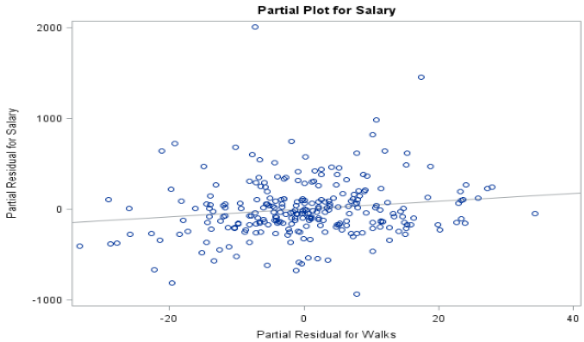

Partial Plot for Salary
Partial Residual for Salary
Partial Residual for Intercept

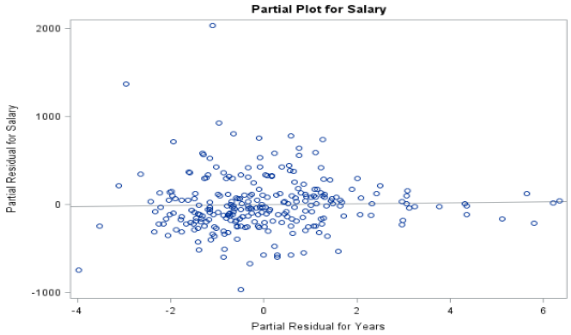

Partial Plot for Salary
Partial Residual for Salary
Partial Residual for Hits

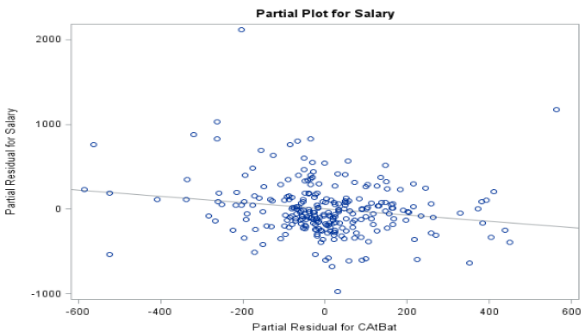

Partial Plot for Salary
Partial Residual for Salary
Partial Residual for HmRun

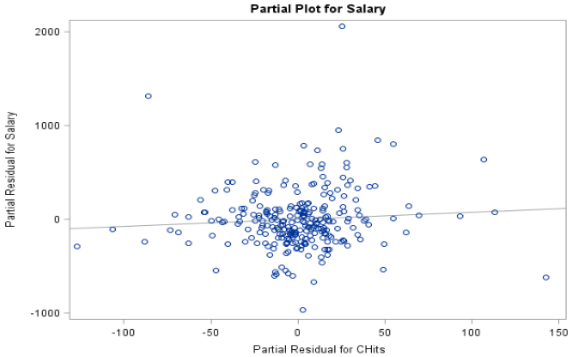

Partial Plot for Salary
Partial Residual for Salary
Partial Residual for Runs

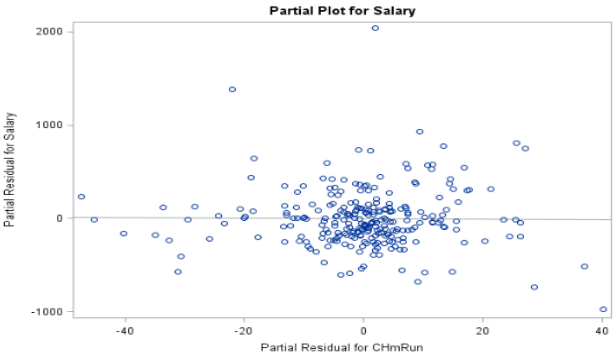

Partial Plot for Salary
Partial Residual for Salary
Partial Residual for RBI


Partial Plot for Salary
Partial Residual for Salary
Partial Residual for Walks


Partial Plot for Salary
Partial Residual for Salary
Partial Residual for Years


Partial Plot for Salary
Partial Residual for Salary
Partial Residual for CAtBat


Partial Plot for Salary
Partial Residual for Salary
Partial Residual for CHits


Partial Plot for Salary
Partial Residual for Salary
Partial Residual for CHmRun

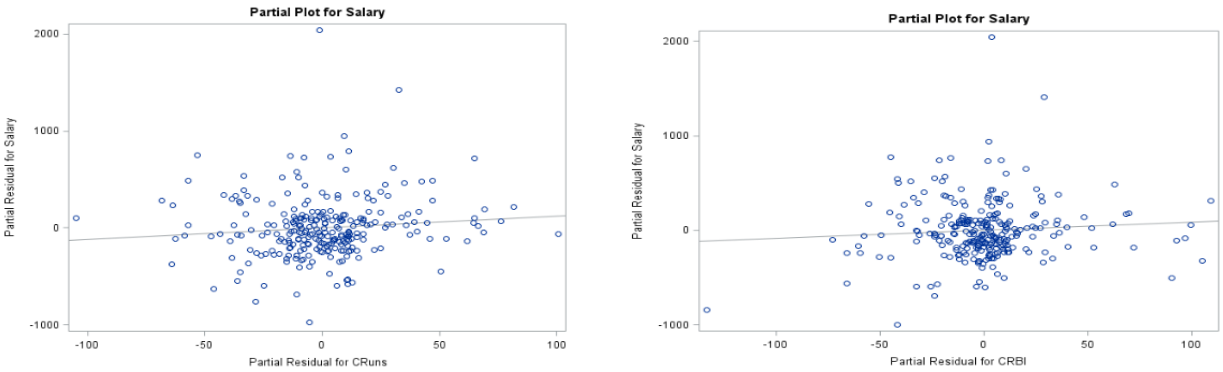


Figure A6: Partial plot for logsalary

Now, we look into whether the deployed model is working acceptably. We see that the overall F-test is significant.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	26921356	1794757	16.79	<.0001
Error	247	26397757	106874		
Corrected Total	262	53319113			

Table A3: Analysis Variance

Its graphical alternative is the following plot, Residual-Fit Spread plot. The left panel is "centered fitted response" ($\hat{y}_i - \bar{y}$) versus "F_n ($\hat{y}_i - \bar{y}$)"; the right panel is "e_i" versus "(e_i)". If the deployed model works, the spread of residuals is expected to be SMALLER (shorter in the graph) than the spread of "centered fitted response". In this example, the line on the right is shorter than the line on the left, the deployed model seems working.

Figure A7: Residual-Fit Spread Plot for Salary

Moreover, one might be wondering if the exponential orders of regressors are specified correctly. To see that, one could look into whether the leftover (residuals) is still related to each of the regressors. IF THERE IS STILL FUNCTIONAL PATTERN (not uniform banded) REVEALED, HIGHER ORDER TERMS MIGHT BE CONSIDERED. We use the following implementation to help us to see that.

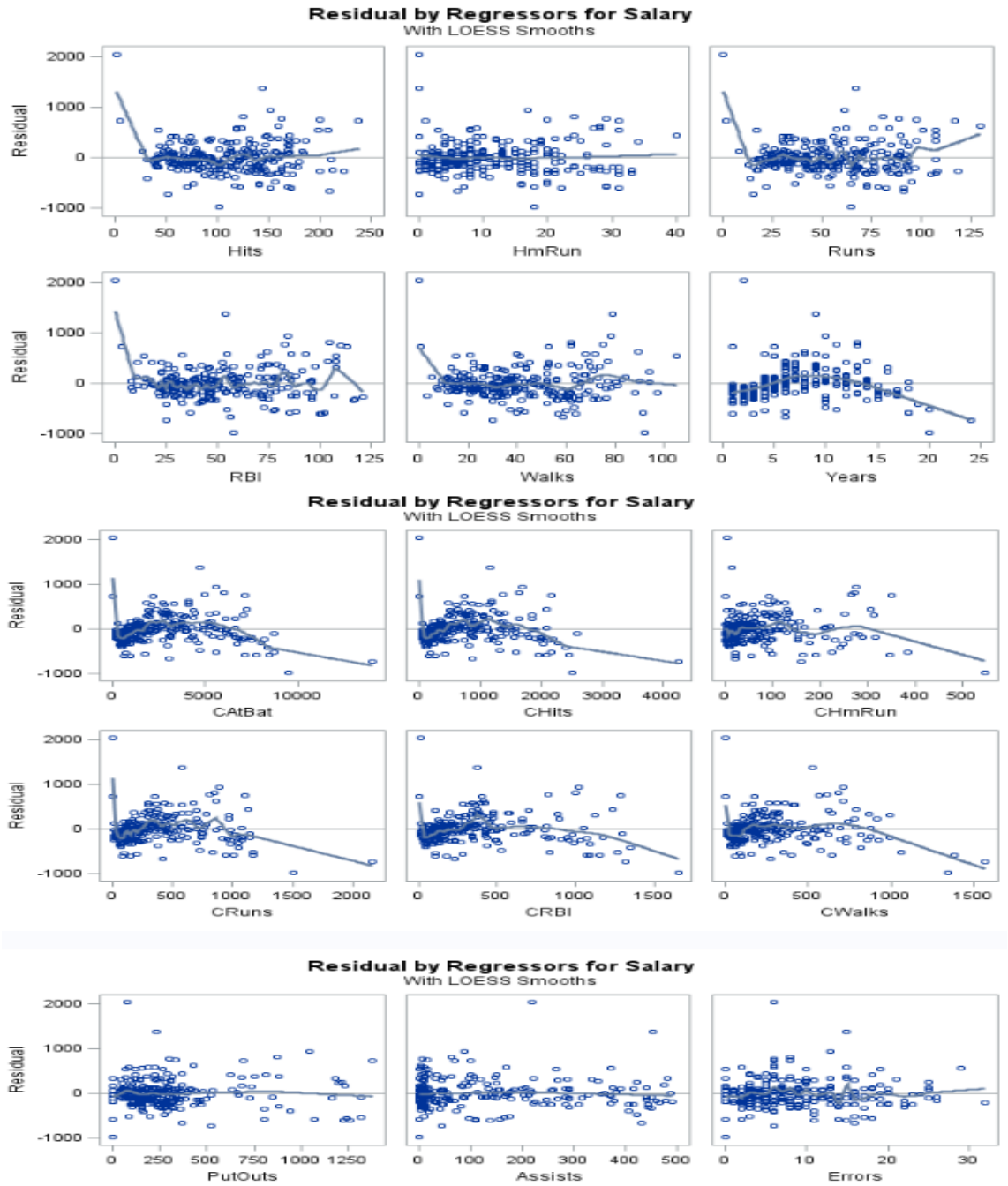


Figure A8: Residual plot for logsalary

Among all those three panels, there is obvious pattern revealed. There seems need to include higher order terms of regressors. If one is to improve the modeling, extra external regressors might be worth of investigation. **Thus, players near the start of their careers and players near the end of their careers get paid less than the model predicts. In order to address this lack of fit, it is needed to use polynomials of degree two for these variables.**

Appendix B: Correlation matrix

Pearson Correlation Coefficients, N = 310 Prob > |r| under H0: Rho=0									
	CBA	**HR_H**	**Years2**	**CAtBat2**	**CHits2**	**CHmRun2**	**CRuns2**	**CRBI2**	**CWalks2**
CBA	1.00000	-0.01206 0.8325	0.25723 <.0001	0.32206 <.0001	0.37719 <.0001	0.16514 0.0035	0.35367 <.0001	0.28996 <.0001	0.18443 0.0011
HR_H	-0.01206 0.8325	1.00000	0.16714 0.0032	0.15348 0.0068	0.12659 0.0258	0.37088 <.0001	0.18810 0.0009	0.25312 <.0001	0.23724 <.0001
Years2	0.25723 <.0001	0.16714 0.0032	1.00000	0.90956 <.0001	0.88167 <.0001	0.67547 <.0001	0.85530 <.0001	0.82788 <.0001	0.73722 <.0001
CAtBat2	0.32206 <.0001	0.15348 0.0068	0.90956 <.0001	1.00000	0.98879 <.0001	0.75011 <.0001	0.96475 <.0001	0.92545 <.0001	0.78727 <.0001
CHits2	0.37719 <.0001	0.12659 0.0258	0.88167 <.0001	0.98879 <.0001	1.00000	0.71890 <.0001	0.96249 <.0001	0.92561 <.0001	0.74106 <.0001
CHmRun2	0.16514 0.0035	0.37088 <.0001	0.67547 <.0001	0.75011 <.0001	0.71890 <.0001	1.00000	0.78834 <.0001	0.89786 <.0001	0.69362 <.0001
CRuns2	0.35367 <.0001	0.18810 0.0009	0.85530 <.0001	0.96475 <.0001	0.96249 <.0001	0.78834 <.0001	1.00000	0.91883 <.0001	0.83929 <.0001
CRBI2	0.28996 <.0001	0.25312 <.0001	0.82788 <.0001	0.92545 <.0001	0.92561 <.0001	0.89786 <.0001	0.91883 <.0001	1.00000	0.74479 <.0001
CWalks2	0.18443 0.0011	0.23724 <.0001	0.73722 <.0001	0.78727 <.0001	0.74106 <.0001	0.69362 <.0001	0.83929 <.0001	0.74479 <.0001	1.00000
Hits	0.57092 <.0001	0.04461 0.4338	-0.00426 0.9405	0.14622 0.0099	0.18259 0.0012	0.11702 0.0395	0.19791 0.0005	0.13628 0.0163	0.06520 0.2524
HmRun	0.24852 <.0001	0.77374 <.0001	0.11221 0.0484	0.20547 0.0003	0.20490 0.0003	0.38576 <.0001	0.26568 <.0001	0.29203 <.0001	0.21257 0.0002
Runs	0.46589 <.0001	0.18267 0.0012	-0.04264 0.4544	0.11163 0.0496	0.13407 0.0182	0.13903 0.0143	0.19535 0.0005	0.11523 0.0426	0.11553 0.0421
RBI	0.41997 <.0001	0.45863 <.0001	0.12015 0.0345	0.25187 <.0001	0.27228 <.0001	0.33810 <.0001	0.29968 <.0001	0.31552 <.0001	0.19040 0.0008
Walks	0.30459 <.0001	0.19726 0.0005	0.06968 0.2212	0.18976 0.0008	0.19587 0.0005	0.21219 0.0002	0.27866 <.0001	0.19505 0.0006	0.34412 <.0001
Years	0.30640 <.0001	0.16549 0.0035	0.96383 <.0001	0.86012 <.0001	0.83470 <.0001	0.61394 <.0001	0.81847 <.0001	0.75689 <.0001	0.69708 <.0001
CAtBat	0.40687 <.0001	0.15757 0.0054	0.90265 <.0001	0.95455 <.0001	0.94021 <.0001	0.68989 <.0001	0.93022 <.0001	0.85001 <.0001	0.76124 <.0001
CHits	0.46112 <.0001	0.13757 0.0154	0.88679 <.0001	0.95194 <.0001	0.95166 <.0001	0.67478 <.0001	0.93468 <.0001	0.85354 <.0001	0.73890 <.0001
CHmRun	0.27975 <.0001	0.46601 <.0001	0.74864 <.0001	0.81770 <.0001	0.79715 <.0001	0.93775 <.0001	0.85368 <.0001	0.91124 <.0001	0.74166 <.0001
CRuns	0.43859 <.0001	0.18405 0.0011	0.86197 <.0001	0.93498 <.0001	0.92792 <.0001	0.71685 <.0001	0.95750 <.0001	0.85007 <.0001	0.79395 <.0001
CRBI	0.39085 <.0001	0.29134 <.0001	0.87118 <.0001	0.94075 <.0001	0.93450 <.0001	0.83070 <.0001	0.93535 <.0001	0.94429 <.0001	0.77012 <.0001
CWalks	0.30678 <.0001	0.23924 <.0001	0.82061 <.0001	0.85407 <.0001	0.81883 <.0001	0.69396 <.0001	0.88845 <.0001	0.78309 <.0001	0.93601 <.0001
PutOuts	0.18789 0.0009	0.11235 0.0481	-0.00047 0.9934	0.06075 0.2863	0.08156 0.1520	0.09268 0.1034	0.06453 0.2573	0.10434 0.0666	0.04678 0.4118
Assists	0.02145 0.7068	-0.30019 <.0001	-0.07869 0.1670	-0.01051 0.8538	-0.01669 0.7698	-0.12447 0.0284	-0.03449 0.5452	-0.08937 0.1164	-0.02128 0.7090
Errors	-0.00188 0.9737	-0.16031 0.0047	-0.12745 0.0248	-0.04661 0.4135	-0.03747 0.5110	-0.09047 0.1119	-0.06223 0.2747	-0.07335 0.1978	-0.08937 0.1163

Appendix C: Ridge regression output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	30.31327	2.52611	78.41	<.0001
Error	246	7.92564	0.03222		
Corrected Total	258	38.23891			

Root MSE	0.17949	R-Square	0.7927
Dependent Mean	2.56814	Adj R-Sq	0.7826
Coeff Var	6.98925		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.54365	0.01142	222.72	<.0001
CBA	1	0.08070	0.01789	4.51	<.0001
HR_H	1	-0.03624	0.03260	-1.11	0.2674
Years2	1	-0.67284	0.05313	-12.66	<.0001
CHmRun2	1	-0.10021	0.05047	-1.99	0.0482
Hits	1	0.03496	0.03125	1.12	0.2644
HmRun	1	0.01365	0.03972	0.34	0.7314
Walks	1	0.05050	0.01476	3.42	0.0007
Years	1	0.72587	0.05418	13.40	<.0001
CHmRun	1	0.23922	0.06113	3.91	0.0001
PutOuts	1	0.02815	0.01228	2.29	0.0227
Assists	1	0.00374	0.01705	0.22	0.8267
Errors	1	0.00901	0.01581	0.57	0.5694

Appendix D: Principal component regression output

Number of Observations Read	318
Number of Observations Used	259
Number of Observations with Missing Values	59

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	21.55294	5.38823	82.02	<.0001
Error	254	16.68597	0.06569		
Corrected Total	258	38.23891			

Root MSE	0.25631	R-Square	0.5636
Dependent Mean	2.56814	Adj R-Sq	0.5568
Coeff Var	9.98021		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2.54923	0.01614	157.91	<.0001	0
z1	1	0.13109	0.00836	15.67	<.0001	1.00588
z2	1	0.05609	0.01038	5.40	<.0001	1.00545
z3	1	0.05875	0.01200	4.89	<.0001	1.01389
z4	1	-0.08908	0.01521	-5.86	<.0001	1.00840

Appendix E: Model Selection Output

- Forward selection

Selected Effects:	Intercept CBA Years2 CWalks2 Hits Years CHmRun CWalks PutOuts

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	31.20669	3.90084	174.60	<.0001
Error	242	5.40678	0.02234		
Corrected Total	250	36.61346			

Root MSE	0.14947
R-Square	0.85233
Adj R-Sq	0.84745
AIC	-692.28773
AICC	-691.37107
SBC	-913.55866
ASE	0.02154

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.555770	0.009585	266.65	<.0001
CBA	1	0.051783	0.013867	3.73	0.0002
Years2	1	-0.664203	0.046422	-14.31	<.0001
CWalks2	1	-0.181793	0.036902	-4.93	<.0001
Hits	1	0.082915	0.013553	6.12	<.0001
Years	1	0.659054	0.048794	13.51	<.0001
CHmRun	1	0.100176	0.018563	5.40	<.0001
CWalks	1	0.279928	0.046790	5.98	<.0001
PutOuts	1	0.046054	0.010177	4.53	<.0001

- Backward selection

Selected Effects:	Intercept CBA Years2 CWalks2 Hits Years CHmRun CWalks PutOuts

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	31.20669	3.90084	174.60	<.0001
Error	242	5.40678	0.02234		
Corrected Total	250	36.61346			

Root MSE	0.14947
R-Square	0.85233
Adj R-Sq	0.84745
AIC	-692.28773
AICC	-691.37107
SBC	-913.55866
ASE	0.02154

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.555770	0.009585	266.65	<.0001
CBA	1	0.051783	0.013867	3.73	0.0002
Years2	1	-0.664203	0.046422	-14.31	<.0001
CWalks2	1	-0.181793	0.036902	-4.93	<.0001
Hits	1	0.082915	0.013553	6.12	<.0001
Years	1	0.659054	0.048794	13.51	<.0001
CHmRun	1	0.100176	0.018563	5.40	<.0001
CWalks	1	0.279928	0.046790	5.98	<.0001
PutOuts	1	0.046054	0.010177	4.53	<.0001

- Stepwise selection

Selected Effects:	Intercept CBA Years2 CWalks2 Hits Years CHmRun CWalks PutOuts

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	31.20669	3.90084	174.60	<.0001
Error	242	5.40678	0.02234		
Corrected Total	250	36.61346			

Root MSE	0.14947
R-Square	0.85233
Adj R-Sq	0.84745
AIC	-692.28773
AICC	-691.37107
SBC	-913.55866
ASE	0.02154

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.555770	0.009585	266.65	<.0001
CBA	1	0.051783	0.013867	3.73	0.0002
Years2	1	-0.664203	0.046422	-14.31	<.0001
CWalks2	1	-0.181793	0.036902	-4.93	<.0001
Hits	1	0.082915	0.013553	6.12	<.0001
Years	1	0.659054	0.048794	13.51	<.0001
CHmRun	1	0.100176	0.018563	5.40	<.0001
CWalks	1	0.279928	0.046790	5.98	<.0001
PutOuts	1	0.046054	0.010177	4.53	<.0001

- LAR

Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept CWalks

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	3.00416	3.00416	22.26
Error	249	33.60930	0.13498	
Corrected Total	250	36.61346		

Root MSE	0.36739
Dependent Mean	2.58113
R-Square	0.0821
Adj R-Sq	0.0784
AIC	-247.67319
AICC	-247.57602
SBC	-493.62228

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	2.581206
CWalks	1	0.028835

- LASSO

Effects:	Intercept CBA Years2 CWalks2 Hits Walks Years CHmRun CWalks League_A Division_E PutOuts Errors

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	12	31.29369	2.60781	116.67
Error	238	5.31978	0.02235	
Corrected Total	250	36.61346		

Root MSE	0.14951
Dependent Mean	2.58113
R-Square	0.8547
Adj R-Sq	0.8474
AIC	-688.35931
AICC	-686.57965
SBC	-895.52842

- ELASTIC NET

Effects:	Intercept CBA Years2 CWalks2 Hits Walks Years CHmRun CWalks League_A Division_E PutOuts Errors

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	12	31.29369	2.60781	116.67
Error	238	5.31978	0.02235	
Corrected Total	250	36.61346		

Root MSE	0.14951
Dependent Mean	2.58113
R-Square	0.8547
Adj R-Sq	0.8474
AIC	-688.35931
AICC	-686.57965
SBC	-895.52842

- Stepwise (Select: SL, Stop: PRESS)

Effects:	Intercept CBA Years2 CWalks2 Hits Years CHmRun CWalks League PutOuts

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	9	31.30915	3.47879	158.06
Error	241	5.30431	0.02201	
Corrected Total	250	36.61346		

Root MSE	0.14836
Dependent Mean	2.58113
R-Square	0.8551
Adj R-Sq	0.8497
AIC	-695.09023
AICC	-693.98563
PRESS	5.87684
SBC	-912.83570

- Adaptive LASSO

Effects:	Intercept CBA Years2 CWalks2 Hits Years CHmRun CWalks PutOuts

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	8	31.20669	3.90084	174.60
Error	242	5.40678	0.02234	
Corrected Total	250	36.61346		

Root MSE	0.14947
Dependent Mean	2.58113
R-Square	0.8523
Adj R-Sq	0.8474
AIC	-692.28773
AICC	-691.37107
SBC	-913.55866
CV PRESS	5.81679

Cross Validation Details

	Observations		
Index	Fitted	Left Out	CV PRESS
1	200	51	1.2622
2	201	50	1.1958
3	201	50	1.2947
4	201	50	0.5928
5	201	50	1.4713
Total			5.8168