Major League Baseball 1987 Salary Regression Model

Kamden Brown, Brian Bruxvoort, and Klein Wang

Department of Statistics, Truman State University

STAT 478: Regression Analysis

Dr. Carol Thatcher

7 December 2023

Introduction

The dataset used is the Hitters dataset, which can be found on Kaggle. The data frame includes 322 observations of major league baseball players. The original dataset includes 20 variables, but the full model used in this analysis will only be looking at 7 of those variables, the career statistics, which are as follows: Years, CAtBat, CHits, CHmRun, CRuns, CRBI, CWalks. Years refers to the number of years the player has been in major league baseball. CAtBat is the total number of at bats for the player in that player's career. CHits is the total number of hits for a player throughout the player's career. CHmRun is the total number of home runs hit by a player during the player's career. CRuns refers to the total number of runs scored by a player over the course of the player's career. CRBI is the total number of runs batted in by a player within that player's career. CWalks refers to the total number of walks by a player during the course of the player's career. The full model is used to predict the response variable, the opening day annual salary (in thousands of dollars) of a major league baseball player in 1987.The objective of this analysis is to find the "best" model for predicting a major league baseball player's 1987 opening day annual salary (in thousands of dollars) using a player's career statistics.

Methods

Several methods were used to determine the "best" model for predicting salary. The methods performed were: Mallow's Cp, backward elimination, forward selection, and stepwise selection.

Mallow's Cp measures the bias in prediction from using a reduced model rather than a full model with all potential predictors.

The backward elimination method is another variable selection method that starts by selecting all variables in the full model and testing each variable for significance. If all variables are found to be significant the test stops, but if one or more variables is found to be not significant, least important, then the variable with the smallest AIC is removed from the model and the new model is run under the same conditions. This method continues to remove one variable at a time until all remaining variables are significant. Another method used was Forward Selection.

Forward selection is an additional method used for finding the “best” model. Forward selection begins by considering all one variable, simple linear models. If none of the variables in any of the models are found significant the method stops; however, if one or more predictors is found significant, smallest p-value or largest AIC, that predictor is added to the model. The process repeats this time looking for the best two variable models with one of the variables being the most significant variable from the one variable models. Continue adding one variable at a time until no more significant variables can be added. The resulting model is the “best” model. The last method used for determining the “best” model was the stepwise selection.

Stepwise selection involves both forward selection and backward elimination techniques. It starts with a null model and tests all one variable models with the predictor with the smallest p-value or largest AIC being added to the null model. This process is repeated with the new model and is tested against all two variable models. The process then alternates between adding a variable, all the variables in that model for significance and then removing the variable with the least significance until no more variables can be added or removed to the model.

All four methods for selecting the “best” model came back with the same five predictors, which made choosing the final model easy as it was just the model with the five variables returned from the tests.

The X^TX matrix assesses the multicollinearity and VIF of the predictor variables in the regression model. It serves as a fundamental tool to examine the relationship among predictor variables. The diagonal elements in the matrix represent the sum of squared values for each individual predictor variable. The information is crucial in understanding the variance of the corresponding coefficients in the linear regression model, which offers a comprehensive view of the total variability associated with each predictor.

The overall F test evaluates the significance of the linear regression model as a whole. The test is conducted by comparing the obtained F statistic to the critical F value, derived from the F distribution with appropriate degrees of freedom. The F test provides a useful assessment of whether the model significantly contributes to explaining the variability in the response variable. It is an important tool for validating the reliability of the regression analysis and informing subsequent interpretations of the individual predictors.

The R-squared and adjusted R-squared are metrics employed to measure the goodness of fit of a linear regression model. R-squared measures the proportion of variability in the response variable explained by the predictor variables in the model. Meanwhile, the adjusted R-squared penalizes the inclusion of irrelevant predictors, providing a more realistic assessment of the model’s explanatory power.

The individual t-values for marginal effects assess the significance of each predictor variable’s impact on the response variable in a linear regression model. These t-values measure whether the estimated coefficient for each predictor is significantly different from zero.

Typically, predictor variables with lower p-values (0.05 as chosen significance level) as statistically significant contributors to the model.

Confidence intervals for mean response assess the reliability of a linear regression model. The output provides a range within which the average response is expected to fall, offering insight into the precision and accuracy of the model's predictions. Unlike the confidence interval for mean response, the prediction interval for a new individual provides a wider range that encompasses the variability of individual observations. It takes into account not only the inherent variability in the data but also the uncertainty associated with predicting the response for a specific individual.

The variance Inflation Factor (VIF) is a statistical measure employed in regression analysis to assess the extent of multicollinearity among predictor variables. VIF values are calculated for each predictor variable, measuring how much the variance of the estimated regression coefficients is inflated due to multicollinearity. Elevated VIF values, typically above 10, suggest a high degree of correlation and may consider further investigation, such as variable selection or standardization.

The Durbin-Watson test is utilized to find out if there is autocorrelation at lag 1 in the residuals (prediction errors) from a regression analysis. A value of two in the Durbin-Watson statistic denotes no autocorrelation. The statistic ranges from 0 to 4. Positive autocorrelation is indicated by values less than 2, and negative autocorrelation is shown by values more than 2. In general, there may be a significant autocorrelation issue if the Durbin-Watson statistic is smaller than 1.5 or larger than 2.5. Otherwise, autocorrelation is probably not a reason for alarm if the statistic falls between 1.5 and 2.5.

Deleted t residuals are used for detecting outliers. Deleted t residuals with a t of ±2 are considered mild outliers, while t residuals exceeding ±3 are deemed moderate outliers.

For detecting influential points, three statistics were calculated: Cook's Distance, DFBETA, DFFIT, with subsequent plots created for each test. Cook's Distance measures the change in the set of coefficients if the observation is removed. Cook's Distance has a cutoff of $D > F_{0.50,\, p\, ,n-p}$. Cook's D greater than the cutoff is deemed an influential point. DFBETA measures the change in each coefficient if the observation is removed. DFBETA uses a cutoff of $|DFBETA| > 2/\sqrt{n}$. DFBETAs larger than the cutoff are regarded as influential points. DFFIT measures the change in the predicted value if the observation is removed. DFFIT has a cutoff criteria of $|DFFIT| > 2\sqrt{(p/n)}$. Any DFFIT value greater than the cutoff value is considered an influential point.

Analysis

The full model uses 7 variables: Years, CAtBat, CHits, CHmRun, CRuns, CRBI, and CWalks to predict Salary. A t-test was performed on the full model to determine which variables were significant in predicting a major league baseball player's salary. Individual t-values are shown in Figure 1. CAtBat is a statistically significant variable because the critical value is 1.969 which is calculated by $t_{0.025,\, 255}$. The comparison between its t-value (-3.285) and the critical value suggests that it is the only significant variable in predicting Salary. Further, model selection technique is being used to find the best model among 7 variables since not much information was being found from the individual t-test from 7 variables in the model.

```
Call:
lm(formula = Salary ~ ., data = hitters)

Residuals:
     Min       1Q   Median       3Q      Max
-1154.24  -238.80   -55.58   141.67  1788.47

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 378.91468   42.65356   8.884  < 2e-16 ***
Years       -22.34744   12.33100  -1.812  0.07112 .
CAtBat       -0.41675    0.12687  -3.285  0.00116 **
CHits         1.06505    0.62017   1.717  0.08713 .
CHmRun       -0.39291    1.55537  -0.253  0.80077
CRuns         1.05626    0.67679   1.561  0.11984
CRBI          0.99440    0.68949   1.442  0.15047
CWalks       -0.09272    0.27779  -0.334  0.73883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 351.3 on 255 degrees of freedom
Multiple R-squared:  0.4098,    Adjusted R-squared:  0.3936
F-statistic: 25.29 on 7 and 255 DF,  p-value: < 2.2e-16
```

Figure 1. Linear regression model using all 7 variables in the dataset to predict Salary

All subset for the 7 variables is shown in Figure 2. Each row in the table represents a unique combination of predictors along with Mallow's C_p. The value around P_R or smaller is being considered as a good model. P_R=k+1, which is the number of coefficients in the candidate model. Mallow's C_p technique for all subsets found the model with Years, CAtBat, CHits, CRuns, and CRBI as predictor variables to be significant which had a C_p value (4.138630) lower than P_R, which is 6 in this model (Figure 2).

		Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	
1	2	0	0	0	0	0	1	0	34.169376
1	2	0	0	0	0	1	0	0	36.262189
1	2	0	0	1	0	0	0	0	42.874542
1	2	0	1	0	0	0	0	0	53.452637
1	2	0	0	0	1	0	0	0	53.999733
1	2	0	0	0	0	0	0	1	69.392246
1	2	1	0	0	0	0	0	0	103.697064
2	3	1	0	0	0	1	0	0	22.253373
2	3	1	0	0	0	0	1	0	22.658676
2	3	1	0	1	0	0	0	0	25.939209
2	3	0	1	1	0	0	0	0	27.243643
2	3	0	1	0	0	1	0	0	29.167453
2	3	0	0	0	0	1	1	0	33.296318
2	3	0	0	0	1	1	0	0	33.315882
2	3	0	0	0	0	1	0	1	35.042555
2	3	0	0	1	1	0	0	0	35.043827
2	3	0	1	0	0	0	1	0	35.421771
3	4	0	1	1	1	0	0	0	7.549015
3	4	0	1	1	0	0	1	0	10.367786
3	4	1	0	0	0	1	1	0	14.024738
3	4	1	0	1	0	0	1	0	15.755954
3	4	1	0	1	1	0	0	0	16.387821
3	4	0	1	1	0	1	0	0	17.434149
3	4	0	1	0	0	1	1	0	18.702181
3	4	1	0	0	1	1	0	0	19.399079
3	4	1	0	0	1	0	1	0	21.492473
3	4	1	1	0	0	0	1	0	21.753865
4	5	1	1	1	1	0	0	0	5.435581
4	5	0	1	1	0	1	1	0	5.556476
4	5	1	1	1	0	0	1	0	6.787453
4	5	0	1	1	1	1	0	0	7.254296
4	5	0	1	1	1	0	0	1	8.876675
4	5	0	1	1	1	0	1	0	9.507871
4	5	0	1	1	0	0	1	1	10.839999
4	5	1	1	0	0	1	1	0	13.092848
4	5	0	1	0	1	1	1	0	13.108995
4	5	1	0	0	0	1	1	1	13.798214
5	6	1	1	1	0	1	1	0	4.138630
5	6	1	1	1	1	1	0	0	6.092879
5	6	1	1	1	1	0	0	1	6.825050
5	6	1	1	1	1	0	1	0	7.167977
5	6	0	1	1	0	1	1	1	7.323115
5	6	1	1	1	0	0	1	1	7.504031
5	6	0	1	1	1	1	1	0	7.554014
5	6	0	1	1	1	1	0	1	9.246386
5	6	1	1	0	1	1	1	0	9.256936
5	6	0	1	1	1	0	1	1	10.778691
6	7	1	1	1	0	1	1	1	6.063814
6	7	1	1	1	1	1	1	0	6.111399
6	7	1	1	1	1	1	0	1	8.079999
6	7	1	1	1	1	0	1	1	8.435765
6	7	1	1	0	1	1	1	1	8.949252
6	7	0	1	1	1	1	1	1	9.284421
6	7	1	0	1	1	1	1	1	16.790655
7	8	1	1	1	1	1	1	1	8.000000

Figure 2. Model Selection: all subsets (Mallow's C_p technique)

A backward elimination was run on the full model to determine which variables were significant (Figure 3). The backward elimination found Years, CAtBat, CHits, CRuns, and CRBI to be significant. This in turn gives the following model: Salary = 380.8004 - 22.6366Years - 0.4343CAtBat + 1.2376CHits + 0.8566CRun + 0.8219CRBI as the "best" model using backward elimination (Figure 4).

```
Start:  AIC=3091.09
Salary ~ Years + CAtBat + CHits + CHmRun + CRuns + CRBI + CWalks

         Df Sum of Sq      RSS    AIC
- CHmRun  1      7875 31477118 3089.2
- CWalks  1     13748 31482990 3089.2
<none>                31469243 3091.1
- CRBI    1    256690 31725933 3091.2
- CRuns   1    300595 31769837 3091.6
- CHits   1    363964 31833206 3092.1
- Years   1    405326 31874569 3092.5
- CAtBat  1   1331662 32800904 3100.0

Step:  AIC=3089.16
Salary ~ Years + CAtBat + CHits + CRuns + CRBI + CWalks

         Df Sum of Sq      RSS    AIC
- CWalks  1      9233 31486351 3087.2
<none>                31477118 3089.2
- Years   1    402226 31879344 3090.5
- CRuns   1    424553 31901671 3090.7
- CHits   1    900774 32377892 3094.6
- CAtBat  1   1441219 32918336 3098.9
- CRBI    1   1709189 33186307 3101.1

Step:  AIC=3087.24
Salary ~ Years + CAtBat + CHits + CRuns + CRBI

         Df Sum of Sq      RSS    AIC
<none>                31486351 3087.2
- Years   1    421792 31908143 3088.7
- CRuns   1    573706 32060056 3090.0
- CHits   1   1351847 32838198 3096.3
- CAtBat  1   1683282 33169633 3098.9
- CRBI    1   1704644 33190994 3099.1
```

Figure 3. Model Selection: Backward Elimination

```
Coefficients:
(Intercept)        Years       CAtBat        CHits        CRuns         CRBI
   380.8004     -22.6366      -0.4343       1.2376       0.8566       0.8219
```

Figure 4. Coefficients Table for Backward Elimination

Another method used for determining the “best” model was forward selection, which found CRBI, Years, CRuns, CAtBat, and CHits to be significant predictors of Salary (Figure 5). Forward selection returned a reduced model: Salary = 380.8004 + 0.8219CRBI - 22.6366Years + 0.8566CRuns - 0.4343CAtBat + 1.2376CHits (Figure 6).

```
Start:  AIC=3215.77
Salary ~ 1

          Df Sum of Sq       RSS    AIC
+ CRBI     1  17139434 36179679 3115.8
+ CRuns    1  16881162 36437951 3117.6
+ CHits    1  16065140 37253973 3123.5
+ CAtBat   1  14759710 38559403 3132.5
+ CHmRun   1  14692193 38626920 3133.0
+ CWalks   1  12792622 40526491 3145.6
+ Years    1   8559105 44760007 3171.7
<none>                 53319113 3215.8

Step:  AIC=3115.78
Salary ~ CRBI

          Df Sum of Sq       RSS    AIC
+ Years    1   1667339 34512340 3105.4
+ CRuns    1    354561 35825119 3115.2
<none>                 36179679 3115.8
+ CAtBat   1     92261 36087418 3117.1
+ CHits    1     75469 36104210 3117.2
+ CWalks   1     51974 36127705 3117.4
+ CHmRun   1       515 36179165 3117.8

Step:  AIC=3105.37
Salary ~ CRBI + Years

          Df Sum of Sq       RSS    AIC
+ CRuns    1   1312321 33200019 3097.2
+ CHits    1   1098674 33413666 3098.9
+ CHmRun   1    390737 34121603 3104.4
+ CAtBat   1    358479 34153861 3104.6
<none>                 34512340 3105.4
+ CWalks   1     28554 34483786 3107.2

Step:  AIC=3097.17
Salary ~ CRBI + Years + CRuns

          Df Sum of Sq       RSS    AIC
+ CAtBat   1    361821 32838198 3096.3
+ CWalks   1    274773 32925246 3097.0
<none>                 33200019 3097.2
+ CHmRun   1     71602 33128416 3098.6
+ CHits    1     30385 33169633 3098.9

Step:  AIC=3096.29
Salary ~ CRBI + Years + CRuns + CAtBat

          Df Sum of Sq       RSS    AIC
+ CHits    1   1351847 31486351 3087.2
+ CHmRun   1    720203 32117995 3092.5
+ CWalks   1    460305 32377892 3094.6
<none>                 32838198 3096.3

Step:  AIC=3087.24
Salary ~ CRBI + Years + CRuns + CAtBat + CHits

          Df Sum of Sq       RSS    AIC
<none>                 31486351 3087.2
+ CWalks   1    9232.9 31477118 3089.2
+ CHmRun   1    3360.6 31482990 3089.2
```

Figure 5. Model Selection: Forward Selection

```
Coefficients:
(Intercept)         CRBI        Years        CRuns       CAtBat        CHits
   380.8004       0.8219     -22.6366       0.8566      -0.4343       1.2376
```

Figure 6. Coefficients Table for Forward Selection

The last technique used for model selection was stepwise selection. This method found CRBI, Years, CRuns, CAtBat, and CHits to be significant predictors of Salary (Figure 7). Stepwise selection determined the following reduced model: Salary = 380.8004 + 0.8219CRBI - 22.6366Years + 0.8566CRuns - 0.4343CAtBat + 1.2376CHits (Figure 8).

```
Start:  AIC=3215.77
Salary ~ 1

         Df Sum of Sq      RSS    AIC
+ CRBI    1  17139434 36179679 3115.8
+ CRuns   1  16881162 36437951 3117.6
+ CHits   1  16065140 37253973 3123.5
+ CAtBat  1  14759710 38559403 3132.5
+ CHmRun  1  14692193 38626920 3133.0
+ CWalks  1  12792622 40526491 3145.6
+ Years   1   8559105 44760007 3171.7
<none>                53319113 3215.8

Step:  AIC=3115.78
Salary ~ CRBI

         Df Sum of Sq      RSS    AIC
+ Years   1   1667339 34512340 3105.4
+ CRuns   1    354561 35825119 3115.2
<none>                36179679 3115.8
+ CAtBat  1     92261 36087418 3117.1
+ CHits   1     75469 36104210 3117.2
+ CWalks  1     51974 36127705 3117.4
+ CHmRun  1       515 36179165 3117.8
- CRBI    1  17139434 53319113 3215.8

Step:  AIC=3105.37
Salary ~ CRBI + Years

         Df Sum of Sq      RSS    AIC
+ CRuns   1   1312321 33200019 3097.2
+ CHits   1   1098674 33413666 3098.9
+ CHmRun  1    390737 34121603 3104.4
+ CAtBat  1    358479 34153861 3104.6
<none>                34512340 3105.4
+ CWalks  1     28554 34483786 3107.2
- Years   1   1667339 36179679 3115.8
- CRBI    1  10247667 44760007 3171.7

Step:  AIC=3097.17
Salary ~ CRBI + Years + CRuns

         Df Sum of Sq      RSS    AIC
+ CAtBat  1    361821 32838198 3096.3
+ CWalks  1    274773 32925246 3097.0
<none>                33200019 3097.2
+ CHmRun  1     71602 33128416 3098.6
+ CHits   1     30385 33169633 3098.9
- CRBI    1   1262304 34462322 3105.0
- CRuns   1   1312321 34512340 3105.4
- Years   1   2625100 35825119 3115.2

Step:  AIC=3096.29
Salary ~ CRBI + Years + CRuns + CAtBat

         Df Sum of Sq      RSS    AIC
+ CHits   1   1351847 31486351 3087.2
+ CHmRun  1    720203 32117995 3092.5
+ CWalks  1    460305 32377892 3094.6
<none>                32838198 3096.3
- CAtBat  1    361821 33200019 3097.2
- Years   1    939059 33777256 3101.7
- CRuns   1   1315663 34153861 3104.6
- CRBI    1   1558024 34396222 3106.5

Step:  AIC=3087.24
Salary ~ CRBI + Years + CRuns + CAtBat + CHits

         Df Sum of Sq      RSS    AIC
<none>                31486351 3087.2
- Years   1    421792 31908143 3088.7
+ CWalks  1      9233 31477118 3089.2
+ CHmRun  1      3361 31482990 3089.2
- CRuns   1    573706 32060056 3090.0
- CHits   1   1351847 32838198 3096.3
- CAtBat  1   1683282 33169633 3098.9
- CRBI    1   1704644 33190994 3099.1
```

Figure 7. Model Selection: Stepwise Selection

```
Coefficients:
(Intercept)         CRBI        Years        CRuns       CAtBat        CHits
   380.8004       0.8219     -22.6366       0.8566      -0.4343       1.2376
```

Figure 8. Coefficients Table for Stepwise Selection

Backward elimination, forward selection, and stepwise selection all returned the same five variables: CRBI, Years, CRuns, CAtBat, and CHits, which made selecting the final model easy. The final model chosen is: Salary = 380.8004 + 0.8219CRBI - 22.6366Years + 0.8566CRuns - 0.4343CAtBat + 1.2376CHits.

The matrix X serves as the foundational structure for our linear regression model, comprising 263 rows and 6 columns. The first column represents the intercept term with a constant value of 1, while the subsequent columns contain the values of predictor variables—CRBI, Years, CRuns, CAtBat, and CHits—for each observation. The X^TX matrix is obtained by multiplying the transpose of X by X, which results in a 6x6 matrix. The off-diagonal elements indicate the covariances between pairs of predictor variables. Meanwhile, the diagonal elements in X^TX matrix are particularly important because they represent the sum of squared values for each individual predictor variable. These values are directly related to the variance of the corresponding coefficients in the linear regression model and indicate the total variability of each predictor variable.

In Figure 9, X^TX [1,1] equals 263, representing the sum of squared ones and symbolizing the total number of observations in the dataset. X^TX [2, 2] equals 56,109,808, which represents the sum of squared CRBI values. This substantial value indicates considerable variability in the CRBI variable across observations, suggesting a wide range of runs batted in during players' careers. X^TX [3, 3] equals 20,081, reflecting the sum of squared Years values. A relatively

smaller value compared to others indicates less variability in the "Years" variable. The value is expected, as the number of years a player has been in the major leagues may not vary as widely as other performance-related variables. X^TX [4, 4] equals 63,055,745, representing the sum of squared CRuns values. It indicates that players have diverse career run totals. X^TX [5, 5] equals 3,227,304,578, which represents the sum of squared CAtBat values. It indicates that players have a wide range of career at-bat counts among the players in the dataset. X^TX [6, 6] equals 247,251,105, which represents the sum of squared CHits values. It suggests that players have a diverse number of hits in their careers.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	263	86900	1923	95001	698934	189935
[2,]	86900	56109808	986210	57925680	415120396	114753247
[3,]	1923	986210	20081	1059279	7740092	2119692
[4,]	95001	57925680	1059279	63055745	447462034	123985841
[5,]	698934	415120396	7740092	447462034	3227304578	891167503
[6,]	189935	114753247	2119692	123985841	891167503	247251105

Figure 9. X^TX matrix for CRBI, Years, CRuns, CAtBat, and CHits variables

Next, the overall F test in regression analysis is being assessed to determine whether there is a significant linear relationship between the predictors variables as a group and the response variable. Null hypothesis states that the coefficients for all predictor variables in the model are equal to zero. Alternative hypothesis states that at least one of the coefficients for the predictor variables is not equal to zero. The result of the overall F test is provided in Figure 10. The F statistic (35.64) is compared to the critical F value (2.249), which was calculated by using $F_{0.005,\,5,\,257}$. After running the overall F Test, the null hypothesis was rejected, showing enough evidence to conclude that the linear regression model is significant.

Further, coefficient of determination, often denoted as R-squared, is used as a method to assess the goodness of fit of the regression model. In Figure 10, the "Multiple R-squared" is 0.4095, indicating that roughly 41% of the variability in "Salary" is explained by the predictors (CRBI, Years, CRuns, CAtBat, CHits) in the model. Additionally, adjusted R-squared is used as modification of the standard R-squared that helps address some of its limitations, especially when dealing with multiple independent variables in the regression model. Again, in Figure 10, the adjusted R-squared of the model is approximately 0.398. The value is relatively lower than the R-squared which suggests the inclusion of some predictors might not contribute significantly to the model. Since the difference between R-squared and adjusted R-squared is not huge, it suggests that the inclusion of predictors is not penalized heavily.

After establishing the model's overall significance through the F test, the analysis shifts towards a more detailed examination of individual predictor variables. Individual t-values for marginal effects are also shown in Figure 10. CRBI, CRuns, CAtBat, and CHits are variables with coefficients that are statistically significant because the critical value is 1.969 which is calculated by $t_{0.025,\,257}$. The comparison between t-value and critical value suggests that they are significant variables in predicting Salary. Meanwhile, absolute t-value (1.855) for the Years variable does not reach the critical value (1.969), suggesting it may not be as important in the model.

```
Call:
lm(formula = Salary ~ CRBI + Years + CRuns + CAtBat + CHits,
    data = hitters)

Residuals:
     Min       1Q   Median       3Q      Max
-1186.26  -237.31   -54.79   147.43  1787.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 380.8004    41.7573   9.119  < 2e-16 ***
CRBI          0.8219     0.2203   3.730 0.000236 ***
Years       -22.6366    12.1999  -1.855 0.064674 .
CRuns         0.8566     0.3958   2.164 0.031389 *
CAtBat       -0.4343     0.1172  -3.707 0.000257 ***
CHits         1.2376     0.3726   3.322 0.001024 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350 on 257 degrees of freedom
Multiple R-squared:  0.4095,    Adjusted R-squared:  0.398
F-statistic: 35.64 on 5 and 257 DF,  p-value: < 2.2e-16
```

Figure 10. Linear regression model using CRBI, Years, CRuns, CAtBat, and CHits variables to predict Salary

Next, the reliability of the model's predictions will be explored by checking the confidence and prediction intervals for each predictor variable. Figure 11 and Figure 12 provide valuable insights into the predicted salary for a hypothetical baseball player with specific career statistics. For a player with 186 career RBIs, 9 years of experience, 192 career runs, 1876 career at-bats, and 467 career hits, the estimated salary is approximately $2,576,498 (Figure 11). The accompanying 95% confidence interval for the mean response suggests that, with 95% confidence, the salary for a new individual with these career statistics will fall within the range of $1,717,714 to $3,441,282 (Figure 11). This interval provides a measure of precision for our estimate.

Unlike confidence intervals that are only concerned with the center of the population distribution, prediction intervals take into account the tails of the distribution as well as the center. It has greater sensitivity to the assumption of normality than do confidence intervals. In Figure 12, the prediction interval is being calculated for a new individual with the same statistics. The prediction interval shows a wider interval, spanning from approximately -$4,370,292 to $9,523,288 (Figure 12). This wider interval reflects the inherent variability in predicting individual salaries and emphasizes the potential influence of outliers or extreme cases in the data.

```
       fit      lwr      upr
1 257.6498 171.1714 344.1282
```

Figure 11. Confidence interval for mean response

```
       fit       lwr      upr
1 257.6498 -437.0292 952.3288
```

Figure 12. Prediction interval for new individual

The Variance Inflation Factor (VIF) values for the predictor variables in the model indicate the presence of multicollinearity. CRBI and Years exhibit moderate levels of multicollinearity with VIF values of 10.86 and 7.31, respectively (Figure 13). However, the variables CRuns, CAtBat, and CHits display substantial to extreme levels of multicollinearity, with VIF values of 36.76, 153.50, and 124.73, respectively (Figure 13). These exceedingly high VIF values suggest that CRuns, CAtBat, and CHits are highly correlated with other predictor variables in the model.

```
     CRBI      Years      CRuns     CAtBat      CHits
10.857492   7.313928  36.757190 153.495846 124.733001
```

Figure 13. VIF for CRBI, Years, CRuns, CAtBat, and CHits variables

Due to the presence of high VIF values, the process of standardization is implemented to address and mitigate the multicollinearity issue in the regression model. The newly fitted linear regression model includes standardized predictor variables (SCRBI, SYears, SCRuns, SCAtBat, and SCHits). The standardized coefficients allow for a direct comparison of the impact of each predictor on the response variable in terms of standard deviations. However, standardizing the predictors has not altered the interpretation of the model substantially (Figure 14). The persistently high VIF values for SCRuns, SCAtBat, and SCHits suggest that multicollinearity remains a concern.

SCRBI	SYears	SCRuns	SCAtBat	SCHits
10.857492	7.313928	36.757190	153.495846	124.733001

Figure 14. VIF for standardized CRBI, Years, CRuns, CAtBat, and CHits variables

In order for the analysis of the multiple linear regression to be valid, several assumptions need to be met. One assumption of multiple linear regression is that the population errors follow a normal distribution. If this assumption is violated, all the inference procedures will be invalid unless normality is achieved through other means, such as the Central Limit Theorem. From the histogram of the residuals and the normal probability plot of the residuals, it can be seen that the assumption of normality for the final model is violated (Figure 15). If the data were normal, the histogram would have a bell-shaped curve and the data points in the normality plot would form a straight line, both of which would indicate normality; however, this is not what is seen. Instead, the histogram appears to be slightly skewed to the right, and the data in the probability plot have strong curvature, which happens when data is from a left or right-skewed population, which implies a non-normal distribution (Figure 15). Recentering the data will help fix issues with

normality in order to meet the assumption of normal data.

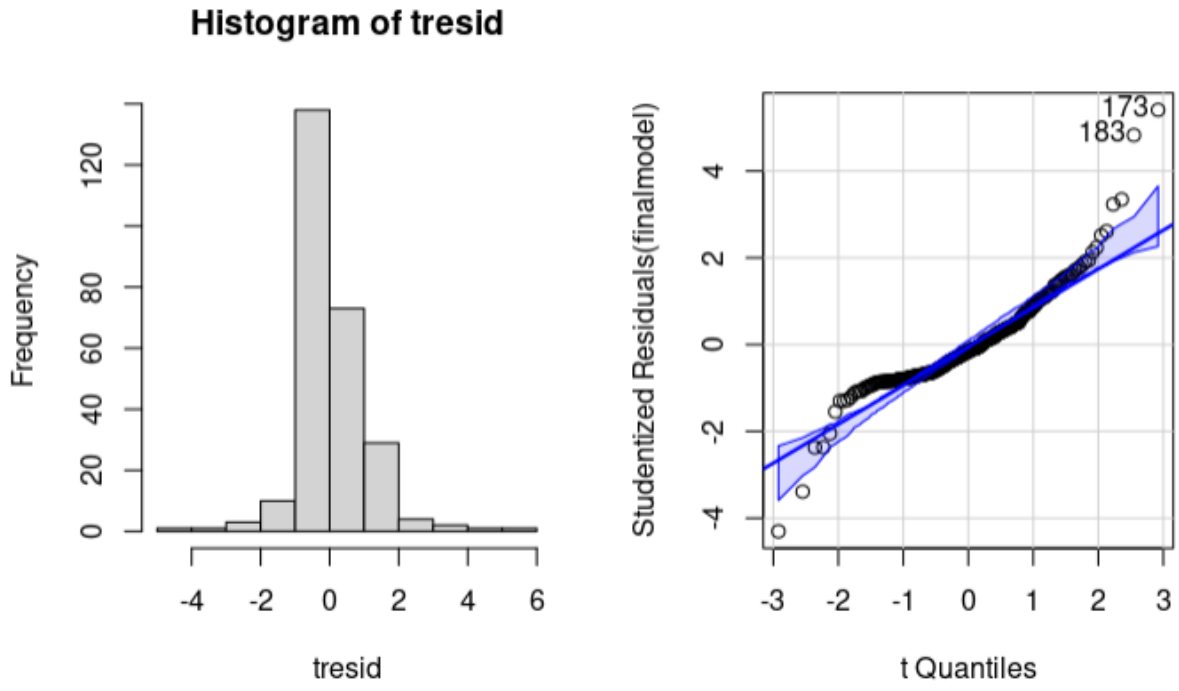

Figure 15. Histogram of Residuals and Probability Plot of Residuals

After standardizing the data and running the final model again, the histogram more closely resembles that of a normal distribution, and the data in the probability distribution has less curvature and is closer to the straight line that is expected when the assumption of normality is met (Figure 16).

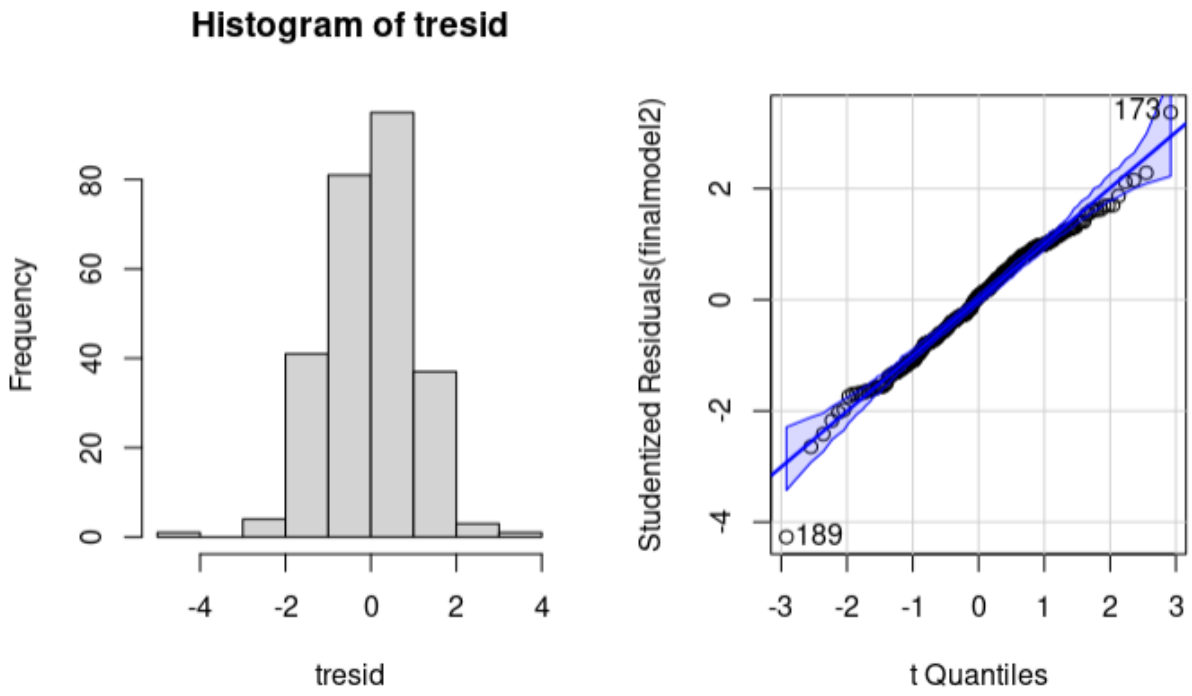

Figure 16. Histogram of Residuals and Normal Probability Plot of Residuals

Another assumption is that the mean of the population errors is zero. In other words, the relationship is linear in the coefficients, making the model correct. If this assumption is violated, the predictors could be biased, and the model may not be a good fit for the data. An additional assumption of multiple linear regression is that the standard deviation of the population errors is constant, while the last assumption of linear regression is that the population errors are independent of each other. If these two assumptions are violated, the estimate of population standard deviation will be biased, all of the inference procedures will be invalid, and the coefficient estimates will no longer be minimum variance estimates. The residuals in the versus fit plot have a fan-shaped effect, which suggests the variation increases with the response (Figure 17). This effect can occur with right-skewed data, which, as previously stated above, is present with the original final model and is a clear indicator of non-constant variation. To fix this, a

natural log transformation can be done on the response variable, salary. While the residuals are fan-shaped in regards to variation, the residuals are centered around zero, indicating linearity. There are a lot of residuals in the versus order plot, making it very cluttered and difficult to determine if the population errors are independent of each other. A Durbin-Watson test was used to check for autocorrelation. The Durbin-Watson test for positive autocorrelation returned a Durbin-Watson statistic of 1.958506 or a p-value of 0.341, which is not significant, indicating no presence of positive autocorrelation (Figure 18). Additionally, a Durbin-Watson test for negative autocorrelation came back with a Durbin-Watson statistic of 1.958506, or a p-value of 0.626, which is not significant, signifying that there is no existence of negative autocorrelation within the model (Figure 19).

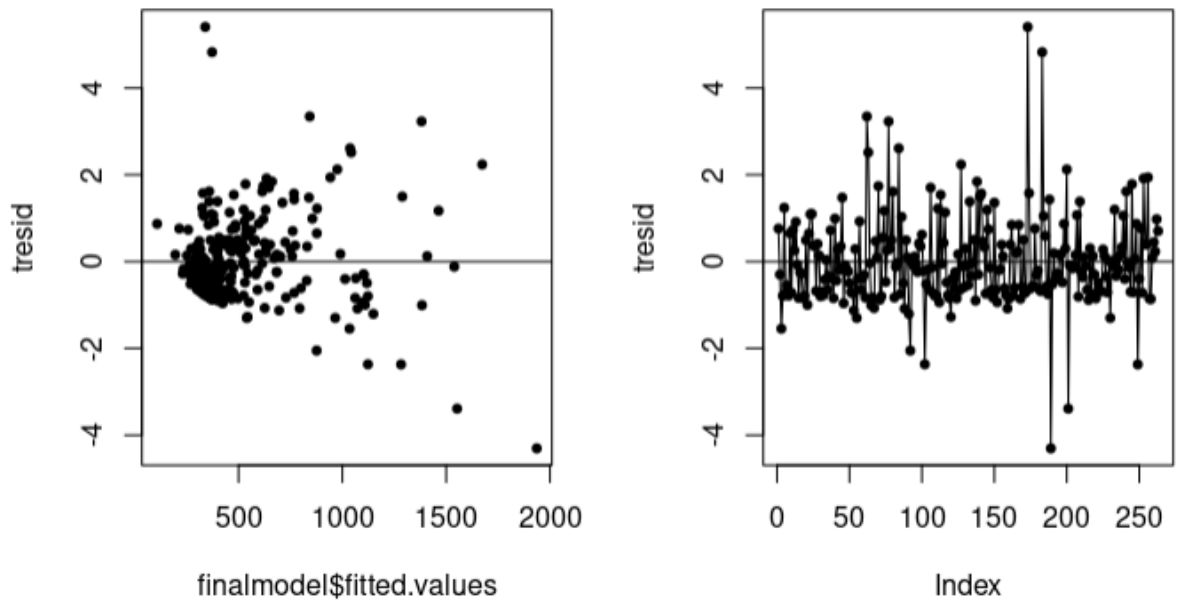

Figure 17. Versus Fits and Versus Order Plots

```
lag Autocorrelation D-W Statistic p-value
  1      0.01974357      1.958506   0.341
Alternative hypothesis: rho > 0
```

Figure 18. Durbin-Watson Test for Positive Autocorrelation

```
lag Autocorrelation D-W Statistic p-value
  1      0.01974357      1.958506   0.626
Alternative hypothesis: rho < 0
```

Figure 19. Durbin-Watson Test for Negative Autocorrelation

Taking the natural log of the response variable, salary, helps to resolve issues with the assumption of constant variation. After performing the transformation and checking for constant variation, the variation appears to be more constant. While there is still a slight curvature of the residuals in the versus fit plot, the data appears to be more consistent than before transforming the response and is therefore an improvement of the final model in regards to variation (Figure 20). Although the transformation of the response improves the variation of the model, it negatively impacts the linearity of the model. The data appears to be less centered around zero than before the transformation took place, making the model less linear than before. One method of fixing this would be to take the natural logs of the predictor variables to help fix this issue with linearity. The versus fits graph is still chaotic, but the Durbin-Watson test confirmed there is no issue with autocorrelation, which reaffirms the assumption that the population errors are independent of one another.

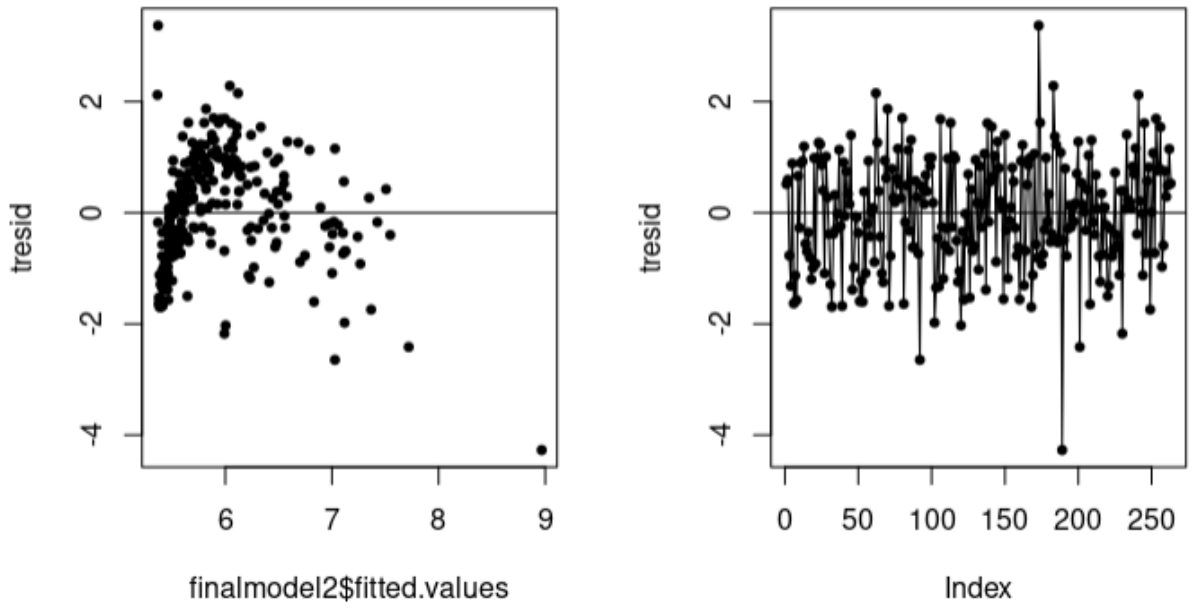

Figure 20. Versus Fits and Versus Order Plots

When checking for unusual points, a vast amount of outliers, leverage points, and influential data points were discovered. Figure 21 shows all the residuals plotted in order. Data points above or below the blue line represent mild outliers, while data points higher or lower than the red line represent moderate outliers. Looking at the plot of the studentized residuals, there appears to be five mild outliers and three moderate outliers.

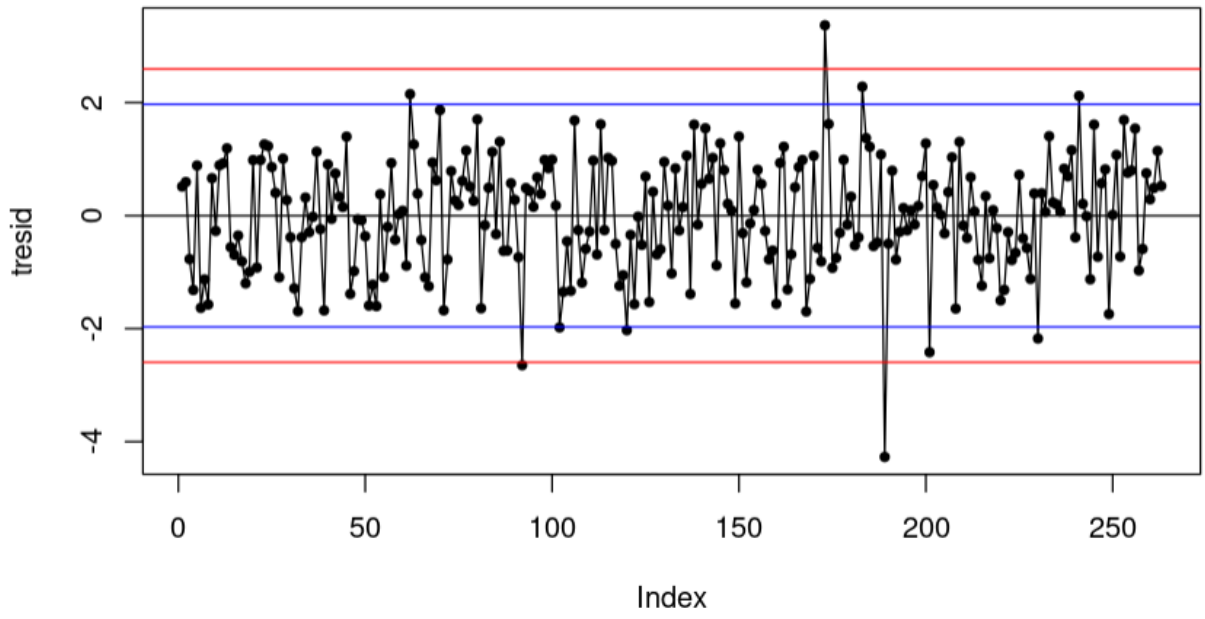

Figure 21. Studentized Residual Plot

When checking for leverages, it can be seen that there are approximately thirty-five points that are considered leverages (Figure 22). All these points have a leverage value greater than ~0.045 and are therefore found to be leverage points.

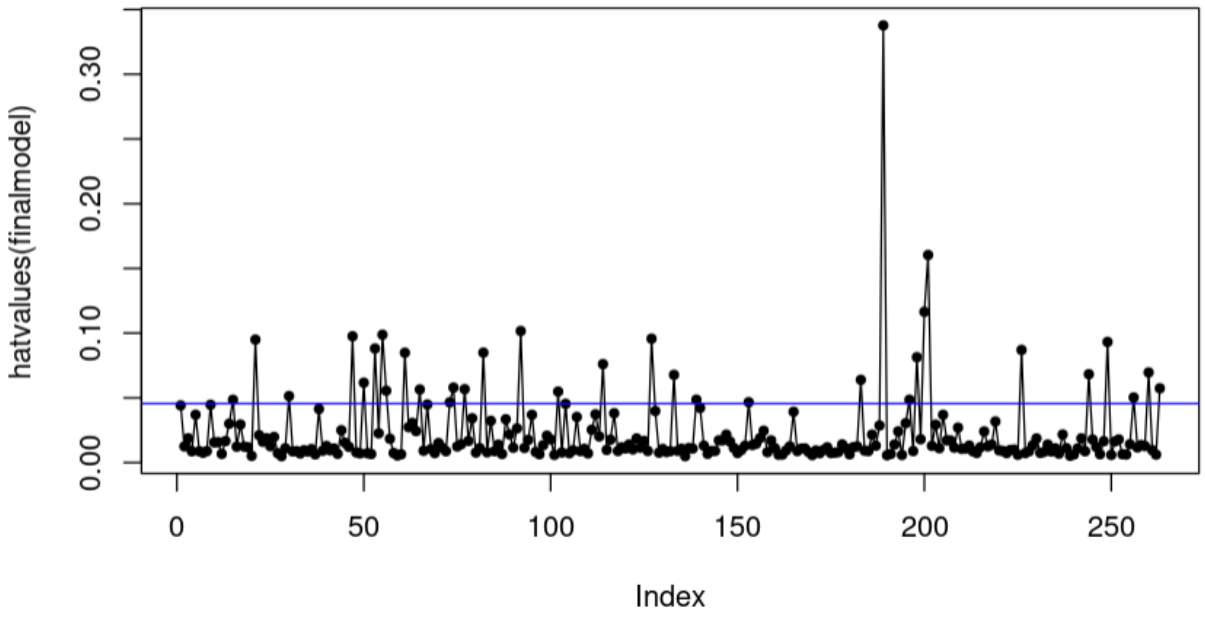

Figure 22. Leverage Plot

When looking for influential points, Cook's Distance found only one point with a value greater than the cutoff of 1, but it was significantly greater (Figure 23).

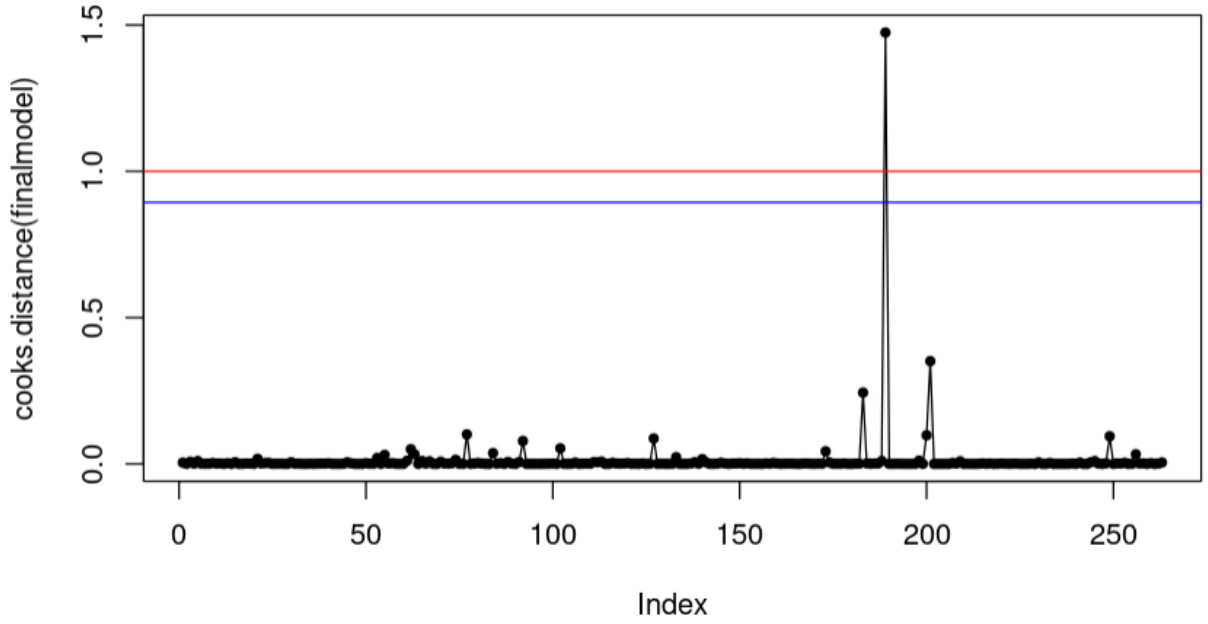

Figure 23. Cook's Distance Plot

The DFBETAs returned approximately fourteen influential points with one very influential point seen in blue at the top of the plot (Figure 24).

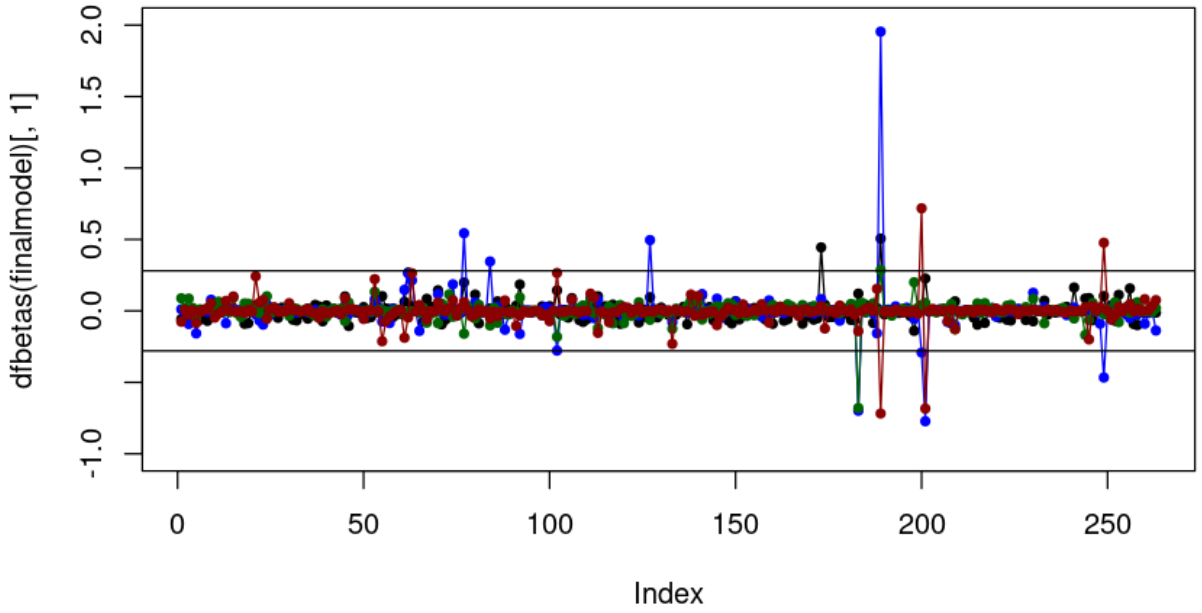

Figure 24. Plot of DFBETAs

DFFITs found thirteen influential points, two of which were highly influential (Figure 25).

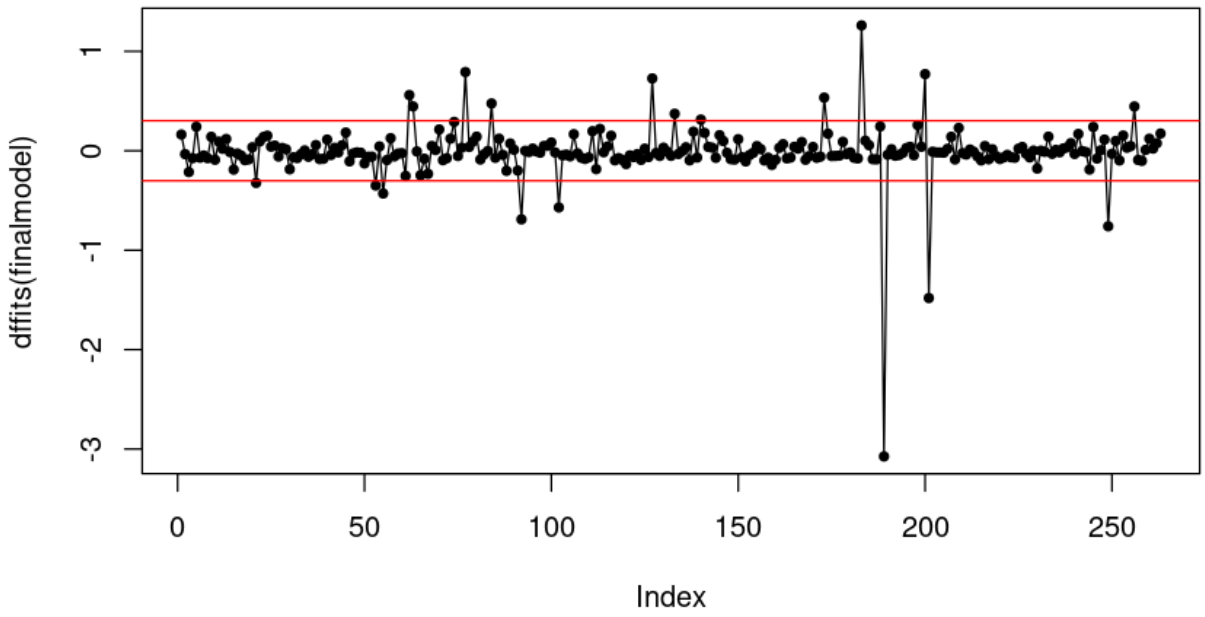

Figure 25. Plot of DFFITs

Although outliers, leverages, and influential points were found in the data, outliers and influential points should not automatically be removed as these points may contain valuable information. Instead, checking the data for errors, collecting more data in the x-range, collecting other explanatory variables, and running the model with and without these unusual points to determine the effect of the points should be done before deciding to remove any of the data points.

Conclusion

The objective of this paper was to predict the opening day annual salary of major league baseball players based on their career statistics, specifically Years, CAtBat, CHits, CHmRun, CRuns, CRBI, and CWalks. Various model selection techniques, including Mallow's C_p, backward elimination, forward selection, and stepwise selection, were employed to identify the "best" model, consistently highlighting the significance of CRBI, Years, CRuns, CAtBat, and

CHits. The analysis utilized statistical tools such as the X^TX matrix to assess multicollinearity, the overall F test to evaluate the model's significance, and R-squared and adjusted R-squared to measure goodness of fit. Individual t-values provided insights into the significance of each predictor variable, while confidence intervals for mean response and prediction intervals for new individuals measured the reliability of the model's predictions.

However, the presence of multicollinearity, outliers, and influential points prompted additional check and the implementation of corrective measures. The transformation of the response variable, natural log transformations, and standardization were employed to address issues related to normality, constant variation, and linearity. Despite these efforts, challenges persisted, and the identification of influential points highlighted the need for cautious interpretation. After applying various model and variable selection technique, the final model, represented by Salary = 380.8004 + 0.8219CRBI - 22.6366Years + 0.8566CRuns - 0.4343CAtBat + 1.2376CHits, offers valuable insights into the factors influencing the opening day annual salary of major league baseball players.

Works Cited

Floser. (n.d.). *Hitters*, Kaggle Dataset. Retrieved December 6, 2023 from https://www.kaggle.com/datasets/floser/hitters.

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York.

Appendix

Data Source:

This dataset was originally taken from the StatLib library which is maintained at Carnegie Mellon University. This is part of the data that was used in the 1988 ASA Graphics Section Poster Session. The salary data were originally from Sports Illustrated, April 20, 1987. The 1986 and career statistics were obtained from The 1987 Baseball Encyclopedia Update published by Collier Books, Macmillan Publishing Company, New York.

Data Dictionary:

A data frame with 322 observations of major league players on the following 20 variables.

Predictor Variables:

AtBat Number of times at bat in 1986

Hits Number of hits in 1986

HmRun Number of home runs in 1986

Runs Number of runs in 1986

RBI Number of runs batted in in 1986

Walks Number of walks in 1986

Years Number of years in the major leagues

CAtBat Number of times at bat during his career

CHits Number of hits during his career

CHmRun Number of home runs during his career

CRuns Number of runs during his career

CRBI Number of runs batted in during his career

CWalks Number of walks during his career

League A factor with levels A and N indicating player's league at the end of 1986

Division A factor with levels E and W indicating player's division at the end of 1986

PutOuts Number of put outs in 1986

Assists Number of assists in 1986

Errors Number of errors in 1986

NewLeague A factor with levels A and N indicating player's league at the beginning of 1987

Response Variable: Salary 1987 annual salary on opening day in thousands of dollars