The City University of New York | School of Professional Studies

DATA621

Homework #1

9/25/2022

Critical Thinking Group 2

William Aiken
Donald Butler
Michael Ippolito
Bharani Nittala
Leticia Salazar
Santiago Torres

For this homework assignment, the goal is to explore, analyze, and model a data set of records representing statistics from baseball teams from the 1871 to 2006. Each record contains the performance statistics of one team for a given year, adjusted for a 162-game season. The data contains a training set and an evaluation set (. We built a multiple linear regression model on the training set to predict the number of wins for the team. Based on the resulting model, we predicted the number of wins for each observation in the evaluation set.

1. DATA EXPLORATION

The data set contains a training set (2,276 observations) and an evaluation set (259 observations). Table 1.1 lists the variables included in the data set, a definition for each variable, and the theoretical effect of the variable on the number of wins the team had that season.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

Table 1.1 Variable names and definitions.

Due to the number of fields in this data, we broke the dataset into intuitive sections and explored each section individually.

1.1 Base Hits by Batter

The following fields comprise the base hits by batter:

1. TARGET_WINS - Number of wins
2. TEAM_BATTING_H - Base hits by batters (1B,2B,3B,HR)
3. TEAM_BATTING_2B - Doubles by batters (2B)
4. TEAM_BATTING_3B - Triples by batters (3B)
5. TEAM_BATTING_HR - Homeruns by batters (4B)

The means and medians are very similar for the base hit variables, implying little skew to the distributions (Table 1.2). R code producing this and all other tables and figures herein is included as Appendix A.

Characteristic	**N = 2,276**[1]
TARGET_WINS	81 82 16
TEAM_BATTING_H	1,469 1,454 145
TEAM_BATTING_2B	241 238 47
TEAM_BATTING_3B	55 47 28
TEAM_BATTING_HR	100 102 61

[1] Mean Median SD

Table 1.2 Means, medians, and standard deviations of base hits by batters.

As shown, we see tight distributions except for all base hits by batters (TEAM_BATTING_H).

Unsurprisingly, all possible base hits (TEAM_BATTING_H) is correlated with winning. As the number of bases achieved by an at-bat increases, the correlation decreases (Figure 1.1).

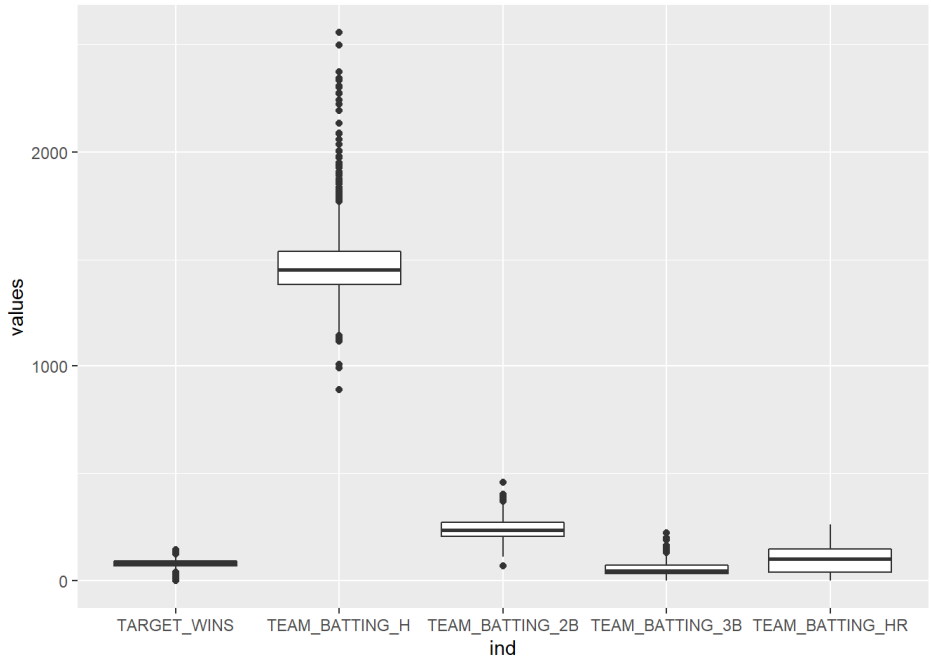

Figure 1.1 Box plots of base hit variables.

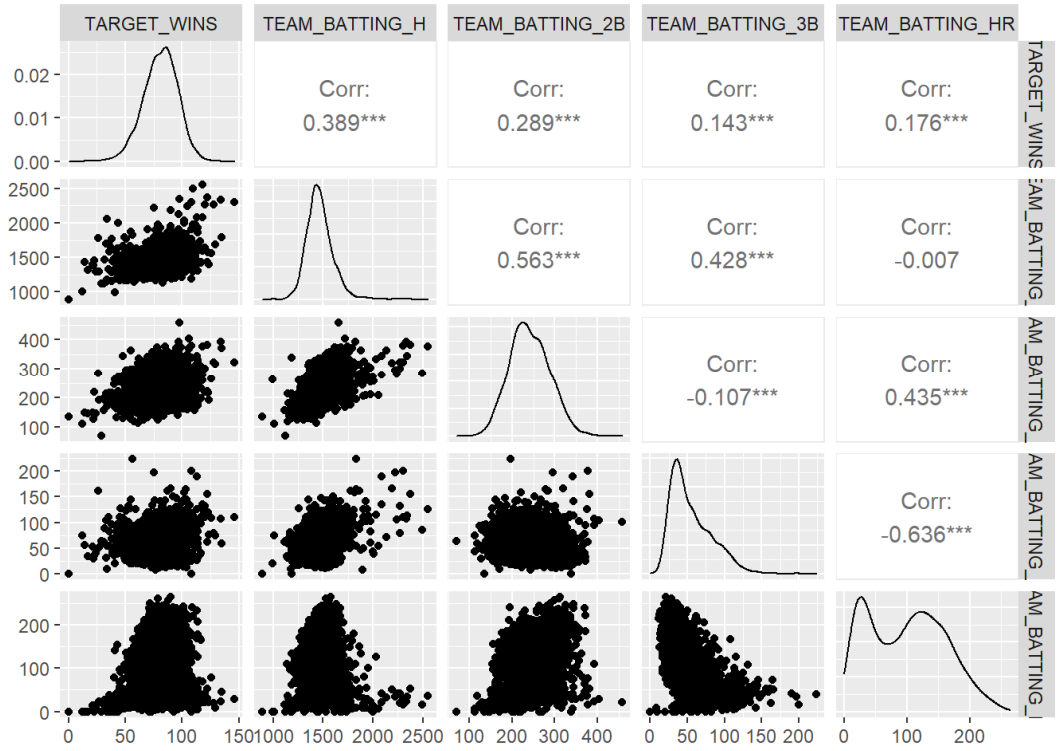

Figure 1.2 Correlation plot of base hit variables.

Interestingly, doubles and triples are correlated with base hits while home runs are not, as shown in Figure 1.2.

1.2 Batting

Next, variables associated with batting were investigated:

1. TARGET_WINS - Number of wins
2. TEAM_BATTING_BB - Walks by batters
3. TEAM_BATTING_HBP - Batters hit by pitch (get a free base)
4. TEAM_BATTING_SO - Strikeouts by batters
5. TEAM_BASERUN_SB - Stolen bases
6. TEAM_BASERUN_CS - Caught stealing

The measures of central tendency show us that most of these variables have slight skew to their distributions. The stolen bases variable has a large right skew to its distribution (Table 1.3 and Figure 1.3). We are missing values for strikeouts, stolen bases, and caught stealing.

Characteristic	N = 2,276[1]
TEAM_BATTING_BB	502 512 123
TEAM_BATTING_SO	736 750 249
Unknown	102
TEAM_BASERUN_SB	125 101 88
Unknown	131
TEAM_BASERUN_CS	53 49 23
Unknown	772

[1] Mean Median SD

Table 1.3 Means, medians, and standard deviations of batting variables.

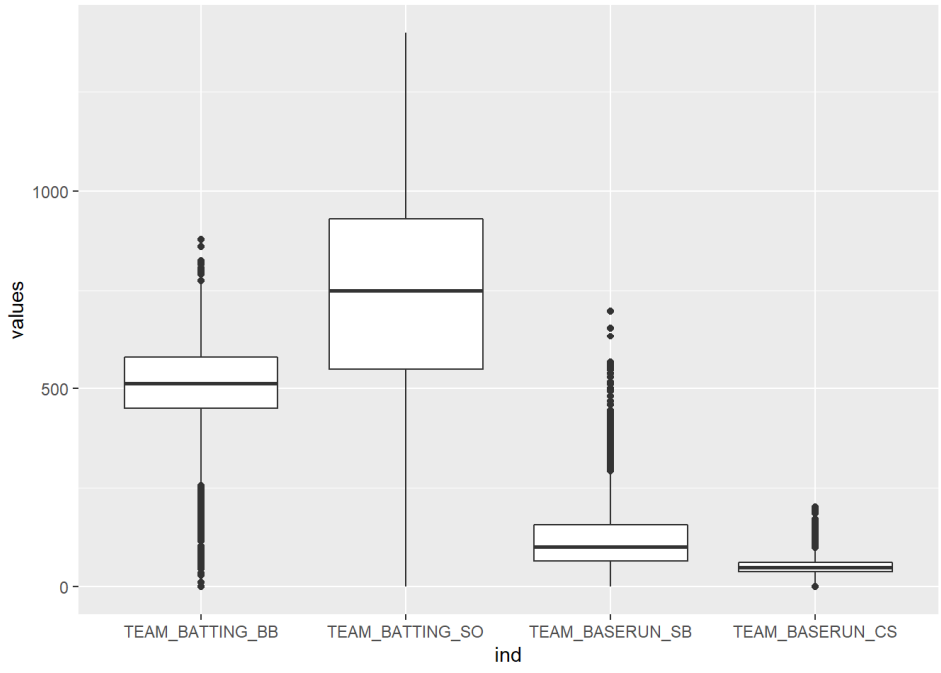

Figure 1.3 Box plots of batting variables.

Of all the batting variables, only walks by batter has a correlation to wins (Figure 1.4).

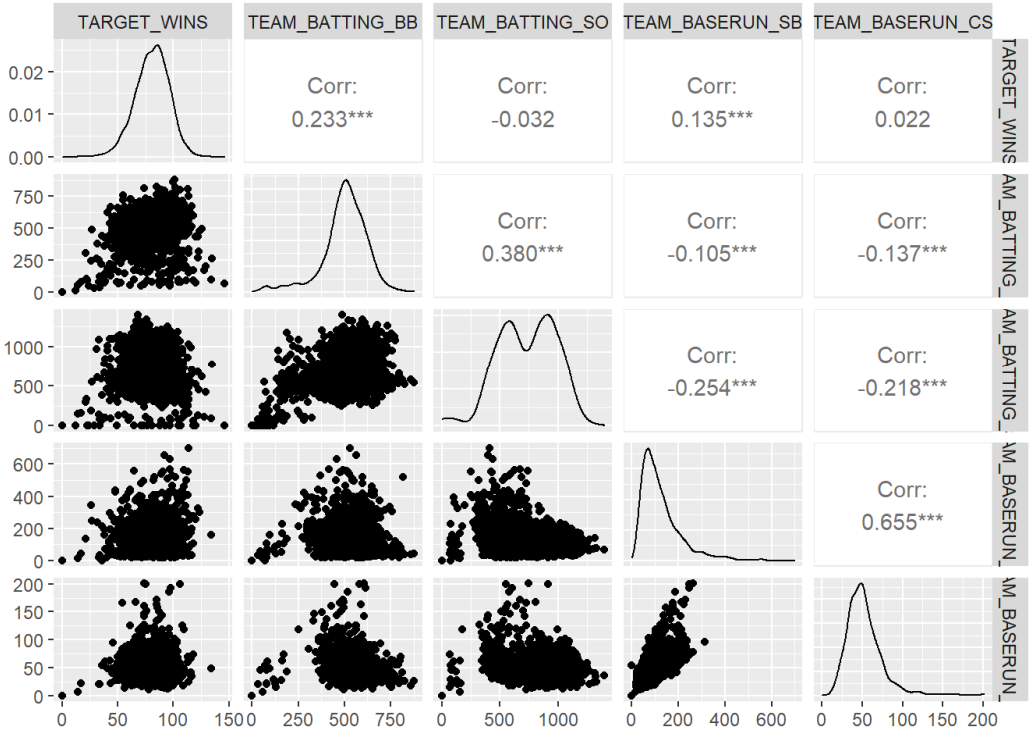

Figure 1.4 Correlation plot of batting variables.

1.3 Fielding

1. TARGET_WINS - Number of wins
2. TEAM_FIELDING_E - Errors
3. TEAM_FIELDING_DP - Double Plays

The errors variable (TEAM_FIELDING_E) has an incredibly right-skewed distribution (Table 1.4 and Figure 1.5). It is also noted that some double plays values are missing.

Characteristic	**N = 2,276**[1]
TEAM_FIELDING_E	246 159 228
TEAM_FIELDING_DP	146 149 26
Unknown	286

[1] Mean Median SD

Table 1.4 Means, medians, and standard deviations of fielding variables.

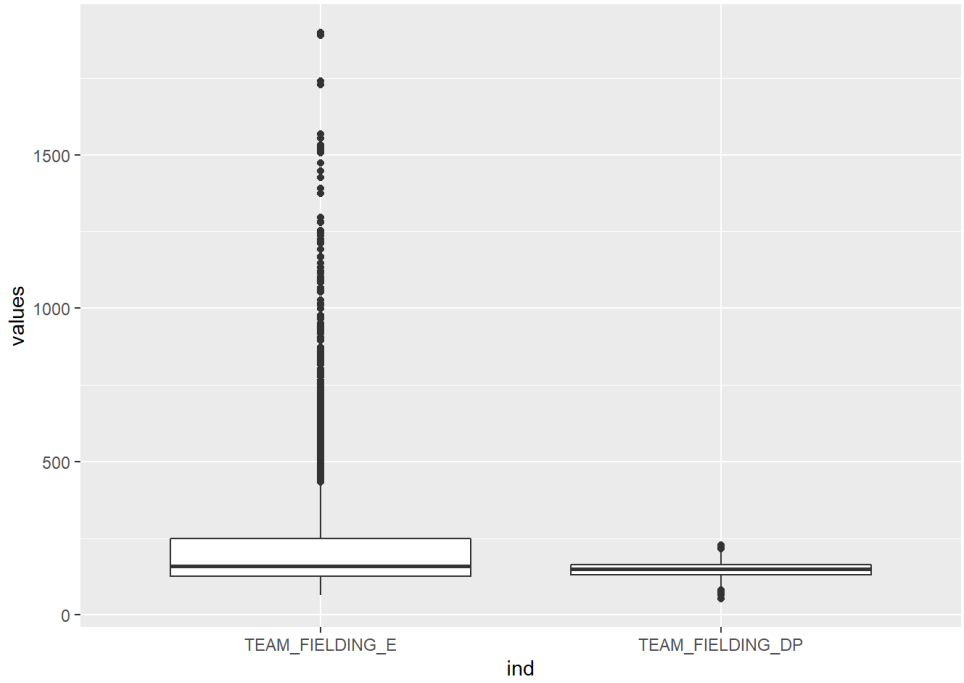

Figure 1.5 Box plots of fielding variables.

Both fielding variables are negatively correlated with wins (Figure 1.6)

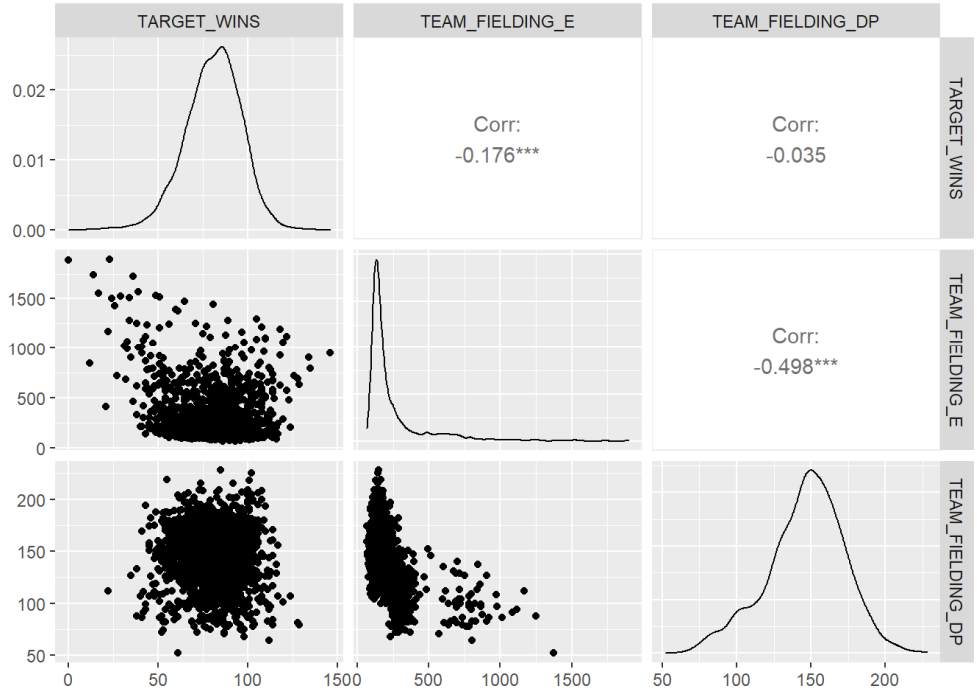

Figure 1.6 Correlation plot of batting variables.

1.4 Pitching

1. TARGET_WINS - Number of wins

2. TEAM_PITCHING_BB - Walks allowed
3. TEAM_PITCHING_H - Hits allowed
4. TEAM_PITCHING_HR - Homeruns allowed
5. TEAM_PITCHING_SO - Strikeouts by pitchers

As shown in Table 1.5 and Figure 1.7, hits allowed (TEAM_PITCHING_H) has a right skew, and some values are missing for strikeouts by pitcher (TEAM_PITCHING_SO).

Characteristic	N = 2,276[1]
TEAM_PITCHING_H	1,779 1,518 1,407
TEAM_PITCHING_HR	106 107 61
TEAM_PITCHING_BB	553 536 166
TEAM_PITCHING_SO	818 814 553
Unknown	102

[1] Mean Median SD

Table 1.5 Means, medians, and standard deviations of pitching variables.

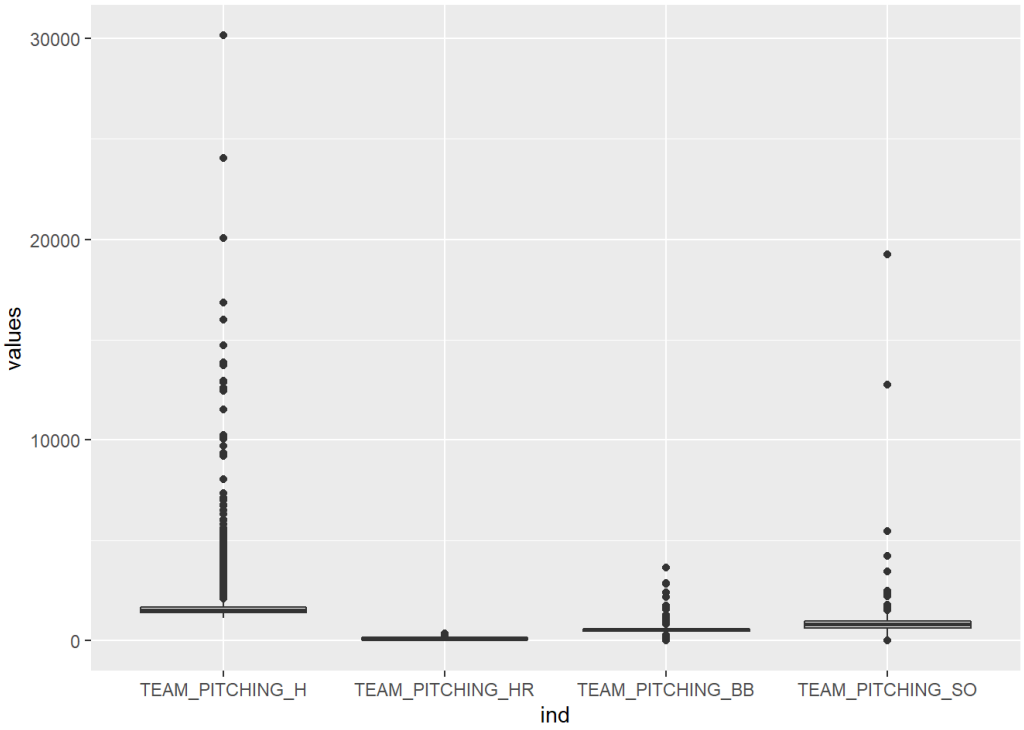

Figure 1.7 Box plots of pitching variables.

The number of hits allowed by pitchers (TEAM_PITCHING_H) is negatively correlated with winning (Figure 1.8). Counter-intuitively, home runs allowed is positively correlated with Winning.

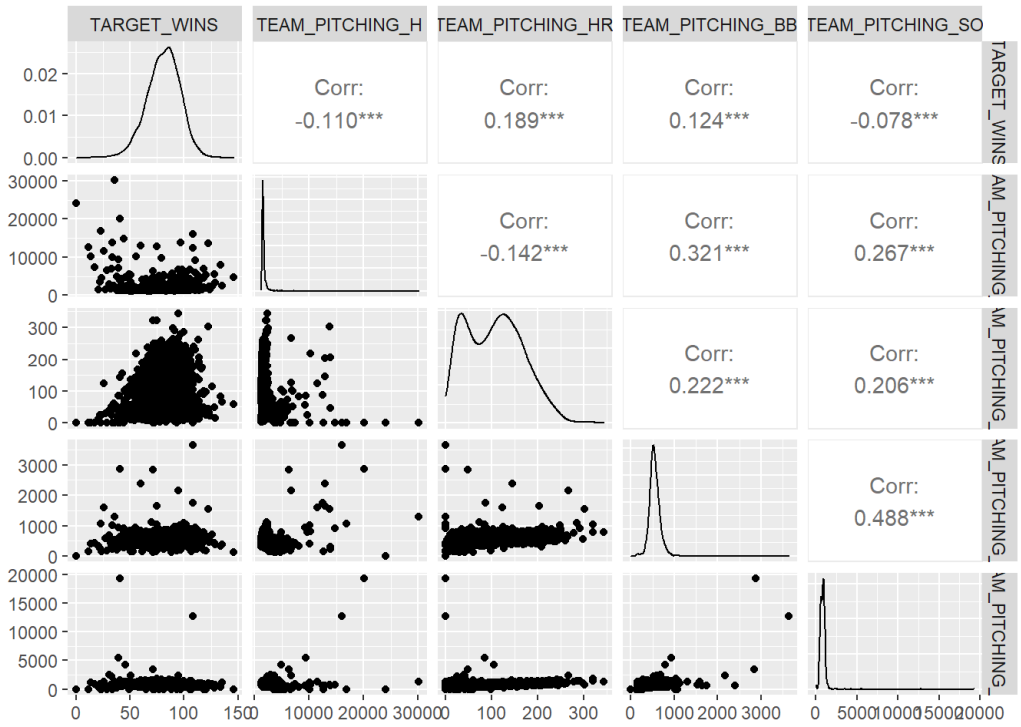

Figure 1.8 Correlation plot of pitching variables.

1.5 Overall Observations

As shown in the correlation plot for all variables (Figure 1.9), there appear to be several strong correlations between explanatory variables and the target. From an initial inspection, it appears the team should focus on getting players on base through hits or walks. Teams can still win if the pitchers allow homeruns, hits, and walks to the other team.

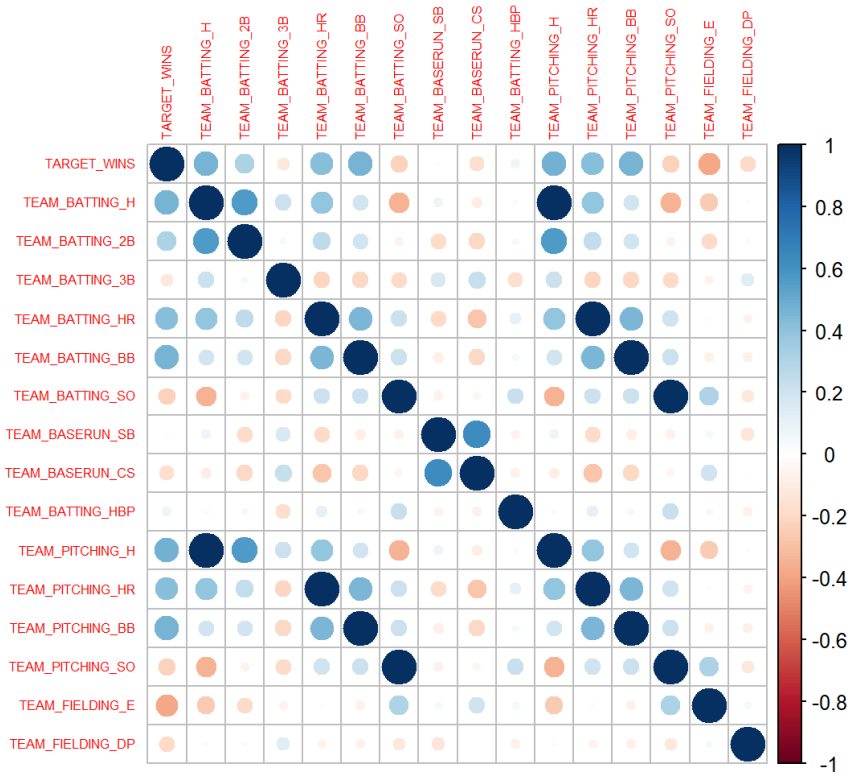

Figure 1.9 Correlation plot of all variables.

The variables with positive correlation to TARGET_WINS are shown in Table 1.6. Hits by batters tops the list at 0.389.

Variable	Correlation to TARGET_WINS
TEAM_BATTING_H	0.389
TEAM_BATTING_2B	0.289
TEAM_BATTING_BB	0.233
TEAM_PITCHING_HR	0.189
TEAM_BATTING_HR	0.176
TEAM_BATTING_3B	0.143
TEAM_PITCHING_BB	0.124

Table 1.6 Variables positively correlated with TARGET_WINS.

Table 1.7 shows variables that are negatively correlated with TARGET_WINS. To win more games, it makes sense the team will need to make fewer errors and yield fewer hits to the opposing team.

Variable	Correlation to TARGET_WINS
TEAM_FIELDING_E	-0.176
TEAM_PITCHING_H	-0.110

Table 1.7 Variables negatively correlated with TARGET_WINS.

2. DATA PREPARATION

Table 2.1 shows the percentages of missing values for each variable. As shown, there are two variables missing many observations: TEAM_BATTING_HBP is missing over 90% of its values, while TEAM_BASERUN_CS is missing just around 30%.

	x
INDEX	0.00
TARGET_WINS	0.00
TEAM_BATTING_H	0.00
TEAM_BATTING_2B	0.00
TEAM_BATTING_3B	0.00
TEAM_BATTING_HR	0.00
TEAM_BATTING_BB	0.00
TEAM_BATTING_SO	4.48
TEAM_BASERUN_SB	5.76
TEAM_BASERUN_CS	33.92
TEAM_BATTING_HBP	91.61
TEAM_PITCHING_H	0.00
TEAM_PITCHING_HR	0.00
TEAM_PITCHING_BB	0.00
TEAM_PITCHING_SO	4.48
TEAM_FIELDING_E	0.00
TEAM_FIELDING_DP	12.57

Table 2.1 Percentages of observations with missing values.

Missing values were imputed using R's multiple imputation by chained equations (mice) package. Imputation was accomplished using predictive mean matching with five multiple imputations over a maximum of five iterations. Figure 2.1 shows the variable summary after imputation, while Figure 2.2 illustrates this information with histograms.

```
     INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB
Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0
1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0
Median :1270.5   Median : 82.00   Median :1454   Median :238.0   Median : 47.00   Median :102.00   Median :512.0
Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25   Mean   : 99.61   Mean   :501.6
3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0
Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00   Max.   :264.00   Max.   :878.0
TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
Min.   :   0.0   Min.   :  0.0   Min.   : 1137   Min.   :  0.0    Min.   :   0.0   Min.   :    0.0   Min.   :  65.0
1st Qu.: 542.0   1st Qu.: 67.0   1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  611.0   1st Qu.: 127.0
Median : 732.0   Median :105.0   Median : 1518   Median :107.0    Median : 536.5   Median :  802.5   Median : 159.0
Mean   : 727.5   Mean   :134.6   Mean   : 1779   Mean   :105.7    Mean   : 553.0   Mean   :  810.3   Mean   : 246.5
3rd Qu.: 925.0   3rd Qu.:169.0   3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  957.2   3rd Qu.: 249.2
Max.   :1399.0   Max.   :697.0   Max.   :30132   Max.   :343.0    Max.   :3645.0   Max.   :19278.0   Max.   :1898.0
```

Figure 2.1 Variable summary after imputation.

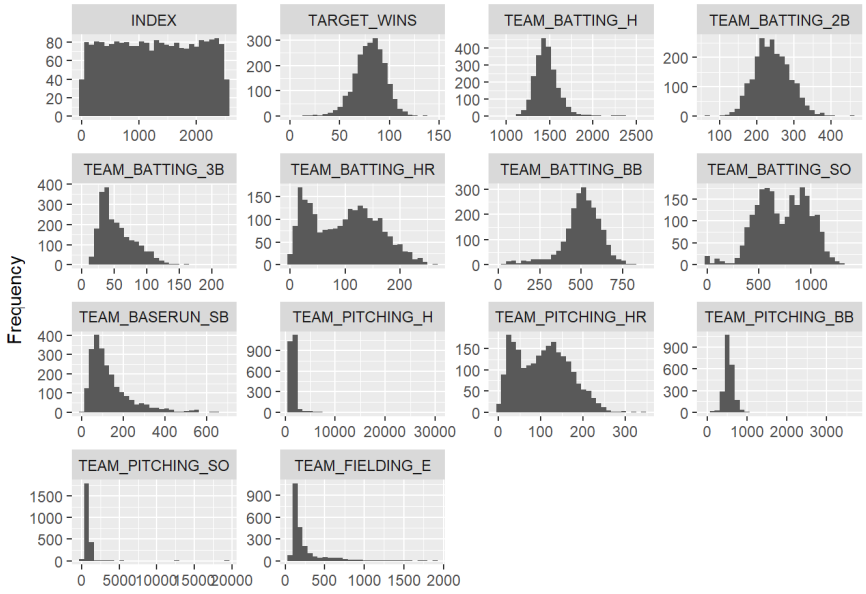

Figure 2.2 Histograms of variables after imputation.

It was noted that following imputation, several variables exhibited impossibly high values, for example TEAM_PITCHING_H, which included values as high as 30,132. To correct for this, any imputed value exceeding three standard deviations was modified to be the median value for that variable. Figure 2.3 illustrates the new variable summary following this cleanup process.

```
     INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B  TEAM_BATTING_HR  TEAM_BATTING_BB
 Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0
 1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00   1st Qu.: 42.00   1st Qu.:451.0
 Median :1270.5   Median : 82.00   Median :1454   Median :238.0   Median : 47.00   Median :102.00   Median :512.0
 Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25   Mean   : 99.61   Mean   :501.6
 3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00   3rd Qu.:147.00   3rd Qu.:580.0
 Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00   Max.   :264.00   Max.   :878.0
 TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
 Min.   :   0.0   Min.   :  0.0   Min.   :1137   Min.   :  0.0    Min.   :  0.0    Min.   :   0.0   Min.   : 65.0
 1st Qu.: 542.0   1st Qu.: 67.0   1st Qu.:1419   1st Qu.: 50.0    1st Qu.:476.0    1st Qu.: 611.0   1st Qu.:127.0
 Median : 732.0   Median :105.0   Median :1518   Median :107.0    Median :536.5    Median : 802.2   Median :159.0
 Mean   : 727.5   Mean   :134.6   Mean   :1605   Mean   :105.7    Mean   :500.4    Mean   : 788.3   Mean   :198.9
 3rd Qu.: 925.0   3rd Qu.:169.0   3rd Qu.:1660   3rd Qu.:150.0    3rd Qu.:536.5    3rd Qu.: 954.0   3rd Qu.:215.0
 Max.   :1399.0   Max.   :697.0   Max.   :4134   Max.   :343.0    Max.   :536.5    Max.   :1600.0   Max.   :681.0
```

Figure 2.3 Variable summary after cleanup.

3. MODEL BUILDING

After imputing and cleaning the data, the process of model building was begun by fitting wins to the full training data set.

3.1 Full Model

By testing all variables in this first model, we are able to see how significant the variables are in our dataset. We will then be able to use this model to serve as a basis for our other models. Model results are given in Table 3.1, and residual plots are illustrated in Figure 3.1.

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(13,2262)	69.33
R^2	0.28
Adj. R^2	0.28

	Est.	S.E.	t val.	p
(Intercept)	9.71	5.40	1.80	0.07
INDEX	-0.00	0.00	-1.41	0.16
TEAM_BATTING_H	0.03	0.00	9.14	0.00
TEAM_BATTING_2B	-0.01	0.01	-0.95	0.34
TEAM_BATTING_3B	0.09	0.02	5.33	0.00
TEAM_BATTING_HR	0.09	0.03	3.51	0.00
TEAM_BATTING_BB	0.04	0.00	9.18	0.00
TEAM_BATTING_SO	0.00	0.00	0.76	0.45
TEAM_BASERUN_SB	0.04	0.00	10.42	0.00
TEAM_PITCHING_H	0.00	0.00	2.01	0.04
TEAM_PITCHING_HR	-0.04	0.02	-1.56	0.12
TEAM_PITCHING_BB	-0.02	0.01	-2.37	0.02
TEAM_PITCHING_SO	-0.01	0.00	-1.86	0.06
TEAM_FIELDING_E	-0.02	0.00	-7.13	0.00

Standard errors: OLS

Table 3.1 Full model results.

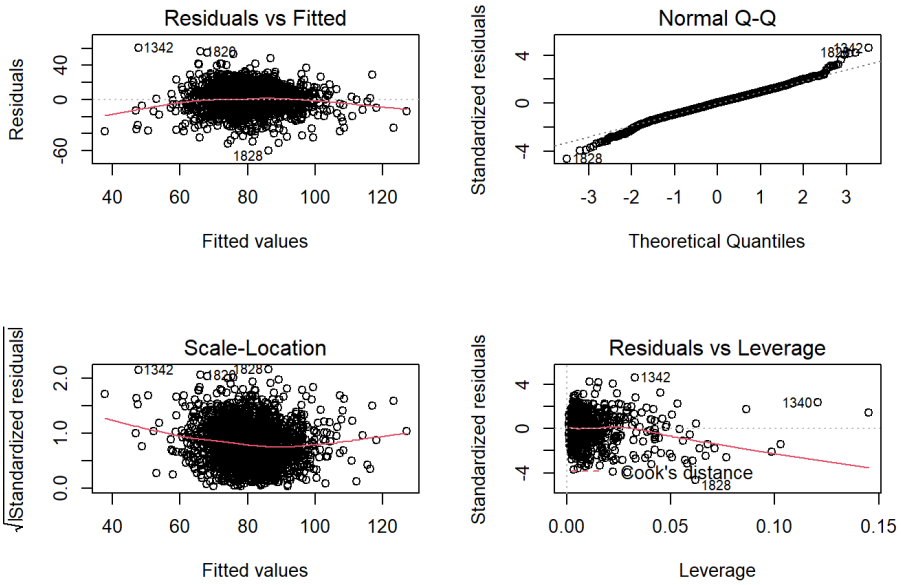

Figure 3.1. Full model residual plots.

3.2 Log Transformation

As shown in Table 3.2, many variables were skewed. To correct for this, we attempted a log transformation, which distributes skewness into a more "normally" distributed shape. We applied log transformations on highly skewed variables (less than -1 or greater than 1).

	x
INDEX	0.0042149
TARGET_WINS	-0.3987232
TEAM_BATTING_H	1.5713335
TEAM_BATTING_2B	0.2151018
TEAM_BATTING_3B	1.1094652
TEAM_BATTING_HR	0.1860421
TEAM_BATTING_BB	-1.0257599
TEAM_BATTING_SO	-0.2241493
TEAM_BASERUN_SB	1.8612344
TEAM_PITCHING_H	3.3490180
TEAM_PITCHING_HR	0.2877877
TEAM_PITCHING_BB	-2.4931634
TEAM_PITCHING_SO	-0.0285753
TEAM_FIELDING_E	2.1084199

Table 3.2. Variable skewness.

It is noted that this model was not successful compared to the first model. There weren't any significant changes between the two models, and we therefore discarded the second model. The results of the log-transformed model are given in Table 3.3. Residuals are plotted in Figure 3.2.

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(13,2262)	69.88
R^2	0.29
Adj. R^2	0.28

	Est.	S.E.	t val.	p
(Intercept)	-282.86	37.55	-7.53	0.00
INDEX	-0.00	0.00	-1.56	0.12
TEAM_BATTING_H	108.88	13.73	7.93	0.00
TEAM_BATTING_2B	-5.19	5.25	-0.99	0.32
TEAM_BATTING_3B	0.11	0.02	6.16	0.00
TEAM_BATTING_HR	0.10	0.03	3.85	0.00
TEAM_BATTING_BB	0.03	0.00	8.65	0.00
TEAM_BATTING_SO	-0.00	0.00	-0.77	0.44
TEAM_BASERUN_SB	13.65	1.24	11.04	0.00
TEAM_PITCHING_H	8.38	5.80	1.44	0.15
TEAM_PITCHING_HR	-0.04	0.02	-1.80	0.07
TEAM_PITCHING_BB	-2.25	4.27	-0.53	0.60
TEAM_PITCHING_SO	-0.00	0.00	-1.25	0.21
TEAM_FIELDING_E	-17.05	2.31	-7.40	0.00

Standard errors: OLS

Table 3.3. Log-transformed model results.

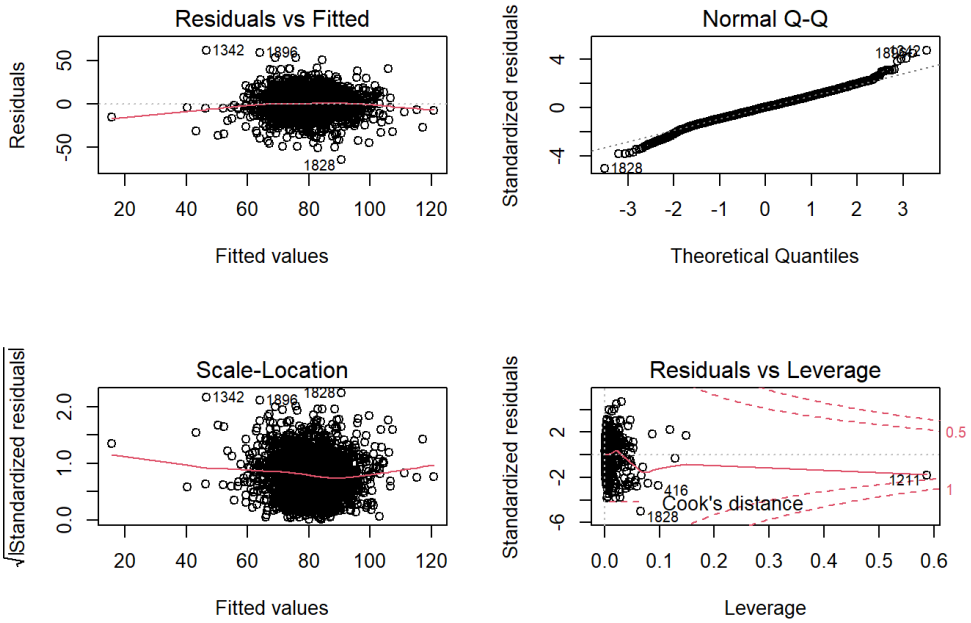

Figure 3.2. Log-transformed model residual plots.

3.3 Statistically Significant Variables

For the third model, we focused on statistically significant variables, chosen primarily from their R output during the data exploration phase. Table 3.4

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(6,2269)	143.71
R²	0.28
Adj. R²	0.27

	Est.	S.E.	t val.	p
(Intercept)	2.27	3.37	0.67	0.50
TEAM_BATTING_H	0.03	0.00	14.60	0.00
TEAM_BATTING_3B	0.09	0.02	5.31	0.00
TEAM_BATTING_HR	0.05	0.01	6.04	0.00
TEAM_BATTING_BB	0.03	0.00	12.25	0.00
TEAM_BASERUN_SB	0.04	0.00	10.94	0.00
TEAM_FIELDING_E	-0.02	0.00	-6.55	0.00

Standard errors: OLS

Table 3.4. Model results using statistically significant variables.

3.4 Backwards Elimination

For this model, variables that are not statistically significant were removed to generate a best-fit model. Table 3.5 includes the results of this model, and residuals are plotted in Figure 3.3.

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(5,2270)	80.37
R^2	0.15
Adj. R^2	0.15

	Est.	S.E.	t val.	p
(Intercept)	44.03	3.44	12.79	0.00
TEAM_BATTING_2B	0.06	0.01	7.29	0.00
TEAM_PITCHING_H	0.01	0.00	8.10	0.00
TEAM_PITCHING_HR	0.06	0.01	8.37	0.00
TEAM_PITCHING_BB	0.03	0.01	6.12	0.00
TEAM_PITCHING_SO	-0.01	0.00	-9.47	0.00

Standard errors: OLS

Table 3.5. Backwards elimination model results.

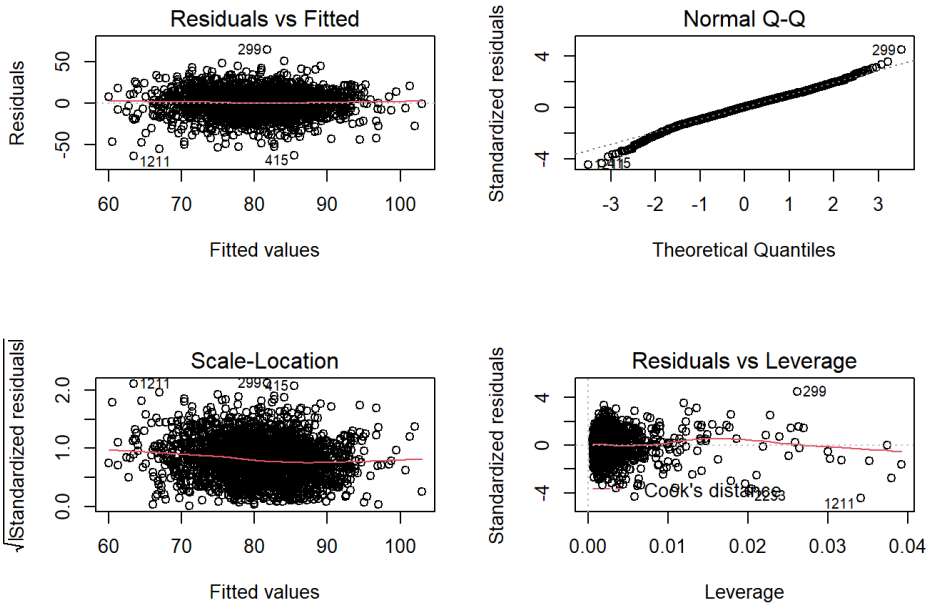

Figure 3.3. Backwards-elimination model residual plots.

3.5 Power Transform

Using a power model may be more effective considering each independent variable doesn't appear to have a truly linear relationship with wins. Here we create a model using a cubit for each independent variable. Partial results are tabulated in Table 3.6; see Appendix A for full results. Residual plots are given in Figure 3.

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(36,2239)	35.32
R^2	0.36
Adj. R^2	0.35

Table 3.6. Power transform results.

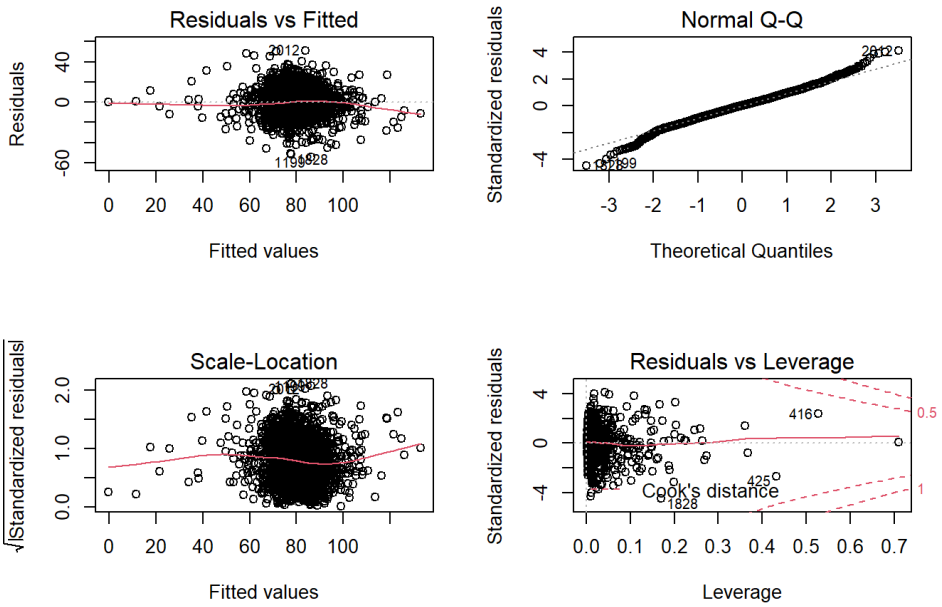

Figure 3.4 Log transformation residual plots.

3.6 Power Transformation with Reverse Elimination

Using reverse elimination on model 5, we removed variables with p-values higher than 0.05. Partial results are shown in Table 3.7, and residual plots are given in Figure 3.5.

Observations	2276
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(26,2249)	48.73
R^2	0.36
Adj. R^2	0.35

Table 3.7. Log transformation with reverse elimination results.

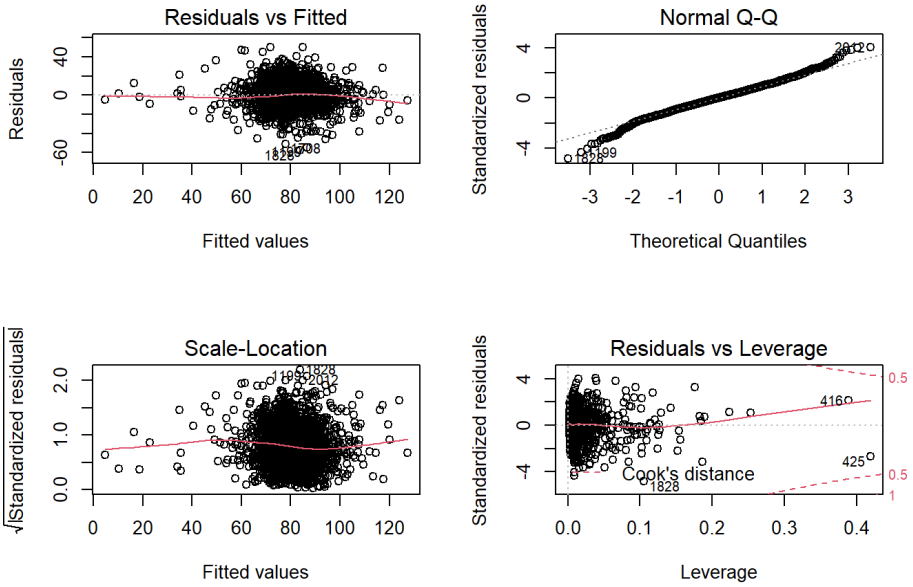

Figure 3.5. Log transformation residual plots.

3.7 Automated Reverse Elimination

Based on residual analysis, all six models produced residuals that are heteroschedastic and not normally distributed. Therefore, we tried a different approach and ran the model without imputation and without transformation. We performed backwards elimination using an automated function to generate a new function call based on removing the parameter with the highest p-value (see Appendix A for R code and full model results). The final model output is given in Table 3.7, with residual plots shown in Figure 3.6

Observations	191 (2085 missing obs. deleted)
Dependent variable	TARGET_WINS
Type	OLS linear regression

F(6,184)	34.07
R²	0.53
Adj. R²	0.51

	Est.	S.E.	t val.	p
(Intercept)	63.47	19.17	3.31	0.00
TEAM_BATTING_H	0.03	0.01	2.54	0.01
TEAM_PITCHING_HR	0.09	0.02	3.81	0.00
TEAM_PITCHING_BB	0.06	0.01	5.93	0.00
TEAM_PITCHING_SO	-0.03	0.01	-4.02	0.00
TEAM_FIELDING_E	-0.17	0.04	-4.36	0.00
TEAM_FIELDING_DP	-0.12	0.04	-3.44	0.00

Standard errors: OLS

Table 3.7. Log transformation with reverse elimination results.

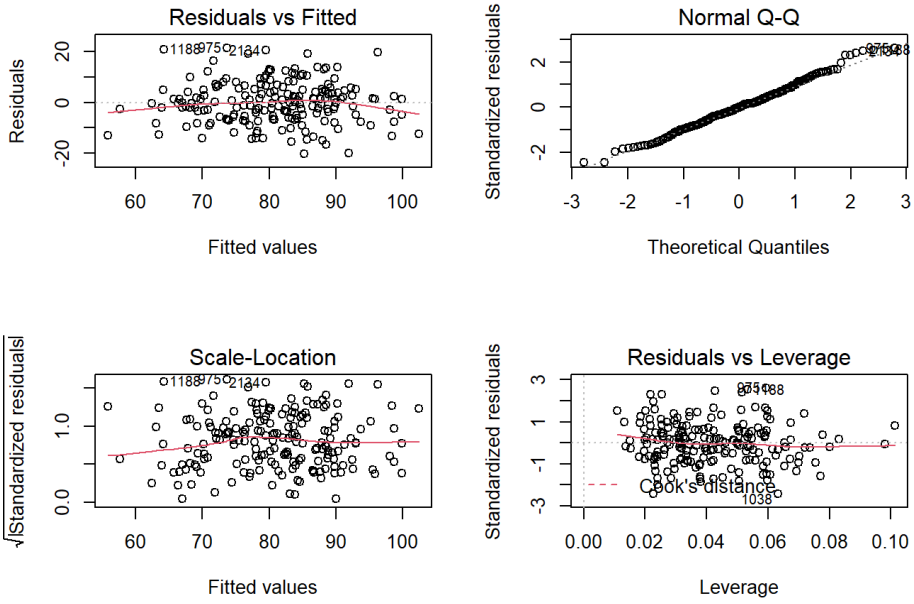

Figure 3.6. Automated reverse elimination residual plots.

4. MODEL SELECTION

Model selection was based on a number of factors, including verifying the validity of the model by checking for homoschedasticity of residuals, verifying that the residuals are normally distributed, and verifying linearity (parameters were assumed to be independent and not colinear).

To aid in residual analysis, we used the Breusch-Pagan test to check for homoschedasticity of residuals, along with the Shapiro test to verify they were normally distributed. We also generated plots to visually examine the residuals.

Based on residual analysis, none of the first six models exhibited residuals that were homoschedastic of residuals, nor were the residuals normally distributed. Therefore, the validity of the models couldn't be ascertained. As a result, we ran a separate set of models programmatically to perform reverse elimination. This yielded a model that produced homoschedastic residuals that were normally distributed.

This model performed better than the previous models, with a higher adjusted R-squared value of 0.511 and lower mean squared error of 69.2. These values indicate a moderate linear relationship between the explanatory and response variables. However, it was noted that only 191 observations remained after specifying that the model remove NaN values. Therefore, despite the better performance, we selected model 6. While this model exhibited heteroschedastic, non-normal residuals, it performed the best in terms of R-squared and mean squared errors, while using the most robust dataset.

To evaluate whether the complexity of model 6 was warranted, we compared it to that of the first model, we an analysis of variance (ANOVA) test between the two models, the results of which are shown in Figure 4.1.

```
Analysis of Variance Table

Model 1: TARGET_WINS ~ INDEX + TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
    TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO + TEAM_BASERUN_SB +
    TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
    TEAM_FIELDING_E
Model 2: TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + I(TEAM_BATTING_2B^2) +
    I(TEAM_BATTING_2B^3) + I(TEAM_BATTING_3B^2) + I(TEAM_BATTING_3B^3) +
    TEAM_BATTING_BB + I(TEAM_BATTING_BB^2) + I(TEAM_BATTING_BB^3) +
    TEAM_BATTING_SO + I(TEAM_BATTING_SO^2) + TEAM_BASERUN_SB +
    I(TEAM_BASERUN_SB^2) + TEAM_PITCHING_H + I(TEAM_PITCHING_H^2) +
    I(TEAM_PITCHING_H^3) + TEAM_PITCHING_HR + I(TEAM_PITCHING_HR^2) +
    I(TEAM_PITCHING_HR^3) + I(TEAM_PITCHING_BB^2) + I(TEAM_PITCHING_BB^3) +
    TEAM_PITCHING_SO + I(TEAM_PITCHING_SO^2) + I(TEAM_PITCHING_SO^3) +
    TEAM_FIELDING_E + I(TEAM_FIELDING_E^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2262 405872
2   2249 360633 13     45239 21.702 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.1. ANOVA results.

As indicated by the low P-value of 2.2e-16, there is enough evidence to reject the null hypothesis that the means are the same. Therefore, we must accept that the means are statistically different. This confirms that there is enough justification in adding complexity to the model. We'll therefore select model 6. The coefficients from this model follow on Table 4.1

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.089e+00  2.382e+01   0.088  0.93011
TEAM_BATTING_H        5.061e-02  4.030e-03  12.557  < 2e-16 ***
TEAM_BATTING_2B       1.106e+00  1.914e-01   5.778 8.59e-09 ***
I(TEAM_BATTING_2B^2) -4.395e-03  7.610e-04  -5.775 8.76e-09 ***
I(TEAM_BATTING_2B^3)  5.596e-06  9.935e-07   5.633 1.99e-08 ***
I(TEAM_BATTING_3B^2)  2.703e-03  3.263e-04   8.283  < 2e-16 ***
I(TEAM_BATTING_3B^3) -1.516e-05  1.951e-06  -7.769 1.20e-14 ***
TEAM_BATTING_BB       2.721e-01  4.648e-02   5.854 5.49e-09 ***
I(TEAM_BATTING_BB^2) -4.758e-04  1.021e-04  -4.658 3.37e-06 ***
I(TEAM_BATTING_BB^3)  3.026e-07  7.204e-08   4.201 2.76e-05 ***
TEAM_BATTING_SO       5.130e-02  1.293e-02   3.967 7.51e-05 ***
I(TEAM_BATTING_SO^2) -3.716e-05  7.371e-06  -5.042 4.97e-07 ***
TEAM_BASERUN_SB       7.394e-02  8.914e-03   8.295  < 2e-16 ***
I(TEAM_BASERUN_SB^2) -7.182e-05  1.601e-05  -4.485 7.66e-06 ***
TEAM_PITCHING_H      -1.423e-01  2.539e-02  -5.607 2.32e-08 ***
I(TEAM_PITCHING_H^2)  5.539e-05  1.087e-05   5.094 3.80e-07 ***
I(TEAM_PITCHING_H^3) -6.520e-09  1.466e-09  -4.448 9.09e-06 ***
TEAM_PITCHING_HR     -1.986e-01  4.090e-02  -4.855 1.29e-06 ***
I(TEAM_PITCHING_HR^2) 2.022e-03  3.216e-04   6.288 3.86e-10 ***
I(TEAM_PITCHING_HR^3) -4.401e-06  7.748e-07  -5.680 1.52e-08 ***
I(TEAM_PITCHING_BB^2) -3.036e-04  8.881e-05  -3.419  0.00064 ***
I(TEAM_PITCHING_BB^3)  3.979e-07  1.281e-07   3.107  0.00192 **
TEAM_PITCHING_SO      -3.527e-02  1.761e-02  -2.003  0.04530 *
I(TEAM_PITCHING_SO^2)  3.219e-05  2.244e-05   1.434  0.15166
I(TEAM_PITCHING_SO^3) -1.183e-08  8.970e-09  -1.319  0.18740
TEAM_FIELDING_E      -1.205e-01  1.547e-02  -7.788 1.03e-14 ***
I(TEAM_FIELDING_E^2)  1.331e-04  2.050e-05   6.493 1.03e-10 ***
```

Table 4.1. Final model coefficients.

Predictions on Evaluation Dataset

Like the training data, the evaluation dataset included a number of NaN values. Therefore, we imputed data using the same method as with the training set. Then, using the coefficients produced by model 6, we ran the predictions against the 259 observations in the evaluation dataset. Table 4.2 shows the first ten observations of the evaluation dataset including the number of predicted wins.

TARGET_WINS <dbl>	INDEX <int>
62	9
64	10
72	14
86	47
74	60
67	63
79	74
62	83
75	98
69	120

Table 4.2. First ten observations of evaluation set with predicted wins.

A full listing of predictions is included in Appendix A.