UNIT 1 Homework
Moneyball OLS Regression Project

Jim Perkins
DL411 Sec 58

<u>80</u> Bingo Bonus Points Attempted

20 Points	Proc GLM----------pgs 23-25
	Proc GENMOD----------pgs 26-31
20 Points	Decision Trees using JMP 11: missing value imput----pgs 8-10
20 Points	R Programing for MoneyBall Model Building----------pgs 32-42
10 Points	SAS MACRO used for Scoring: At the end of code file
10 Points	Scored test file handed in as a SAS DATA SET

Moneyball OLS Regression Project

Table of Contents:

Moneyball OLS Regression Project

Introduction

Predicting the number of wins by a baseball team in a regular season was the goal of this exciting project. To accomplish this task, a data set containing information on 2276 baseball teams was utilized. This data set first was examined and then prepared prior to using this "modified" data to build predictive models. Using the model building technique of linear regression, three different predictive models were constructed. The best model was selected. The best model predicting team wins in a regular season will be graded against other models.

Data Exploration

The data set has performance information on 2276 professional baseball teams from 1871 to 2006. Please note that statistics were adjusted to a 162 game season for the early seasons that had fewer games. The data set contained an index variable as a marker for the teams' identity and a dependent or target variable, called Target_Wins or games the team won in a season. The set also contained 15 predictor variables (Table 1). All variables were in numerical form.

The label or description, N (number of observations), N Miss (number of missing observations), mean, Std Dev (standard deviation), minimum, and maximum for each variable are given in Table 1. For example, Team_Batting_H or base hits by batters had all 2276 values for each team and thus, no missing observations. The average number of hits for each team was 1469 (rounded) plus or minus 145(rounded) hits. The fewest hits by a team was 891 and 2554 was the most hits by each team. Team_Batting_HBP had 2085 (91.6%) missing data and this variable was eliminated from further analysis. The variables of Team_Batting_SO, Team_Baserun_SB, Team_Baserun_CS, Team_Fielding_DP and Team_Pitching_SO had missing data. To include these observations in model development, the missing values for these observations were imputed (See Data Preparation) Also, please note that several variables had 0 as the minimum. For example, Team_Pitching_BB or walks for a team received in a season was 0. This is difficult to understand and possibly 0 was recorded when the data was missing. Variables were screened for this possibility and some data with 0 had values imputed similar to the missing values.

TABLE 1: Label, N (number of observations), N Miss (number of missing observations), mean , Std Dev (standard deviation), minimum and maximum for each variable.

Variable	Label	N	N Miss	Mean	Std Dev	Minimum	Maximum
TARGET_WINS		2276	0	80.7908612	15.7521525	0	146.0000000
TEAM_BATTING_H	Base Hits by batters	2276	0	1469.27	144.5911954	891.0000000	2554.00
TEAM_BATTING_2B	Doubles by batters	2276	0	241.2469244	46.8014146	69.0000000	458.0000000
TEAM_BATTING_3B	Triples by batters	2276	0	55.2500000	27.9385570	0	223.0000000
TEAM_BATTING_HR	Homeruns by batters	2276	0	99.6120387	60.5468720	0	264.0000000
TEAM_BATTING_BB	Walks by batters	2276	0	501.5588752	122.6708615	0	878.0000000
TEAM_BATTING_HBP	Batters hit by pitch	191	2085	59.3560209	12.9671225	29.0000000	95.0000000
TEAM_BATTING_SO	Strikeouts by batters	2174	102	735.6053358	248.5264177	0	1399.00
TEAM_BASERUN_SB	Stolen bases	2145	131	124.7617716	87.7911660	0	697.0000000
TEAM_BASERUN_CS	Caught stealing	1504	772	52.8038564	22.9563376	0	201.0000000
TEAM_FIELDING_E	Errors	2276	0	246.4806678	227.7709724	65.0000000	1898.00
TEAM_FIELDING_DP	Double Plays	1990	286	146.3879397	26.2263853	52.0000000	228.0000000
TEAM_PITCHING_BB	Walks allowed	2276	0	553.0079086	166.3573617	0	3645.00
TEAM_PITCHING_H	Hits allowed	2276	0	1779.21	1406.84	1137.00	30132.00
TEAM_PITCHING_HR	Homeruns allowed	2276	0	105.6985940	61.2987469	0	343.0000000
TEAM_PITCHING_SO	Strikeouts by pitchers	2174	102	817.7304508	553.0850315	0	19278.00

The distribution of the target variable, TARGET_WINS was basically a normal distribution (Figure 1). This allowed linear regression analysis to build predictive models for this data set (See Build Model).

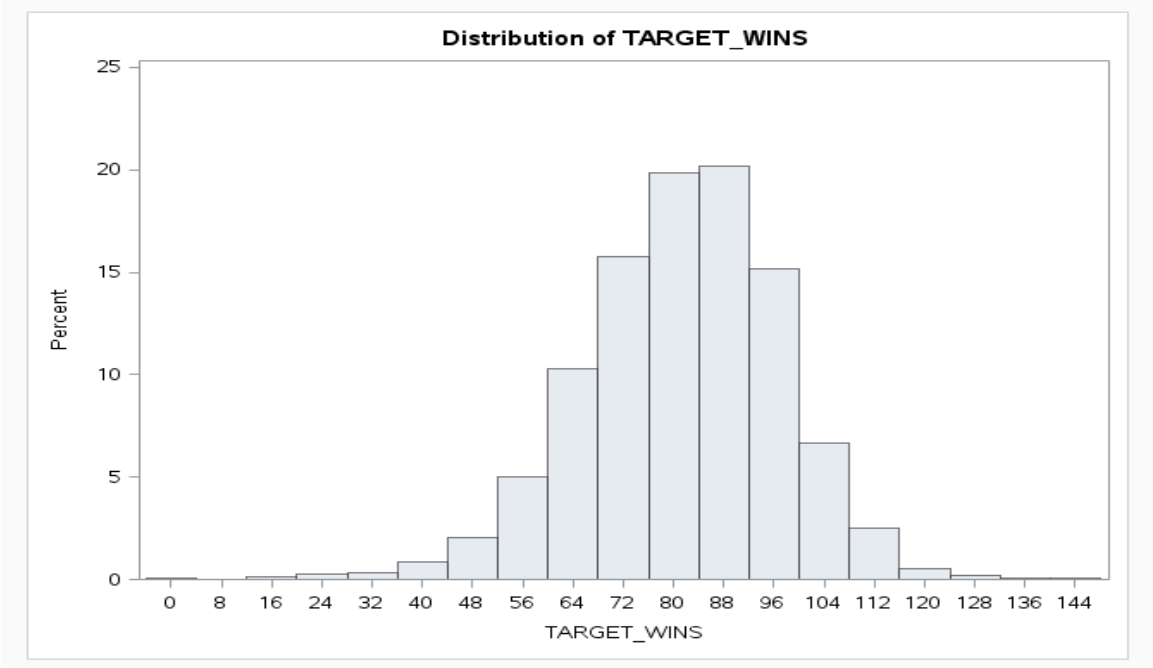


Figure 1: Normal distribution of the dependent variable, Target_Wins.

In linear regression analysis, strong correlations between predictor variables can result in inaccurate models. This is called multicollinearity and the assumption in linear regression is that no variables in the model are strongly correlated. All correlations of the predictor variables were evaluated (data not shown). Those correlations that may cause problems with model development are highlighted. There was an unusually high correlation (0.97 rounded) between Team_Batting_HR and Team_Pitching_HR. This strong correlation was hard to explain. This correlation could interfere with model development and one of these two variables were eliminated in the build model section before a full model was developed (Figure 2).

There are other correlations that are interesting but may or may not lead to problems in model development. Those variables that had a correlation less than 0.80 were not be eliminated prior to model building. There was a positive correlation (0.73 rounded) between Team_Batting_HR and Team_Batting_SO (Figure 2). This was understandable. The batters for the teams trying to hit more home runs are swinging more freely and thus, also strike out more. There is a negative correlation (-0.64 rounded) between Team_Batting_HR and Team_Batting_3B (Figure 2). It appears the teams that hit more triples do not hit more home runs.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	TEAM_BATTING_HR
TEAM_BATTING_SO Strikeouts by batters	0.72707 <.0001 2174
TEAM_PITCHING_HR Homeruns allowed	0.96937 <.0001 2276
TEAM_BATTING_3B Triples by batters	-0.63557 <.0001 2276

Figure 2: Pearson Correlation Coefficients (a measure of a relationship between variables) and the Prob (probability of significance) between Team_Batting_HR compared to Team_Batting_SO, Team_Pitching_HR and Team_Batting_3B.

Another interesting correlation (0.66 rounded) was that Team_Baserun_SB, or those teams that steal more bases were also those teams that get caught stealing or Team_Baserun_CS (Figure 3A). This was understandable for those teams that attempted more stolen bases actually stole more bases and got caught more while stealing. Likewise, the correlation of 0.67 (rounded) between Team_Pitching_H and Team_Fielding_E was understandable (Figure 3B). The team that gives up more hits is also the team that has more opportunities to make errors fielding the ball.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	TEAM_BASERUN_SB
TEAM_BASERUN_CS Caught stealing	0.65524 <.0001 1504

3A

Pearson Correlation Coefficients, N = 2276 Prob > |r| under H0: Rho=0	
	TEAM_PITCHING_H
TEAM_FIELDING_E Errors	0.66776 <.0001

3B

Figure 3: Pearson Correlation Coefficients (a measure of a relationship between variables) and the Prob (probability of significance) between Team_Baserun_SB compared to Team_Baserun_CS(3A) and Team_Pitching_H to Team_Fielding_E (3B).

Scatter plots were used to reveal any association between the target variable, Target_Wins and each predictor variable. Likewise, scatter plots are useful to visualize outliers (values that are greatly smaller or larger that the rest of the values) of the predictor variables. Scatter plots with simple regression and loess lines were determined for the offensive (Figure 4) and defensive (Figure 5) statistics. Simple regression lines reveal a linear relationship between the predictor and the target variable. Loess means local regression. Loess lines indicate when the relationship between the predictor variable and the dependent variable may not be linear or outliers are present that may cause inaccuracy in linear regression models. Scatter plots with a great disparity between the regression and loess lines or clumping of data on one end of the plot reveal variables that need to have the data transformed to improve the variable importance and thus, accuracy of the model. For example, in Figure 4, the association between Team_Batting_H or base hits by batters to Target_Wins have both the regression line and loess line mostly progressing in the same directions. It was reasonable that more hits by a team relates to more wins for the team. This variable was not modified or transformed. All the variables for the offensive statistics appeared to have appropriate relationships with Target_Wins. However, Team_Batting_SO appeared to have many observations with 0s and not associated with Target_Wins. During data preparation these 0s were imputed using the same rules as the missing observation for Team_Batting_SO. No variable for the offensive statistics needed transformation.

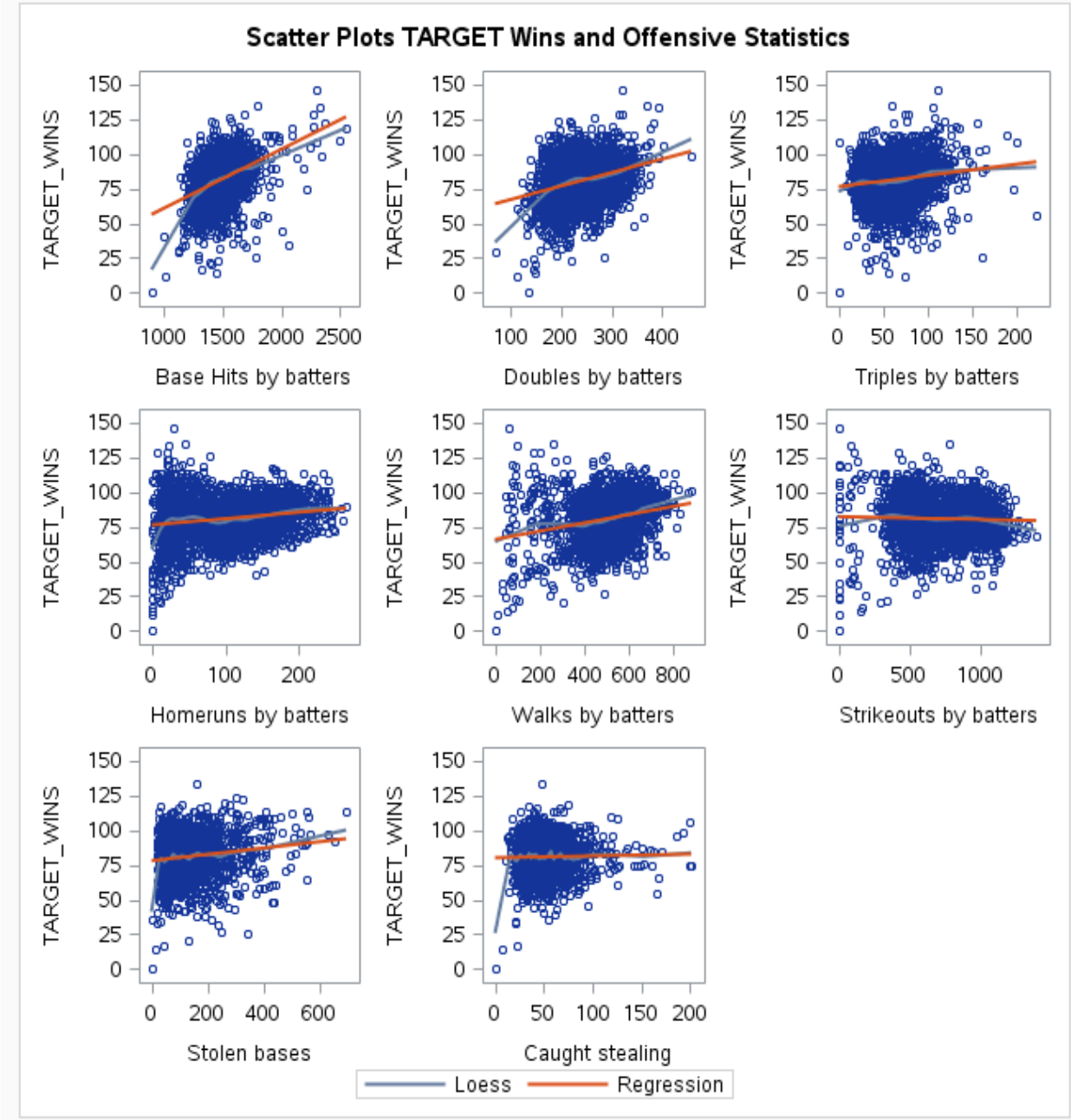


Figure 4: Scatter Plots between Target_Wins and the Offensive Variables with Loess and Regression Lines.

The scatter plots for the defensive statistics revealed several plots with interesting relationships between Target_Wins and the predicting variable (Figure 5). The relationship between Target_Wins with Team_Fielding_DP (Double Plays) and Team_Pitching_HR (Homeruns allowed) seemed appropriate and was not modified. The relationship between Target_Wins with Team_Fielding_E (Errors), and Team_Pitching_H (Hits Allowed) had data far from the main group of data but in the appropriate direction (more errors, less wins or more hits allowed, less wins) and the loess and regression lines are basically in the same direction. At this point these variables were not transformed. The Team_Pitching_BB (Walks allowed) and Team_Pitching_SO had only a few points greatly separated from the main group. There was also a greater difference in the loess and regression line for walks allowed to Target_Wins. Thus, these two variables were transformed so that the spread between variables was less. A natural log (LN) transformation was used before these variables were entered into the model. Also, one team had 0 for Team_Pitching_BB. This one observation was deleted. There also appeared to be many Team_Pitching_SO that are O's. There is little chance that these values would be 0. Those Team_Pitching_SO with O were replaced with the mean before being transformed.

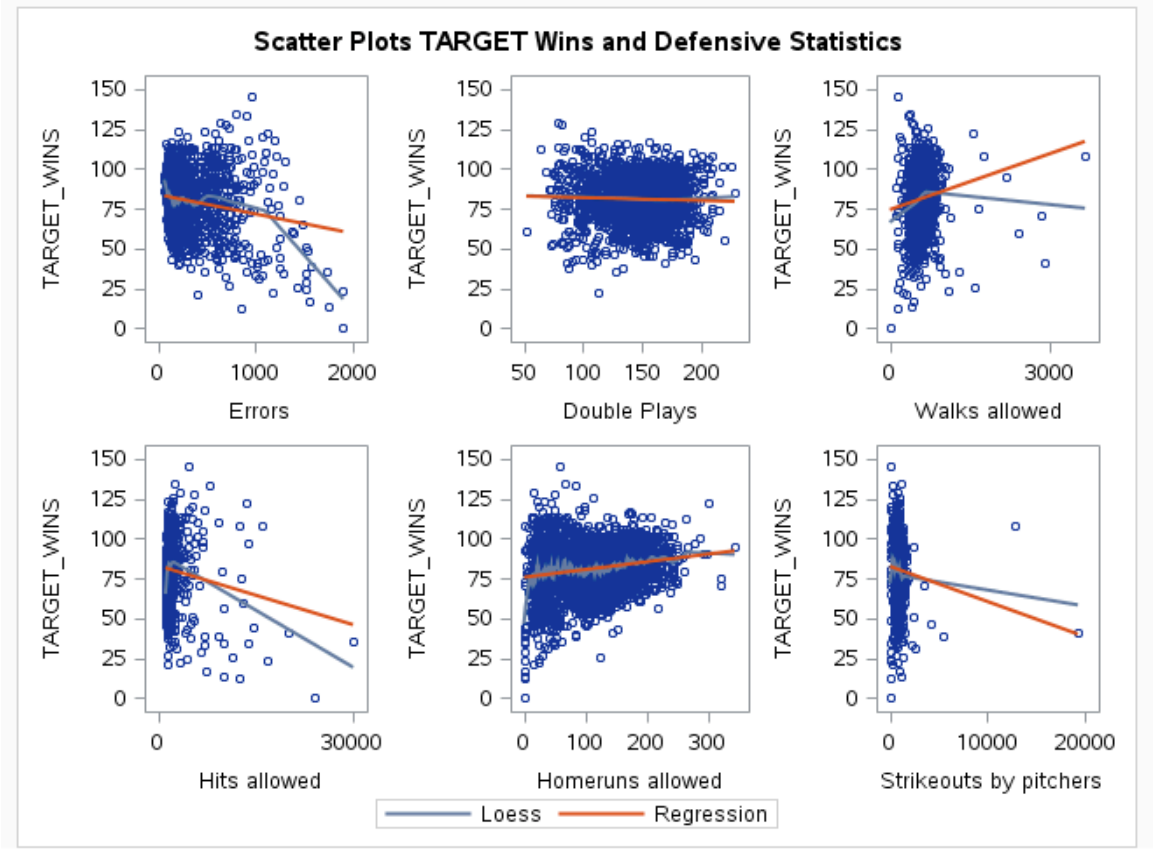


Figure 5: Scatter Plots between Target_Wins and the Defensive Variables with Loess and Regression Lines.

To summarize, after the data exploration step, the variable Team_Batting_HBP with 91.6% missing observations was eliminated at this point from further analysis. The five variables with missing data (Team_Batting_SO, Team_Baserun_SB, Team_Baserun_CS, Team_Fielding_DP, and Team_Pitching_SO) had data imputed. Due to strong correlation, either Team_Batting_HR or Team_Pitching_HR was eliminated during the model development stage. Before model development, Team_Pitching_BB and Team_Pitching_SO were modified or transformed.

Data Preparation

Imputation Values for Variables

The first step in data preparation was imputation of the missing observations for the five variables (Team_Batting_SO, Team_Baserun_SB, Team_Baserun_CS, Team_Fielding_DP, and Team_Pitching_SO). Each variable with missing data had a decision tree performed using JMP Pro 11 statistical software. Only the variables of Team_Batting_SO (Figure 6) and Team_Baserun_SB (Figure 7) had decisions trees developed with three or fewer variables with a model that predicted greater than 50% of the variability. This prediction was measure by an R-Square value (a measure of how strong the model predicts the results) of around 0.50 or greater. Also, the decision trees for these two variables agreed with common sense rules for baseball.

Team_Batting_SO:

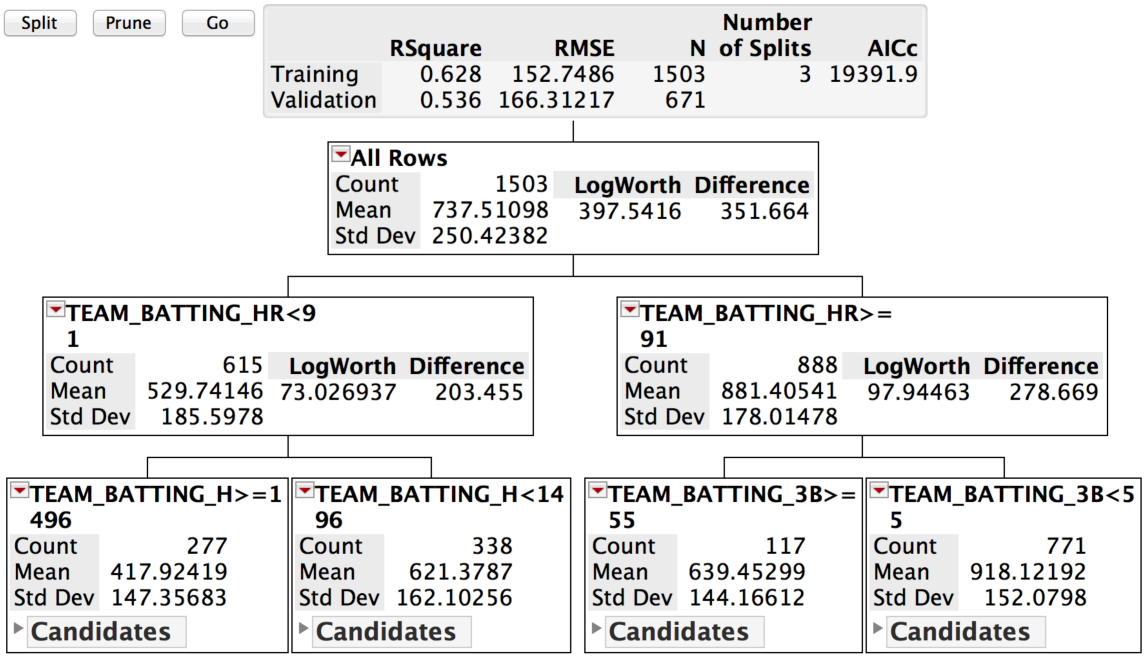


Figure 6: Decision Tree developed for Team_Batting_SO using 1503 observation in the training set and 671 observations in the validation set. The R-squared on the validation was 0.536.

The data rules to impute the missing value for Team_Batting_SO was as follows-

If Team_Batting_HR <91 and Team_Batting_H >=1496 then 418
If Team_Batting_HR <91 and Team_Batting_H < 1496 then 621
If Team_Batting_HR >= 91 and Team_Batting_3B >=55 then 639
If Team_Batting_HR >=91 and Team_Batting_3B < 55 then 918

Common sense dictates that teams that hit fewer homeruns strike out fewer times. Probably these players are not swinging as freely trying to hit home runs and thus strike out fewer times. Likewise, the teams that hit more also strike out fewer times. Thus, the team that hits the fewest homeruns and has more hits strike out the least when compared to teams that hit more homeruns, fewer hits, and strike out the most. There are many 0s recorded for this variable. It is hard to believe that no team struck out. Those observations with 0s were assumed to be missing and thus where replaced with the above rules.

Team_Baserun_SB:

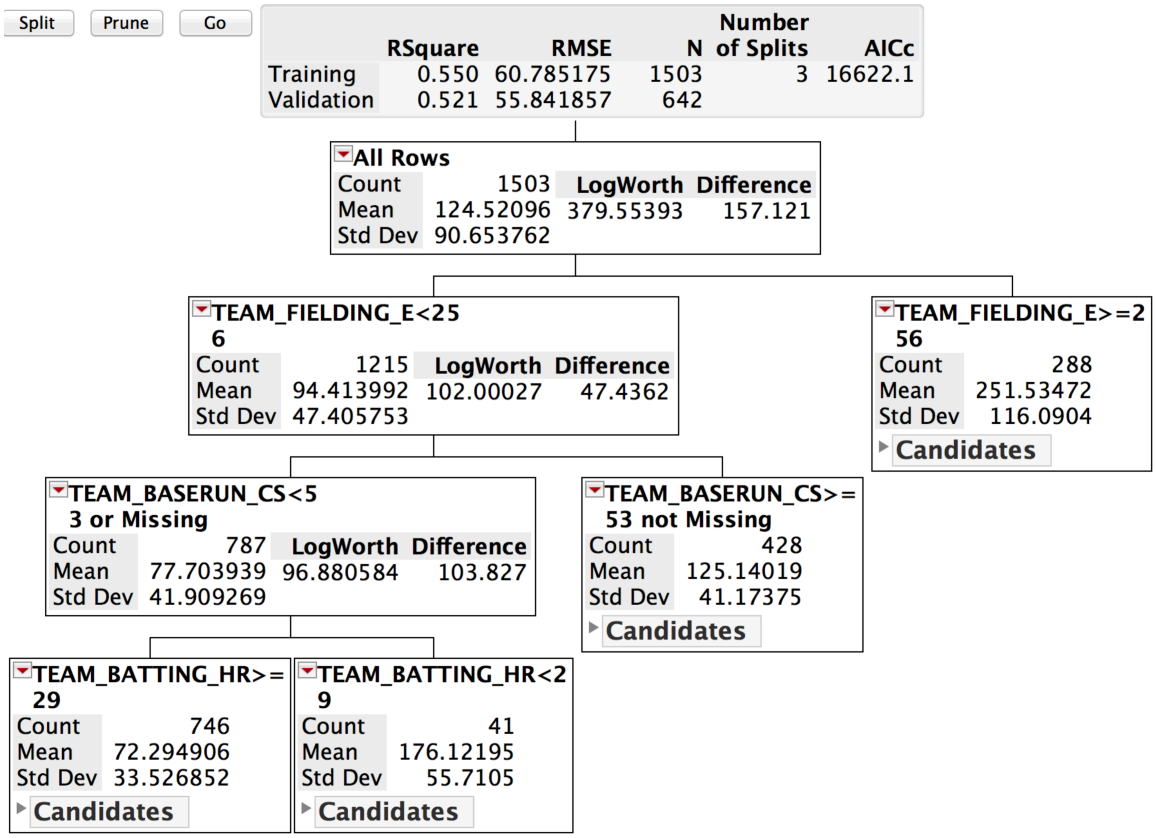


Figure 7: Decision Tree developed for Team_Baserun_SB using 1503 observation in the training set and 642 observations in the validation set. The R-squared on the validation was 0.521.

The data rules to impute missing data for Team_Baserun_SB was as follows—

If Team_Fielding_E > 256 then 251
If Team_Fielding_E <256 and Team_Baserun_CS >=53 then 125
If Team Fielding_E <256 and Team_Baserun_CS < 53 and Team_Batting_HR >=29 then 72
If Team_Fielding_E <256 and Team_Baserun_CS <53 and Team_Batting_HR <29 then 176.

It appears that teams that make the most errors steal the most bases. Common sense probably tells us that players on these teams are selected to run fast and not for their ability to catch and throw the ball. Teams that get caught stealing more bases also steal more bases probably because they attempted more times to steal bases. Finally, teams that hit more homeruns steal fewer bases because those on base do not attempt to steal bases but wait on base until the batter hits a homerun.

Other variables had the mean impute for their missing values. The variable of Team_Baserun_CS had the mean of 52.8 imputed for missing values.
Team_Fielding_DP had the mean of 146.4 imputed for missing values.
Team_Pitching_SO had the mean of 817 used for the missing values. There were also several 0s for this variable. This was unlikely and those with 0 were assumed to be missing. Those variables with 0 were replaced with the average for this variable.

Transforming Variables

Team_Pitching_BB and Team_Pitching_SO had extreme outliers from the majority of the other values. These two variables were transformed with a natural log (LN) function in order to reduce the large spread in the values for these variables (Fig 8).

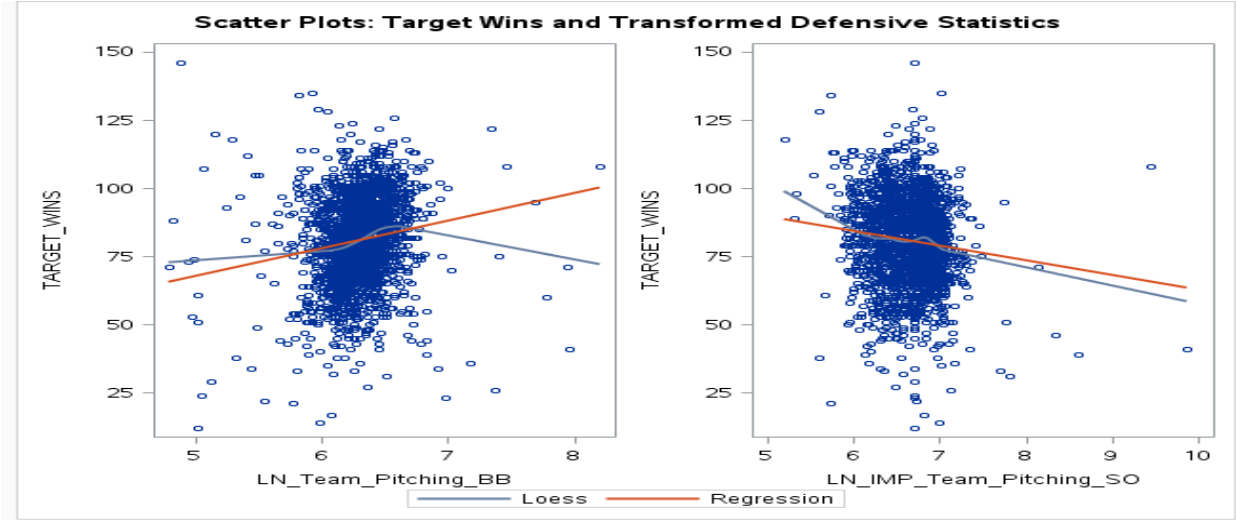


Figure 8: New scatter plots for the two variables that were transformed using natural log.

In summary, the new predictor variables to be used in model building following the data preparation step are as follows:

New	Old
Team_Batting_H	
Team_Batting_HR	
Team_Batting_2B	
Team_Batting_3B	
Team_Batting_BB	
IMP_Team_Baserun_CS	Team_Baserun_CS with mean value for imputation
IMP_Team_Baserun_SB	Team_Baserun_SB with decision rules for imputation
IMP_Team_Batting_SO	Team_Batting_SO with decision rules for imputation
Team_Pitching_H	
Team_Fielding_E	
IMP_Team_Fielding_DP	Team_Fielding_DP with mean value for imputation
LN_Team_Pitching_BB	Team_Pitching_BB with natural log(LN) transformation One value of 0 was deleted before transformation
LN_IMP_Team_Pitching_SO	Team_Pitching_SO with mean for imputation and LN Transformation (all 0 imputation with mean)
M_Team_Baserun_CS	Team_Baserun_CS missing values flagged as missing
M_Team_Fielding_DP	Team_Fielding_DP missing values flagged as missing
M_Team_Pitching_SO	Team_Pitching_SO missing values flagged as missing
M_Team_Baserun_SB	Team_Baserun_SB missing values flagged as missing
M_Team_Batting_SO	Team_Batting_SO missing values flagged as missing

(Please note the two variables undergoing a LN transformation did not have any values that were 0 after imputation and/or deletion of the 0 values.)

Build Models

Linear regression analysis was used to build a predictive model. Linear regression assumes an association between a set of predictor variables and a single, continuous, dependent variable. The linear regression equation describes a linear relationship and is written as

$$Y= \beta_0 + \beta_1*X_1 + \beta_2 *X_2 + \beta_3 * X_3 + \ldots \beta_k * X_k$$

Y is the dependent variable or variable to be predicted like TARGET_WINS. K is the total number of predictor variables in the model. The above formula is very helpful to determine the influence of a predictor variable on the target variable. β_0 is called the intercept or the value of Y when all the Xs equal 0. This may or may not be useful. The remaining β's are the coefficients or multipliers of the predictor variables. If all other values are held constant, then X_1 times β_1 is the effect of the X_1 variable on Y. For example if β_1 equals 0.05 and X_1 equals Team_Batting_H, then a team having 900 hits would have 45 wins (0.05 * 900 = 45) associated with Team_Batting_H. The sign of the β is also very important. If the β is positive the increasing value of the variable results in a higher Y and if the β is negative the increasing value of the

variable results in a lower Y. For example, if the coefficient for TEAM_FIELDING_E was negative, then the more errors the team commits the fewer wins the team has in a season.

Before formal model development, the two variables, Team_Batting_HR and Team_Pitching_HR, that had almost perfect correlation were evaluated. Inserting both variables in the model would result in problems with multicollinearity and thus, possibly leading to inaccurate coefficients. One of these two variables was removed prior to modeling building. A simple linear regression model was performed with each variable against the Target_Win variable. Due to the strong correlation between the variables, the resulting models coefficients were very similar. The coefficient for Team_Batting_HR was a positive 0.0449. The fit plot or plot to review the predictions from Team_Batting_HR indicated that the more homeruns a team hits the more wins (Figure 9). The coefficient for Team_Pitching_HR was also a positive 0.0476 and the fit plot revealed that the more homeruns given up by the pitching staff the more wins for the team (Figure 10). This is not a logical conclusion and thus, the variable Team_Pitching_HR was eliminated from further analysis.

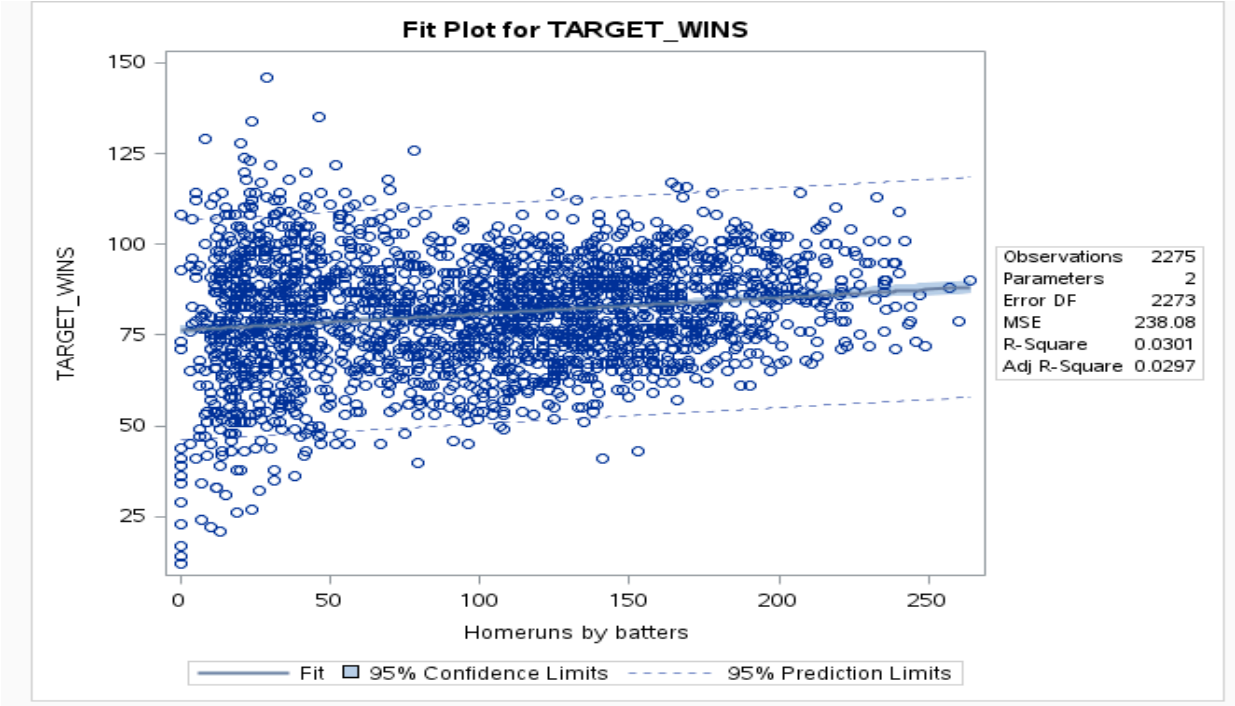


Figure 9: Plot of the predicted Target Wins to Team_Batting_HR.

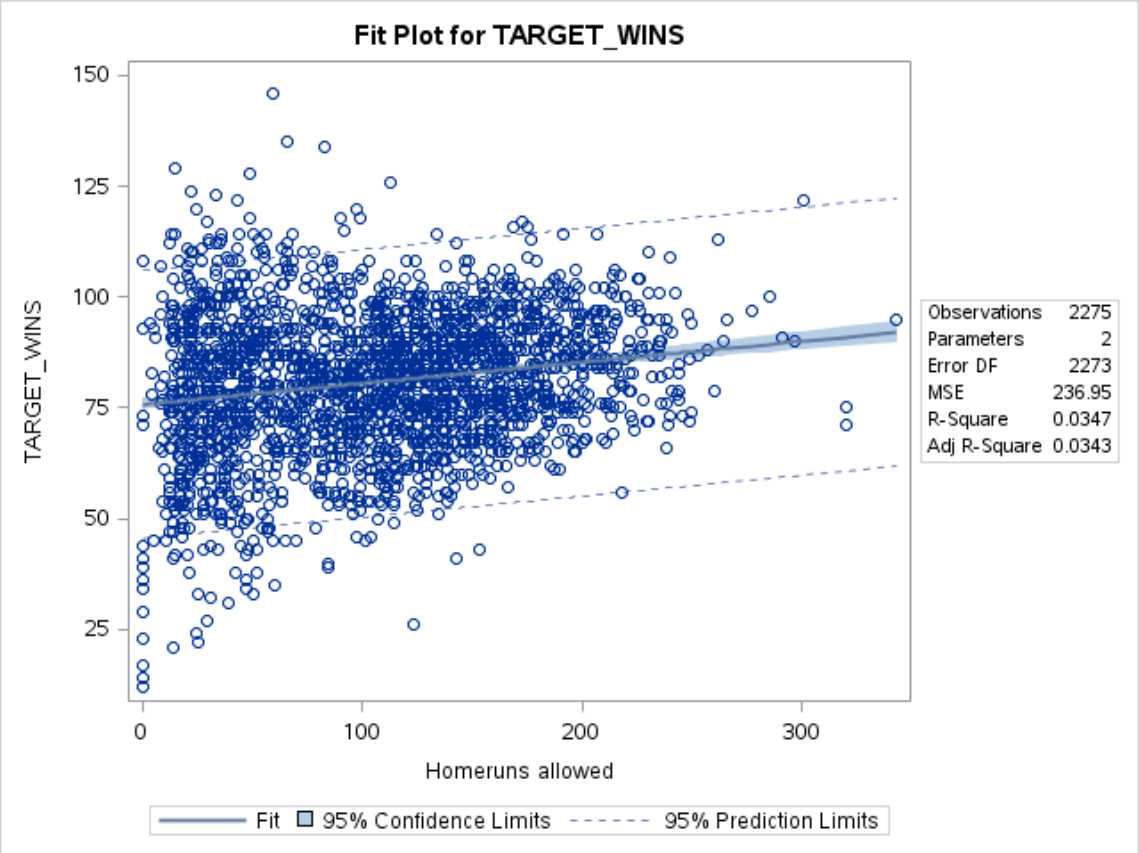


Figure 10: Plot of the predicted Target Wins to Team_Pitching_HR.

Three Different Models Built

Three different multiple linear regression models were developed. The first model (Model 1) was developed utilized both backward and stepwise selection methods. Backward selection enters all variables into the model. Then, the variables are deleted from the model one by one until all the variables in the model are significant or contribute to calculation of Y. Stepwise selection adds variables to the model one by one. After adding a new variable, this method looks at all variables presently in the model to see if they significantly influence Y. A variable can be added and kept in one step but later removed if it does not remain significant.

Both selection methods produced the same significant variables or model (Table 2). The parameter estimates or coefficients indicate the strength and direction the variable has on Target_Wins with the standard error indicating the spread of that strength. The t value was used to indicate the Pr >|t| or the significance of the variable. The variance inflation indicates the relationship or correlation of this variable to other variables in this model. For example, Team_Batting_H or the more base hits by the team had a positive 0.049(rounded) for the coefficient. This indicated the more hits for the team the more wins. The variable was strongly significant on the results of the model as indicated by Pr >|t| value of <.0001. A value of less than 5 for the variance inflation indicated that this variable has only

limited association with other variables. Another example is Team_Fielding_E. A coefficient of minus 0.056(rounded) meant that the more errors a team committed the fewer Target_Wins they had. The variable was moderately associated with another variable with a VIF of 8.29(rounded). Please note, that Team_Batting_2B, Team_Pitching_H and IMP_Team_Fielding_DP had coefficients that go against common logic for baseball. Team_Batting_2B parameter estimate is minus 0.04(rounded). This indicated that the more a team hits doubles the less the team wins. Also the parameter estimate for Team_Pitching_H was a positive 0.002(rounded), indicating the more hits a team gave up the more wins it had. Likewise, not following common sense was the IMP_Team_Fielding_DP. A coefficient of minus 0.1 indicated the more double plays the team made, the less wins for the team. Also note, that the missing values for Team_Fielding_DP, Team_Baserun_SB and Team_Batting_SO were all significantly associated with winning more games with parameter estimates for 4.17, 27.1 and 7, respectively. This indicated that missing values for these variables were associated with more Target_Wins.

Table 2: Significant Variables in Model 1 Created by both Backward and Stepwise Regression.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	42.39418	13.92625	3.04	0.0024	0
TEAM_BATTING_H	Base Hits by batters	1	0.04860	0.00353	13.77	<.0001	4.01240
TEAM_BATTING_HR	Homeruns by batters	1	0.07020	0.00928	7.56	<.0001	4.89673
TEAM_BATTING_2B	Doubles by batters	1	-0.03638	0.00868	-4.19	<.0001	2.55259
TEAM_BATTING_3B	Triples by batters	1	0.05937	0.01578	3.76	0.0002	3.01139
TEAM_BATTING_BB	Walks by batters	1	0.03446	0.00591	5.83	<.0001	8.09436
IMP_Team_BaseRun_SB		1	0.05025	0.00464	10.83	<.0001	2.70968
IMP_Team_Batting_SO		1	-0.01139	0.00226	-5.03	<.0001	4.43711
TEAM_PITCHING_H	Hits allowed	1	0.00211	0.00039393	5.35	<.0001	4.23909
TEAM_FIELDING_E	Errors	1	-0.05638	0.00325	-17.37	<.0001	8.28991
IMP_Team_Fielding_DP		1	-0.10042	0.01325	-7.58	<.0001	1.63672
LN_Team_Pitching_BB		1	-4.40487	2.38650	-1.85	0.0651	5.41912
M_Team_Fielding_DP		1	4.17203	1.44090	2.90	0.0038	3.52932
M_Team_BaseRun_SB		1	27.05476	1.66947	16.21	<.0001	2.34626
M_Team_Batting_SO		1	7.04392	1.31513	5.36	<.0001	1.35112

Another model, Model 2 was created using the all variables except the flagged variables for the missing values. It was theorized that the developed predicted model might be used only with new data and all the data would be complete without missing values. After entering all the variables into the model, the variables of Team_Baserun_CS and Team_Pitching_H were removed manually since these variables were not significant. The significant variables in this model are in Table 3.

The variables that are contrary to common sense are again the IMP_Team_Fielding_DP and Team_Batting_2B. IMP_Team_Fielding_DP with a coefficient of minus 0.1 (rounded) indicated that the more double plays a team turns the less wins for that team. Likewise, with a coefficient of minus 0.022 for Team_Batting_2B indicated that the more doubles a team hits was associated with less wins. Note, vif for IMP_Team_Batting_SO is greater than 10 and thus, is associated with another variable. From the data exploration phase it was noted that Team_Batting_SO and Team_Batting_HR were correlated at 0.72. This could cause a problem with predictions if this model was chosen as the best model.

Table 3: Significant Variables for Model 2 Created without Missing Variables being Flagged.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-37.32703	12.57377	-2.97	0.0030	0
TEAM_BATTING_H	Base Hits by batters	1	0.05243	0.00363	14.46	<.0001	3.91073
TEAM_BATTING_HR	Homeruns by batters	1	0.07264	0.00959	7.57	<.0001	4.82278
TEAM_BATTING_2B	Doubles by batters	1	-0.02211	0.00886	-2.50	0.0127	2.45779
TEAM_BATTING_3B	Triples by batters	1	0.04990	0.01598	3.12	0.0018	2.84977
TEAM_BATTING_BB	Walks by batters	1	0.03885	0.00581	6.69	<.0001	7.21989
IMP_Team_BaseRun_SB		1	0.04913	0.00428	11.47	<.0001	2.13392
IMP_Team_Batting_SO		1	-0.03561	0.00377	-9.44	<.0001	11.36386
TEAM_FIELDING_E	Errors	1	-0.03249	0.00229	-14.18	<.0001	3.80792
IMP_Team_Fielding_DP		1	-0.09016	0.01258	-7.17	<.0001	1.36197
LN_Team_Pitching_BB		1	-12.87879	2.26124	-5.70	<.0001	4.49083
LN_IMP_Team_Pitching_SO		1	20.85150	2.24917	9.27	<.0001	7.67915

Model 3 or a common sense model was developed. Model 3 assumed that missing data will always occur and as shown in Model 1, there is a pattern associated with missing data. Thus, the missing data flags significant in the Model 1 were entered into the model. Common sense tells us that hitting more doubles or creating more double plays should led to more wins and not less. The variables that conflicted with common sense were left out of the model. Thus, a model entering the 3 significant missing flag variables was included with all the other variables minus Team_Batting_2B and Team_Fielding_DP. While creating this model, the variables of

LN_Team_Pitching_BB and M_Team_ Fielding_DP were manually removed from the model due to the fact these variables were not significant.

The significant variables for Model 3 are in Table 4. Two variables flagged for missing, M_Team_Baserun_SB and M_Team_Batting_SO had positive coefficients indicating, if these variables were missing the team had more wins. All the other variables made common sense for baseball. For example, with a coefficient of positive 0.019(rounded) for Team_Batting_BB, the more walks a team earned the more wins for that team. Again Team_Fielding_E had a negative coefficient of 0.044(rounded). The more errors a team committed the fewer games they won in a season.

Table 4: Significant Variables for Model 3 Created with Missing Variables Flagged and Variable that did not make common baseball sense being removed.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-15.06699	11.79690	-1.28	0.2017	0
TEAM_BATTING_H	Base Hits by batters	1	0.04231	0.00263	16.11	<.0001	2.13357
TEAM_BATTING_HR	Homeruns by batters	1	0.06772	0.00939	7.21	<.0001	4.81237
TEAM_BATTING_3B	Triples by batters	1	0.06037	0.01560	3.87	0.0001	2.82693
TEAM_BATTING_BB	Walks by batters	1	0.01918	0.00320	5.99	<.0001	2.28558
IMP_Team_BaseRun_SB		1	0.05572	0.00411	13.56	<.0001	2.04466
IMP_Team_Batting_SO		1	-0.01808	0.00328	-5.51	<.0001	8.95392
TEAM_FIELDING_E	Errors	1	-0.04388	0.00227	-19.30	<.0001	3.90289
LN_IMP_Team_Pitching_SO		1	4.35369	1.85734	2.34	0.0192	5.44955
M_Team_BaseRun_SB		1	21.85883	1.68004	13.01	<.0001	2.28241
M_Team_Batting_SO		1	9.02765	1.27519	7.08	<.0001	1.22022

Select Model

Selecting the best model of the three created models depended on evaluating measurements of goodness-of-fit or accuracy of each model while balancing the number of predictor variables. Accurate models with few variables are considered better models. They are easier to implement and are associated with less work by collected less data. Also, the model must be believable for predicting Target_Wins or

the model will not be used. Finally, the model will need to not violate assumptions for proper linear regression model building.
Each model was assigned values for adjusted R_Square (Adj-R-Square), Akaike Information Criterion (AIC), and Schwarz's Bayesian Criterion (SBC)(Table 5). The Adj-R-Square value determines how well the model predicts the dependent variable or Target_Wins. For example, a value of 0.30 determines that 30% of the dependent variable was described by the model. The larger the value the better the model predicts the target. The adj-R-Square takes into account the number of variables in the model. AIC is a measure of model accuracy that considers the number of variables. SBC is another measure of model accuracy that places a more severe penalty on number of variables in the model than does AIC. For AIC and SBC, the lower the number the better the model.

Table 5: The adjusted R_Square (Adj-R-Square), Akaike Information Criterion (AIC), and Schwarz's Bayesian Criterion (SBC) values for the three built models.

Model	Adj-R-Square	AIC	SBC
Model 1	0.40231	11362.82	11448.77
Model 2	0.35248	11542.01	11610.76
Model 3	0.37778	11450.34	11513.37

Model 1 with 14 variables had the highest adj R_Square (0.4023) and the lowest AIC (11362.82) and SBC (11448.77). Thus, this model was slightly more accurate than the other models. However, this model included the variables that teams that hit more doubles won less games, that teams with pitchers that gave up more hits won more games, and teams that turned more double plays won less games. It would be difficult to tell a coach that his pitchers need to let the opposing batters hit more or that his infielders need not practice turning the double play. Coaches would find it difficult to tell their base runners not to get doubles, but to stop on first base even if they could get a double. This model violates common sense for baseball.

Model 2 had an adj-R-Square of 0.3525(rounded), and AIC of 11542 and a SBC of 11611(rounded) and included 11 predictor variables. This model was the least accurate of the three models. This model was developed with the theory that all new data would not have missing variables. This was probably an incorrect assumption. This model also included variables not making common sense about baseball including hitting more doubles leads to winning less games, and making more double plays lead to losing more games.

Model 3 with an adj-R-Square of 0.3778, AIC of 11450(rounded) and SBC of 11513(rounded) was a slightly less accurate model that model 1. However this model only included 10 predictor variables and followed common assumptions of baseball. On the offensive side, more hits including homeruns and triples led to winning more games. Batters being walked more led to winning games while

batters striking out more led to losing more games. Successfully stealing more bases led to more wins. On the defensive side, pitching more strikeouts and creating fewer errors led to winning more games. Since all data will not be without missing variables, this model included missing data for stealing bases or teams striking out. Missing data for these variables associated to more games having been won.

Model 3 was the chosen model even though this model was slightly less accurate than Model 1. Model 3 had fewer variables making data collection and model deployment easier to perform. Most importantly, Model 3 was more believable and followed common baseball knowledge.

The linear formula for model 3 was the following:

Target_Wins = -15.06699
+ 0.04231 * Team_Batting_H
+ 0.06772 * Team_Batting_HR
+ 0.06037 * Team_Batting_3B
+ 0.01918 * Team_Batting_BB
+ 0.05572 * Imp_Team_BaseRun_SB
+ -0.01808 * Imp_Team_Batting_SO
+ -0.04388 * Team_Fielding_E
+ 4.35369 * LN_IMP_Team_Pitching_SO
+ 21.85883 * M_Team_BaseRun_SB
+ 9.02765 * M_Team_Batting_SO

Model Assumptions

Certain assumptions must be meet in order to use a linear regression model for proper interpretation of the data. The best method to view these assumptions is with graphical representation.

One assumption is that the residuals are normally distributed. Residuals are the errors of the prediction by the model from the actual values. If the residuals are normally distributed, their data distribution should match a theoretical distribution and thus, form a straight line when plotted. The graph is a quantile-quantile or Q-Q Plot. The Q-Q plot for Model 3 revealed that this model met the assumption that the residuals were normally distributed (Figure 11). Likewise, a histogram of the residuals revealed the distribution of the residuals. The residuals from Model 3 had a normal distribution. (Figure 12)

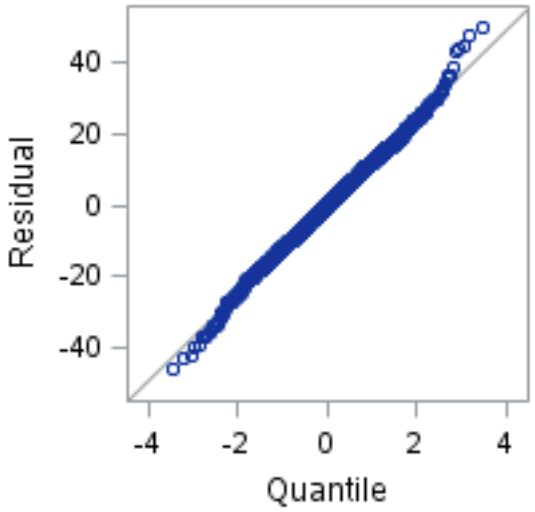


Figure 11: Q-Q Plot for Model 3 revealing that the residuals have a normal distribution.

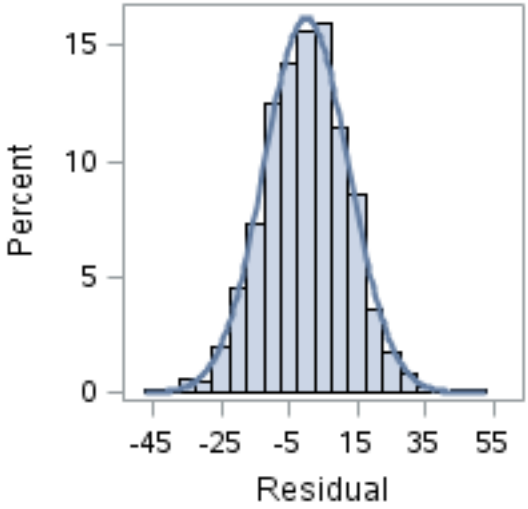


Figure 12. Normal distribution of residuals from Model 3.

Another assumption for linear regression is that the residuals are uncorrelated with the predictor variables. This is best viewed by plotting the residuals against each predictor variable. The scatter of points should be random not to violate this assumption. Any pattern would indicate that this assumption was violated. There was no discernable pattern in the residual by predictors for Model 3 (Figure 13).

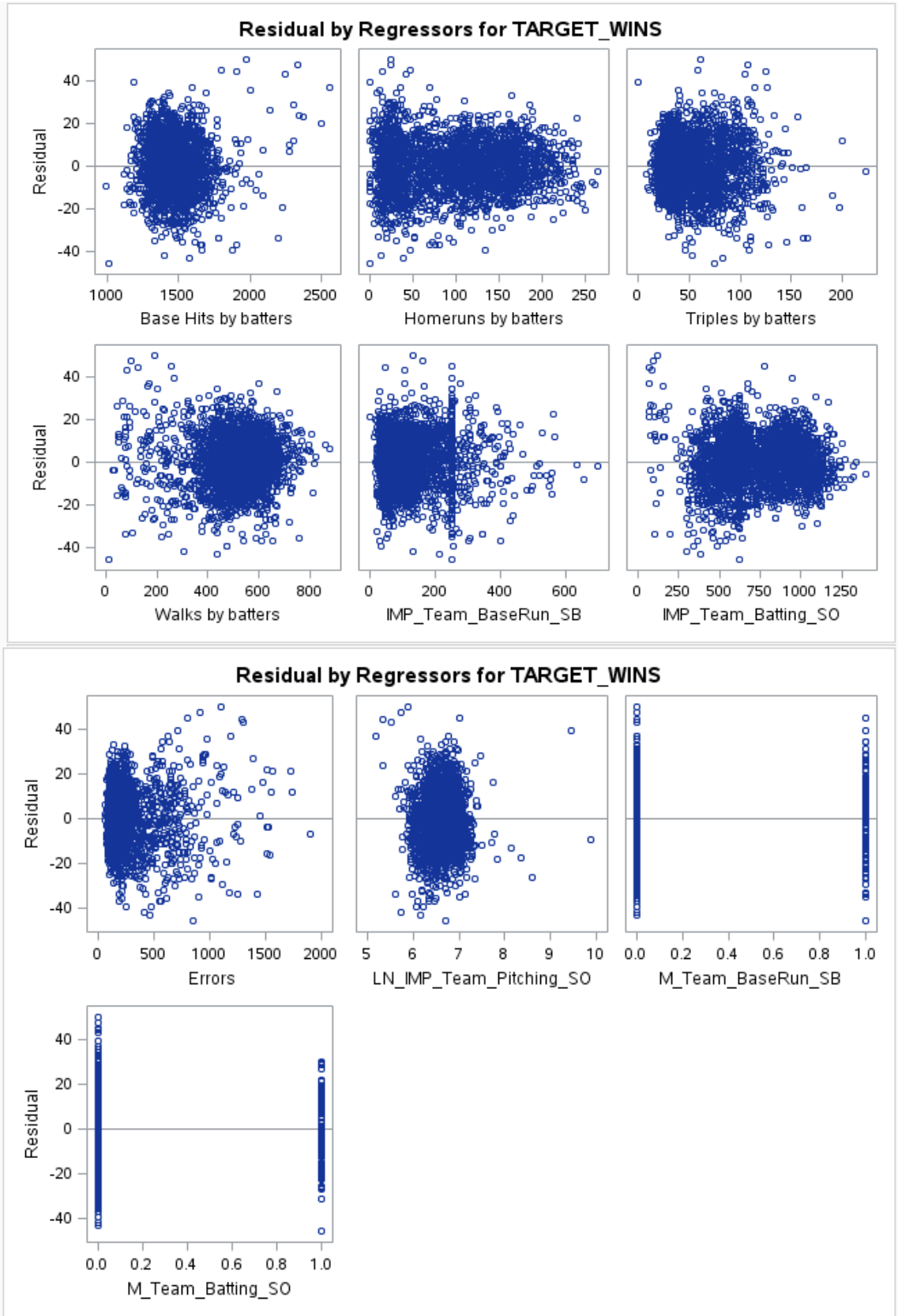


Figure 13: Residuals Plotted Against Predictor Variables for Model 3.

A third assumption is that the residuals or standardized residuals, RStudent (mean value of all the residual is substracted from the residual and this value is then divided by the standard deviation of all the residuals) are uncorrelated to the predicted values. Again, the plot of residuals or RStudent to predicted values should reveal a random scatter of points and display no pattern. These plots for Model 3 revealed no pattern (Figure 14).

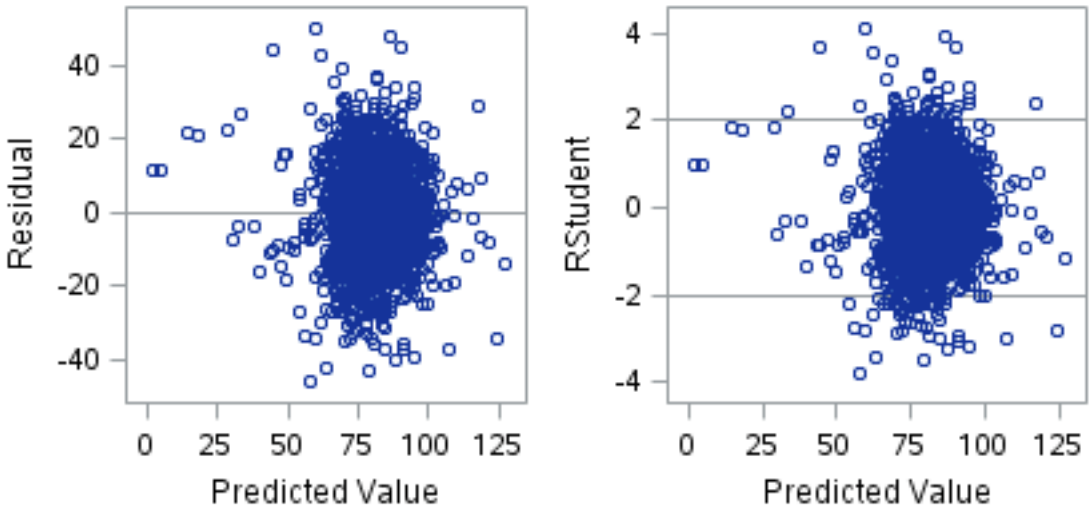


Figure 14. Residuals and Standardized Residuals Plotted Against Predicted Values.

These plots revealed that Model 3 did not violate any major model assumptions. Finally, a plot of the actual Target_Wins to the predicted Target_wins from Model 3 revealed a linear relationship between the values (Figure 15). Model 3 was a good fit for the data. Model 3 was the model entered into the contest to Predict_Target_Wins.

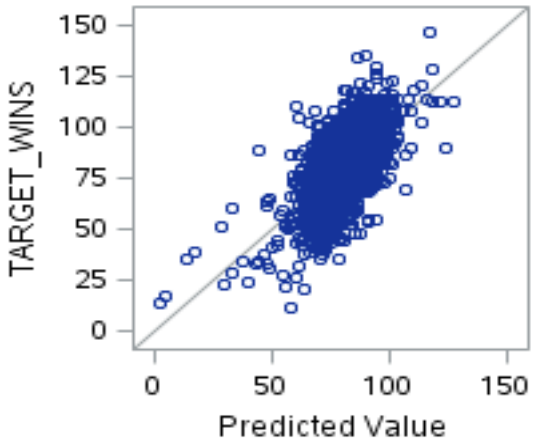


Figure 15. Model 3's Predicted Target Wins have a linear relationship with the actual Target_Wins.

Conclusion

After exploring and preparing a data set for baseball teams from 1871 to 2006, three linear regression models were built to predict the number of wins in a season for professional baseball teams. A model was selected that had fewer variables than the other models and met all the assumptions for linear regression analysis. Even though the selected model had slightly less strength in prediction than another model, the selected model met common sense rules for baseball and was easily deployed. This model was believable.

Bingo Bonus Points:

Using Proc GLM

The SAS results using Proc GLM and the same variables as in Model 3 resulted in exactly the same betas or coefficients including the intercept. Likewise, the ANOVA table was exactly the same. The diagnostic graphs were the same in both the Proc Reg and Proc GLM procedures. These results are not surprising since PROC GLM use the method of least squares to fit general linear models as does PROC REG.

PROC GLM analyzes data using the concept of generalized linear models. This procedure allows for looking for interactions between variables (the presence of one variable influences the effect of another variable in the model).

The Proc GLM has a Type 1 sequential sums of squares procedure. A type 1SS value is the incremental improvement in error sums of squares for each variable as that variable is entered into the model. Also Proc GLM has a Type 3 analysis or sums of squares not depending on the order of how the variables were entered into the model.

Best Model with Proc GLM

The GLM Procedure
Dependent Variable: TARGET_WINS

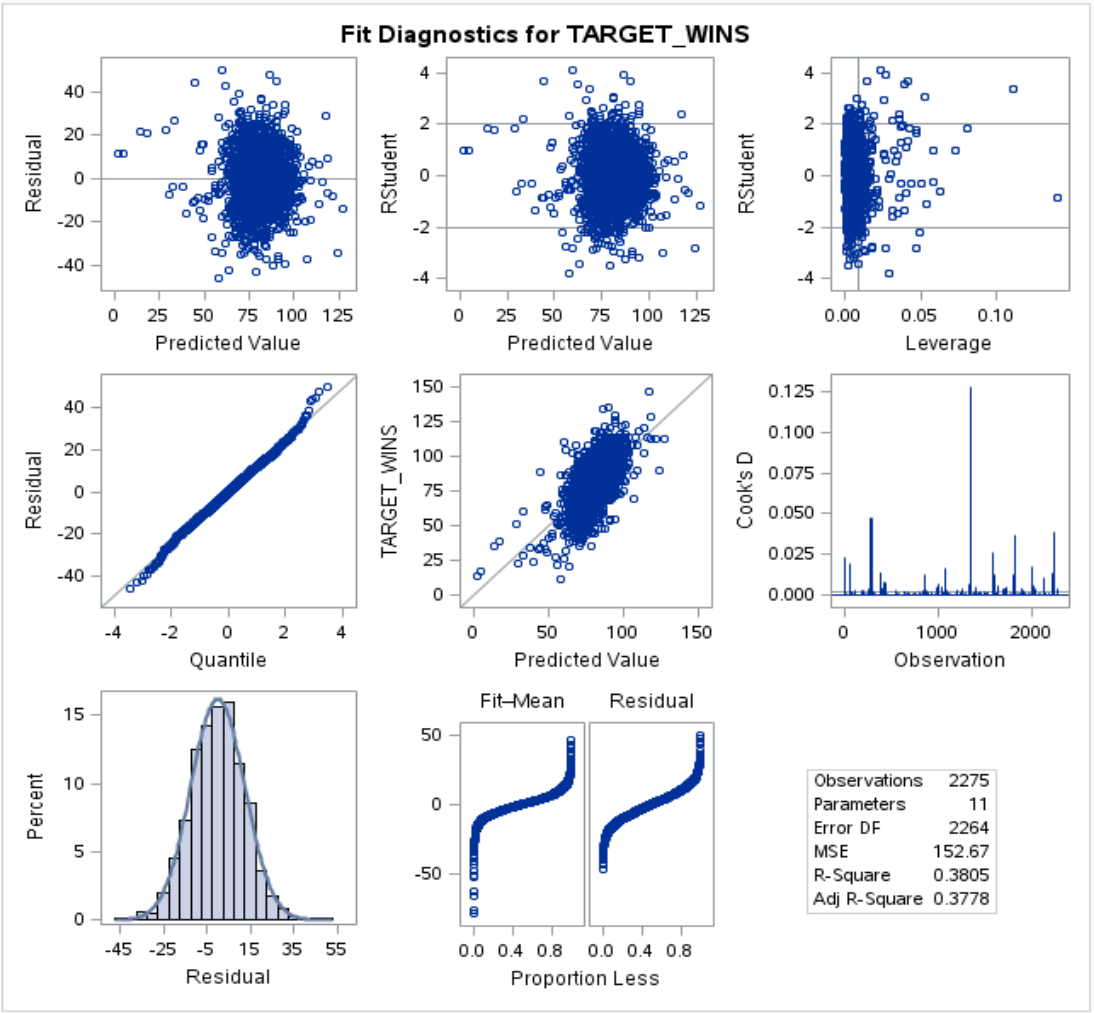
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	212316.5735	21231.6573	139.07	<.0001
Error	2264	345649.8441	152.6722		
Corrected Total	2274	557966.4176			

R-Square	Coeff Var	Root MSE	TARGET_WINS Mean
0.380519	15.28716	12.35606	80.82637

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TEAM_BATTING_H	1	81981.51909	81981.51909	536.98	<.0001
TEAM_BATTING_HR	1	17518.60644	17518.60644	114.75	<.0001
TEAM_BATTING_3B	1	10389.46391	10389.46391	68.05	<.0001
TEAM_BATTING_BB	1	17738.43021	17738.43021	116.19	<.0001
IMP_Team_BaseRun_SB	1	12087.45336	12087.45336	79.17	<.0001
IMP_Team_Batting_SO	1	472.51402	472.51402	3.09	0.0787
TEAM_FIELDING_E	1	33193.56064	33193.56064	217.42	<.0001
LN_IMP_Team_Pitching	1	8943.40654	8943.40654	58.58	<.0001
M_Team_BaseRun_SB	1	22339.87663	22339.87663	146.33	<.0001
M_Team_Batting_SO	1	7651.74264	7651.74264	50.12	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TEAM_BATTING_H	1	39619.98367	39619.98367	259.51	<.0001
TEAM_BATTING_HR	1	7937.47545	7937.47545	51.99	<.0001
TEAM_BATTING_3B	1	2285.32445	2285.32445	14.97	0.0001
TEAM_BATTING_BB	1	5470.57394	5470.57394	35.83	<.0001
IMP_Team_BaseRun_SB	1	28051.77800	28051.77800	183.74	<.0001
IMP_Team_Batting_SO	1	4635.35323	4635.35323	30.36	<.0001
TEAM_FIELDING_E	1	56892.58278	56892.58278	372.65	<.0001
LN_IMP_Team_Pitching	1	838.86901	838.86901	5.49	0.0192
M_Team_BaseRun_SB	1	25844.86270	25844.86270	169.28	<.0001
M_Team_Batting_SO	1	7651.74264	7651.74264	50.12	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-15.06698776	11.79690422	-1.28	0.2017
TEAM_BATTING_H	0.04230686	0.00262624	16.11	<.0001
TEAM_BATTING_HR	0.06771688	0.00939152	7.21	<.0001
TEAM_BATTING_3B	0.06036848	0.01560329	3.87	0.0001
TEAM_BATTING_BB	0.01918161	0.00320441	5.99	<.0001
IMP_Team_BaseRun_SB	0.05571926	0.00411060	13.56	<.0001
IMP_Team_Batting_SO	-0.01807820	0.00328091	-5.51	<.0001
TEAM_FIELDING_E	-0.04387948	0.00227307	-19.30	<.0001
LN_IMP_Team_Pitching	4.35369259	1.85733677	2.34	0.0192
M_Team_BaseRun_SB	21.85883327	1.68004082	13.01	<.0001
M_Team_Batting_SO	9.02765388	1.27518940	7.08	<.0001



Best Model using Proc GENMOD

Using the Proc Genmod with variables in Model 3 revealed that the parameters are the same as for Proc Reg, but the Standard errors are slightly different. Since Proc Genmod uses the maximum likelihood method to determine the coefficients, a Wald Chi-Square is used and the P values are slightly different between the two Proc procedures. The maximum likelihood estimation is an iterative fitting process.

Proc Genmod can perform a generalized linear model on many different probability distributions including normal, binomial, Poisson, gamma, geometric, multinomial and others.

The diagnostic plots are different. Proc Genmod plots are DFBETA per observation. This gives an excellent view of what observation has the largest error and leverage. This is excellent to focus on what observations may need some attention. DFBETA is a method of detecting outliers. DFBETA is the difference between the coefficient for a variable calculated with all the data and the coefficient for a variable calculated when that observation is not included. These errors can be standardized or not.

Best Model with Proc GenMod

The GENMOD Procedure

Model Information	
Data Set	WORK.ADD_MONEYBALL
Distribution	Normal
Link Function	Identity
Dependent Variable	TARGET_WINS

Number of Observations Read	2275
Number of Observations Used	2275

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	2264	345649.8441	152.6722
Scaled Deviance	2264	2275.0000	1.0049
Pearson Chi-Square	2264	345649.8441	152.6722
Scaled Pearson X2	2264	2275.0000	1.0049
Log Likelihood		-8942.2552	
Full Log Likelihood		-8942.2552	
AIC (smaller is better)		17908.5104	
AICC (smaller is better)		17908.6484	
BIC (smaller is better)		17977.2673	

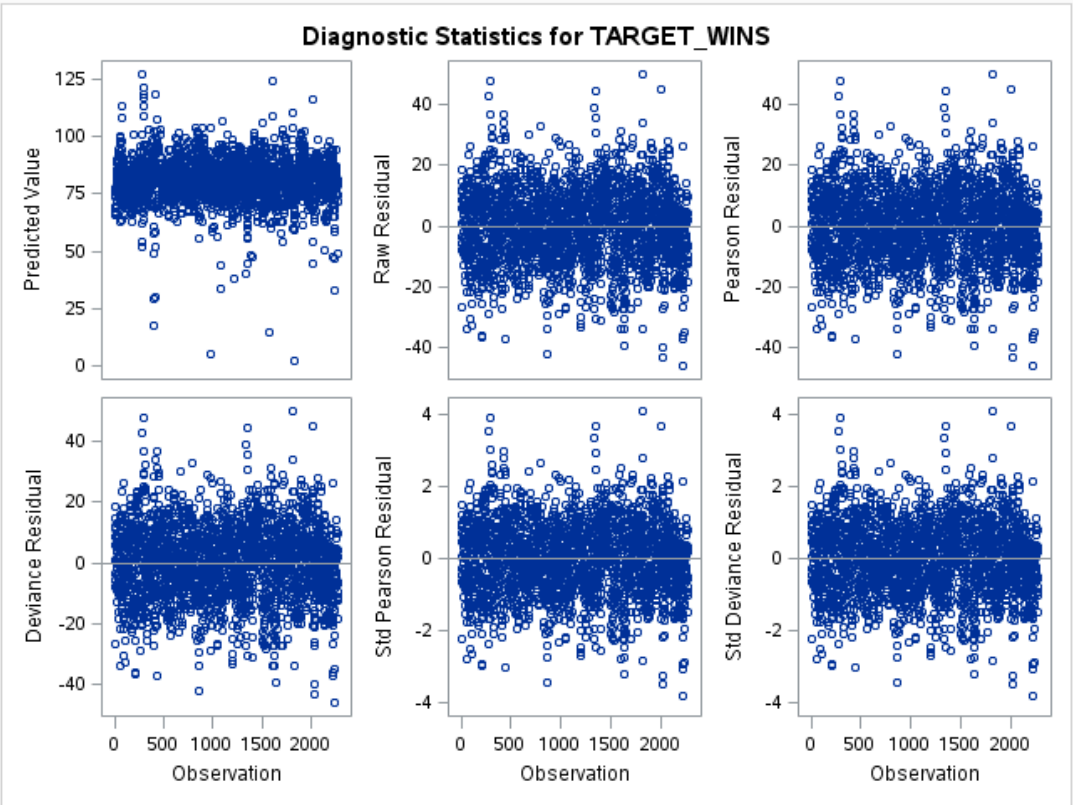
Algorithm converged.

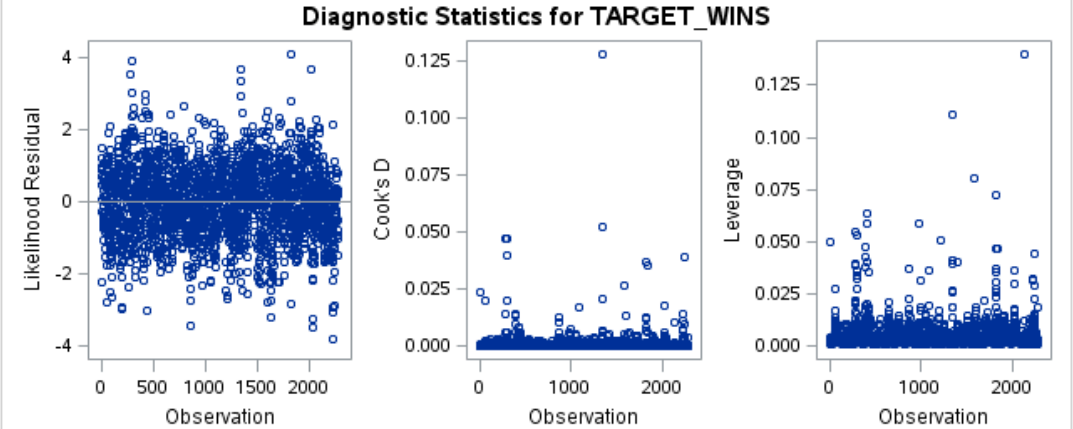
Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.0670	11.7683	-38.1325	7.9986	1.64	0.2004
TEAM_BATTING_H	1	0.0423	0.0026	0.0372	0.0474	260.77	<.0001
TEAM_BATTING_HR	1	0.0677	0.0094	0.0494	0.0861	52.24	<.0001
TEAM_BATTING_3B	1	0.0604	0.0156	0.0299	0.0909	15.04	0.0001
TEAM_BATTING_BB	1	0.0192	0.0032	0.0129	0.0254	36.01	<.0001
IMP_Team_BaseRun_SB	1	0.0557	0.0041	0.0477	0.0638	184.63	<.0001
IMP_Team_Batting_SO	1	-0.0181	0.0033	-0.0245	-0.0117	30.51	<.0001
TEAM_FIELDING_E	1	-0.0439	0.0023	-0.0483	-0.0394	374.46	<.0001
LN_IMP_Team_Pitching	1	4.3537	1.8528	0.7222	7.9852	5.52	0.0188
M_Team_BaseRun_SB	1	21.8588	1.6760	18.5740	25.1437	170.11	<.0001
M_Team_Batting_SO	1	9.0277	1.2721	6.5344	11.5209	50.36	<.0001
Scale	1	12.3262	0.1827	11.9732	12.6896		

Note: The scale parameter was estimated by maximum likelihood.

Best Model with Proc GenMod

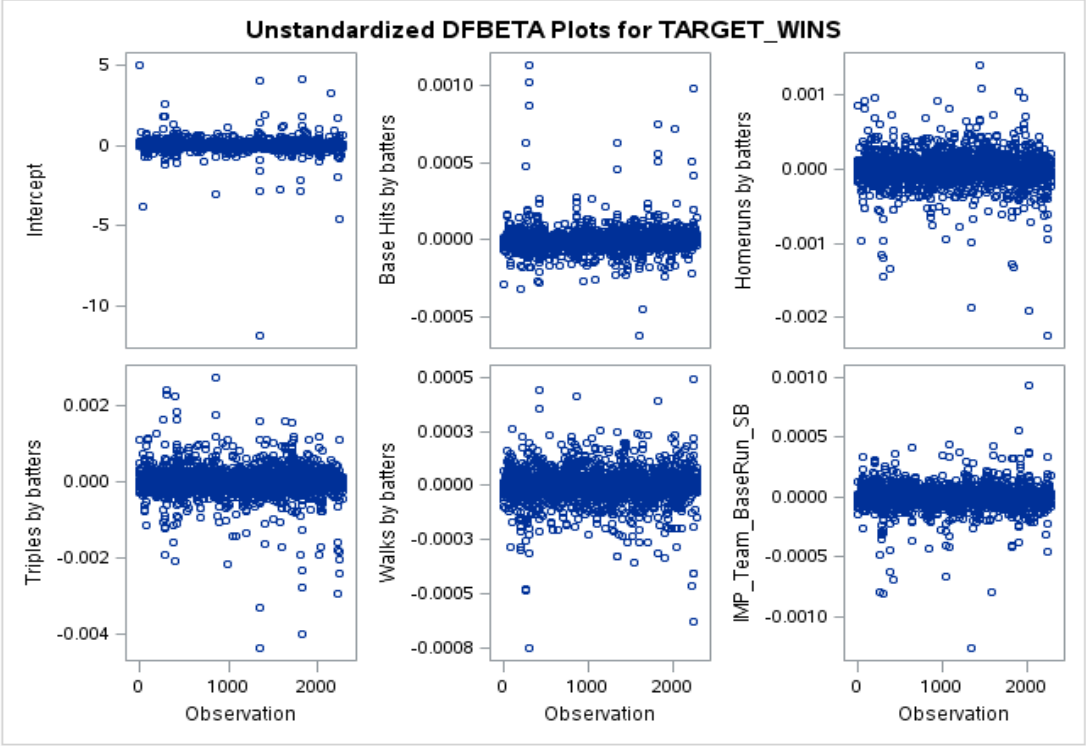
The GENMOD Procedure

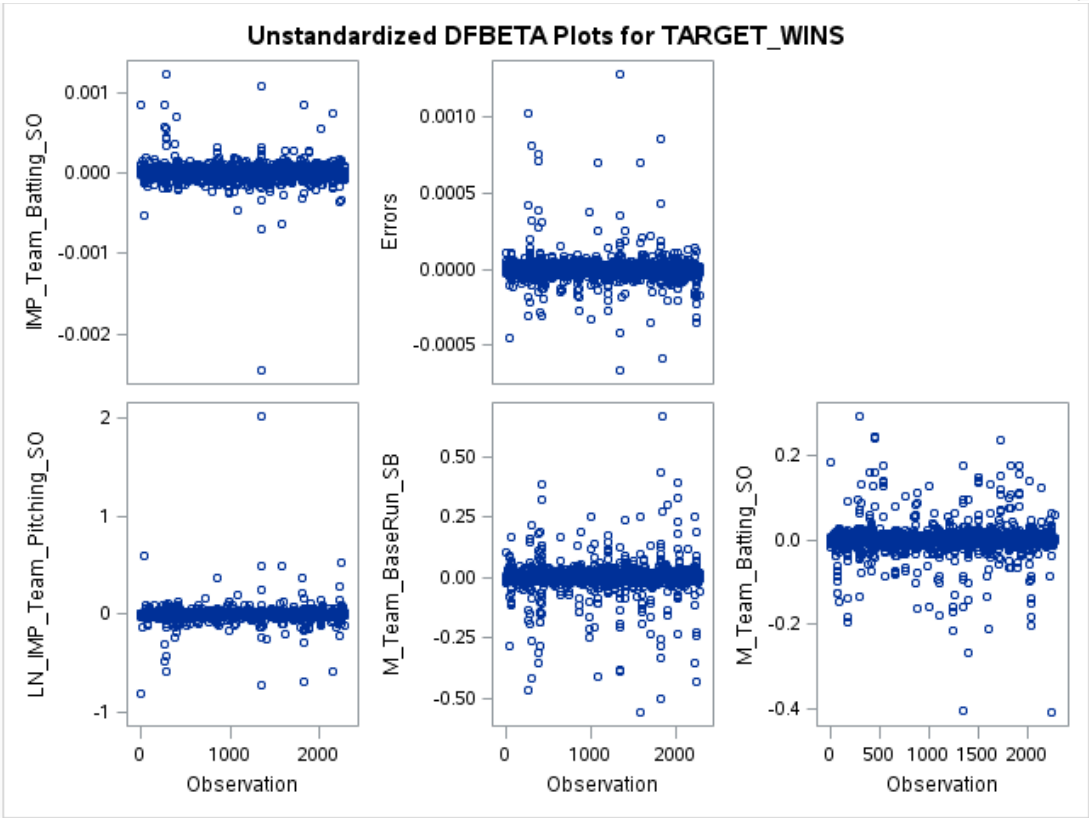




Best Model with Proc GenMod

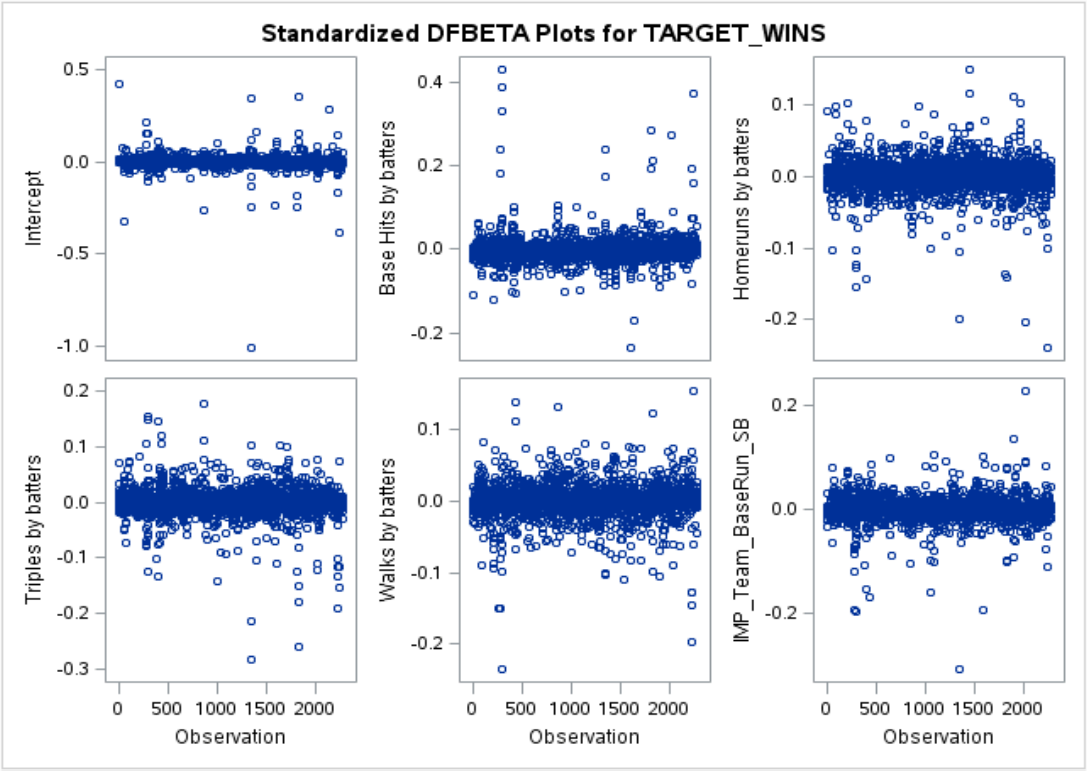
The GENMOD Procedure

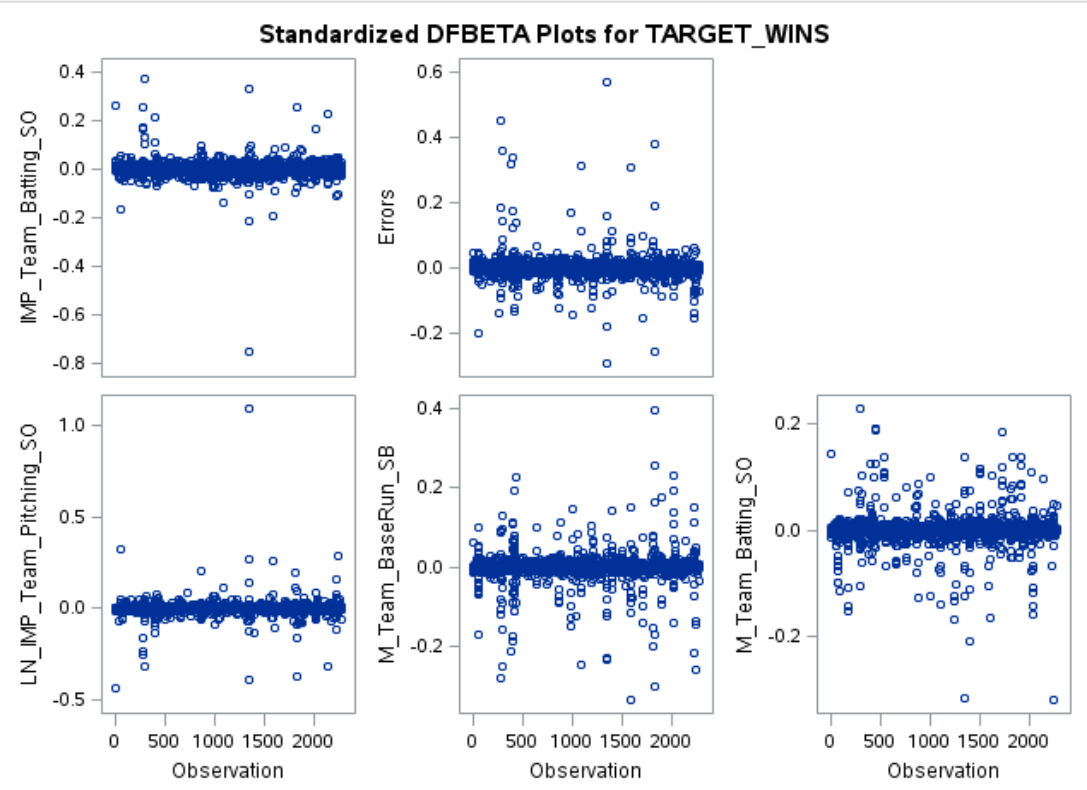


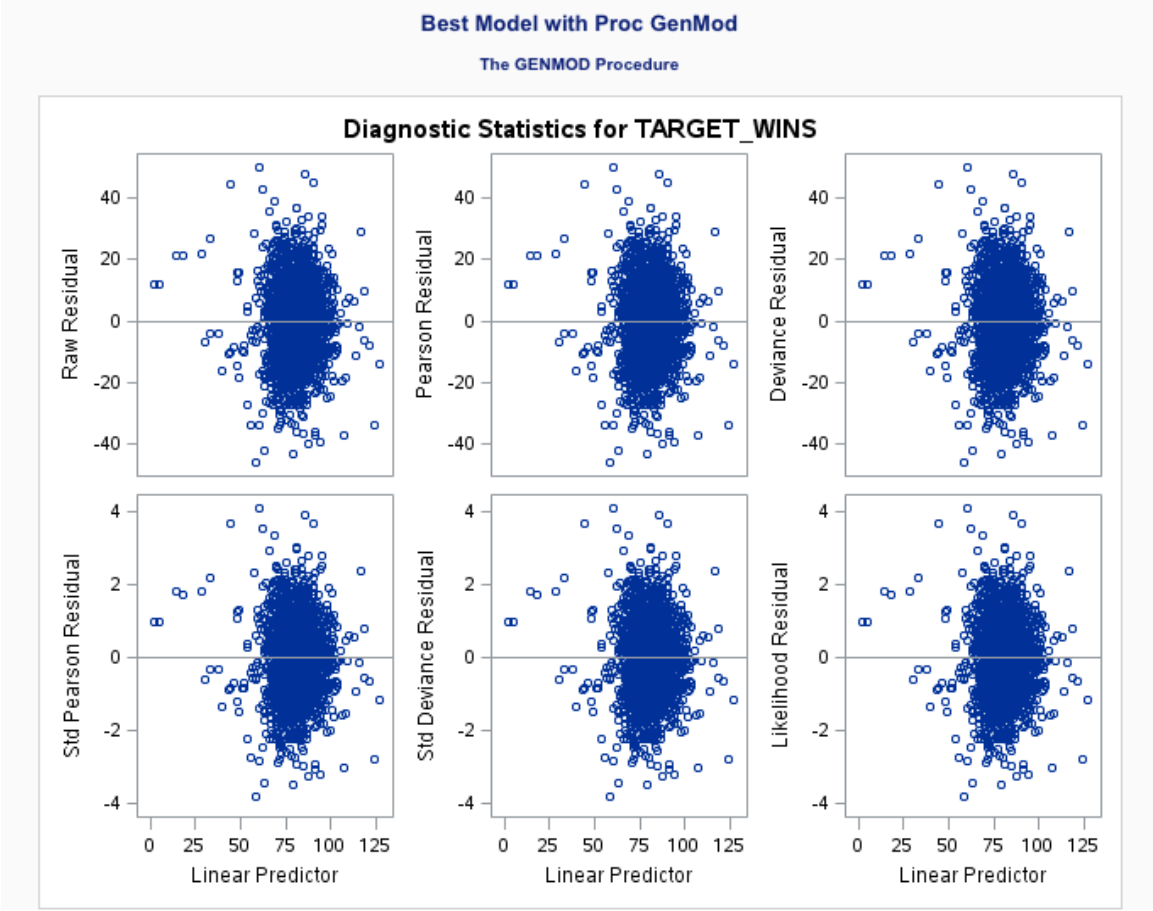


Best Model with Proc GenMod

The GENMOD Procedure






Best Model with Proc GenMod
The GENMOD Procedure
Diagnostic Statistics for TARGET_WINS
Raw Residual
Pearson Residual
Deviance Residual
Std Pearson Residual
Std Deviance Residual
Likelihood Residual
Linear Predictor
Linear Predictor
Linear Predictor

Programming Moneyball in R

Code:

```
> moneyball <- read.csv("~/Desktop/moneyball.dat")
>   View(moneyball)
> attach (moneyball)
> summary (moneyball)
```

Summary Data

```
     INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
Median :1270.5   Median : 82.00   Median :1454   Median :238.0   Median : 47.00
Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00

 TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_BASERUN_CS
Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0   Min.   :  0.0
1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0   1st Qu.: 38.0
Median :102.00   Median :512.0   Median : 750.0   Median :101.0   Median : 49.0
Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8   Mean   : 52.8
3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0   3rd Qu.: 62.0
Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0   Max.   :201.0
                                 NA's   :102      NA's   :131     NA's   :772
TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
Min.   :29.00    Min.   : 1137   Min.   :  0.0    Min.   :   0.0   Min.   :    0.0
1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  615.0
Median :58.00    Median : 1518   Median :107.0    Median : 536.5   Median :  813.5
Mean   :59.36    Mean   : 1779   Mean   :105.7    Mean   : 553.0   Mean   :  817.7
3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  968.0
Max.   :95.00    Max.   :30132   Max.   :343.0    Max.   :3645.0   Max.   :19278.0
NA's   :2085                                                      NA's   :102
TEAM_FIELDING_E  TEAM_FIELDING_DP
Min.   :  65.0   Min.   : 52.0
1st Qu.: 127.0   1st Qu.:131.0
Median : 159.0   Median :149.0
Mean   : 246.5   Mean   :146.4
3rd Qu.: 249.2   3rd Qu.:164.0
Max.   :1898.0   Max.   :228.0
                 NA's   :286
```

From this data, Team_Batting_SO has 102 missing values, Team_Baserun_SB has 131 missing values, Team_Baserun_CS has 772 missing values, Team_Batting_HBP has 2085 missing values, Team_Pitching_SO has 102 missing values, and Team_Fielding_DP has 286 missing values.

TEAM_BATTING_HBP was removed.

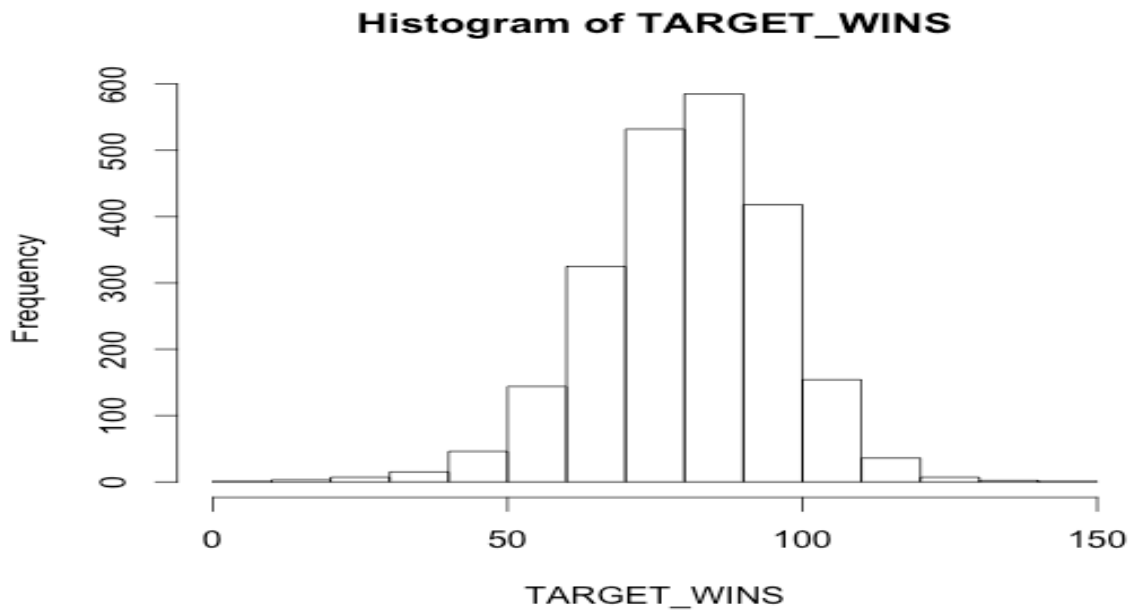
```
> moneyball$TEAM_BATTING_HBP <- NULL
     INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
Median :1270.5   Median : 82.00   Median :1454   Median :238.0   Median : 47.00
Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00

 TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_BASERUN_CS
Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0   Min.   :  0.0
1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0   1st Qu.: 38.0
Median :102.00   Median :512.0   Median : 750.0   Median :101.0   Median : 49.0
Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8   Mean   : 52.8
3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0   3rd Qu.: 62.0
Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0   Max.   :201.0
                                 NA's   :102      NA's   :131     NA's   :772
TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E
Min.   : 1137   Min.   :  0.0    Min.   :   0.0   Min.   :    0.0   Min.   :  65.0
1st Qu.: 1419   1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0
Median : 1518   Median :107.0    Median : 536.5   Median :  813.5   Median : 159.0
Mean   : 1779   Mean   :105.7    Mean   : 553.0   Mean   :  817.7   Mean   : 246.5
3rd Qu.: 1682   3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2
Max.   :30132   Max.   :343.0    Max.   :3645.0   Max.   :19278.0   Max.   :1898.0
                                                  NA's   :102

TEAM_FIELDING_DP
Min.   : 52.0
1st Qu.:131.0
Median :149.0
Mean   :146.4
3rd Qu.:164.0
Max.   :228.0
NA's   :286
```

Since Team_Baserun_CS also has many missing variables,this variable was also eliminated.

```
> moneyball$TEAM_BASERUN_CS <-NULL
> summary (moneyball)
     INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
 Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
 1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
 Median :1270.5   Median : 82.00   Median :1454   Median :238.0   Median : 47.00
 Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
 3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
 Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00

 TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_PITCHING_H
 Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0   Min.   : 1137
 1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0   1st Qu.: 1419
 Median :102.00   Median :512.0   Median : 750.0   Median :101.0   Median : 1518
 Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8   Mean   : 1779
 3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0   3rd Qu.: 1682
 Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0   Max.   :30132
                                  NA's   :102      NA's   :131
 TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
 Min.   :  0.0    Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
 1st Qu.: 50.0    1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
 Median :107.0    Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0
 Mean   :105.7    Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
 3rd Qu.:150.0    3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
 Max.   :343.0    Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
                                   NA's   :102                        NA's   :286
```

In order to perform linear regression, the dependent variable or TARGET_WINS should have a normal distribution.

```
> hist (TARGET_WINS)
```



The dependent variable has basically a normal distribution and thus, linear regression can be used as a predictor model.

```
> cor(moneyball)
                       INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B
INDEX             1.000000000 -0.02105643   -0.017920241      0.01118301
TARGET_WINS      -0.021056435  1.00000000    0.388767521      0.28910365
TEAM_BATTING_H   -0.017920241  0.38876752    1.000000000      0.56284968
TEAM_BATTING_2B   0.011183013  0.28910365    0.562849678      1.00000000
TEAM_BATTING_3B  -0.005814683  0.14260841    0.427696575     -0.10730582
TEAM_BATTING_HR   0.051481047  0.17615320   -0.006544685      0.43539729
TEAM_BATTING_BB  -0.026567236  0.23255986   -0.072464013      0.25572610
TEAM_BATTING_SO           NA          NA             NA              NA
TEAM_BASERUN_SB           NA          NA             NA              NA
TEAM_PITCHING_H   0.017103148 -0.10993705    0.302693709      0.02369219
TEAM_PITCHING_HR  0.050985897  0.18901373    0.072853119      0.45455082
TEAM_PITCHING_BB -0.015287513  0.12417454    0.094193027      0.17805420
TEAM_PITCHING_SO          NA          NA             NA              NA
TEAM_FIELDING_E  -0.009233126 -0.17648476    0.264902478     -0.23515099
TEAM_FIELDING_DP          NA          NA             NA              NA
                 TEAM_BATTING_3B TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO
INDEX              -0.005814683     0.051481047     -0.02656724              NA
TARGET_WINS         0.142608411     0.176153200      0.23255986              NA
TEAM_BATTING_H      0.427696575    -0.006544685     -0.07246401              NA
TEAM_BATTING_2B    -0.107305824     0.435397293      0.25572610              NA
TEAM_BATTING_3B     1.000000000    -0.635566946     -0.28723584              NA
TEAM_BATTING_HR    -0.635566946     1.000000000      0.51373481              NA
TEAM_BATTING_BB    -0.287235841     0.513734810      1.00000000              NA
TEAM_BATTING_SO              NA              NA              NA               1
TEAM_BASERUN_SB              NA              NA              NA              NA
TEAM_PITCHING_H     0.194879411    -0.250145481     -0.44977762              NA
TEAM_PITCHING_HR   -0.567836679     0.969371396      0.45955207              NA
TEAM_PITCHING_BB   -0.002224148     0.136927564      0.48936126              NA
TEAM_PITCHING_SO             NA              NA              NA              NA
TEAM_FIELDING_E     0.509778447    -0.587339098     -0.65597081              NA
TEAM_FIELDING_DP             NA              NA              NA              NA
                 TEAM_BASERUN_SB TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
INDEX                         NA      0.01710315       0.05098590     -0.015287513
TARGET_WINS                   NA     -0.10993705       0.18901373      0.124174536
TEAM_BATTING_H                NA      0.30269371       0.07285312      0.094193027
TEAM_BATTING_2B               NA      0.02369219       0.45455082      0.178054204
TEAM_BATTING_3B               NA      0.19487941      -0.56783668     -0.002224148
TEAM_BATTING_HR               NA     -0.25014548       0.96937140      0.136927564
TEAM_BATTING_BB               NA     -0.44977762       0.45955207      0.489361263
TEAM_BATTING_SO               NA              NA              NA              NA
TEAM_BASERUN_SB                1              NA              NA              NA
TEAM_PITCHING_H               NA      1.00000000      -0.14161276      0.320676162
TEAM_PITCHING_HR              NA     -0.14161276       1.00000000      0.221937505
TEAM_PITCHING_BB              NA      0.32067616       0.22193750      1.000000000
TEAM_PITCHING_SO              NA              NA              NA              NA
TEAM_FIELDING_E               NA      0.66775901      -0.49314447     -0.022837561
TEAM_FIELDING_DP              NA              NA              NA              NA
```

	TEAM_PITCHING_SO	TEAM_FIELDING_E	TEAM_FIELDING_DP
INDEX	NA	-0.009233126	NA
TARGET_WINS	NA	-0.176484759	NA
TEAM_BATTING_H	NA	0.264902478	NA
TEAM_BATTING_2B	NA	-0.235150986	NA
TEAM_BATTING_3B	NA	0.509778447	NA
TEAM_BATTING_HR	NA	-0.587339098	NA
TEAM_BATTING_BB	NA	-0.655970815	NA
TEAM_BATTING_SO	NA	NA	NA
TEAM_BASERUN_SB	NA	NA	NA
TEAM_PITCHING_H	NA	0.667759010	NA
TEAM_PITCHING_HR	NA	-0.493144466	NA
TEAM_PITCHING_BB	NA	-0.022837561	NA
TEAM_PITCHING_SO	1	NA	NA
TEAM_FIELDING_E	NA	1.000000000	NA
TEAM_FIELDING_DP	NA	NA	1

The correlation revealed that Team_Batting_HR and Team_Pitching_HR was correlated with a value of 0.97. One of this variables should be removed.

```
> lm(formula = TARGET_WINS ~ TEAM_BATTING_HR)

Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_HR)

Coefficients:
    (Intercept)  TEAM_BATTING_HR
       76.22576          0.04583

> lm (formula = TARGET_WINS ~ TEAM_PITCHING_HR)

Call:
lm(formula = TARGET_WINS ~ TEAM_PITCHING_HR)

Coefficients:
     (Intercept)  TEAM_PITCHING_HR
        75.65692          0.04857
```

The Team_Batting_HR has an understandable affect on wins. The more homeruns a team hits, the more wins. However, the Team_Pitching_HR has a confusing affect on wins. The more homeruns a team gives up the more games they win. This makes no baseball sense. Thus, Team_Pitching_HR was removed from the data set.

```
> moneyball$TEAM_PITCHING_HR <-NULL
> summary (moneyball)
     INDEX         TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
 Min.   :   1.0   Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
 1st Qu.: 630.8   1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
 Median :1270.5   Median : 82.00   Median :1454   Median :238.0   Median : 47.00
 Mean   :1268.5   Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
 3rd Qu.:1915.5   3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
 Max.   :2535.0   Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00

 TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB TEAM_PITCHING_H
 Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0   Min.   : 1137
 1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0   1st Qu.: 1419
 Median :102.00   Median :512.0   Median : 750.0   Median :101.0   Median : 1518
 Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8   Mean   : 1779
 3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0   3rd Qu.: 1682
 Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0   Max.   :30132
                                  NA's   :102      NA's   :131
 TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
 Min.   :   0.0   Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
 1st Qu.: 476.0   1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
 Median : 536.5   Median :  813.5   Median : 159.0   Median :149.0
 Mean   : 553.0   Mean   :  817.7   Mean   : 246.5   Mean   :146.4
 3rd Qu.: 611.0   3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
 Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
                  NA's   :102                        NA's   :286
```

```
> summary (lm (TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP))

Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
    TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BASERUN_SB +
    TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E +
    TEAM_FIELDING_DP)

Residuals:
    Min      1Q  Median      3Q     Max
-32.323  -7.236   0.128   7.042  29.769
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      58.4285604  6.0329903   9.685  < 2e-16 ***
TEAM_BATTING_H   -0.0005582  0.0106585  -0.052  0.95824
TEAM_BATTING_2B  -0.0500707  0.0088869  -5.634 2.03e-08 ***
TEAM_BATTING_3B   0.1811378  0.0190036   9.532  < 2e-16 ***
TEAM_BATTING_HR   0.1015750  0.0091610  11.088  < 2e-16 ***
TEAM_BATTING_SO   0.0506646  0.0164782   3.075  0.00214 **
TEAM_BASERUN_SB   0.0701629  0.0055398  12.665  < 2e-16 ***
TEAM_PITCHING_H   0.0284838  0.0092462   3.081  0.00210 **
TEAM_PITCHING_BB  0.0317941  0.0029789  10.673  < 2e-16 ***
TEAM_PITCHING_SO -0.0699744  0.0155588  -4.497 7.31e-06 ***
TEAM_FIELDING_E  -0.1156166  0.0070447 -16.412  < 2e-16 ***
TEAM_FIELDING_DP -0.1127450  0.0122943  -9.170  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.2 on 1823 degrees of freedom
  (441 observations deleted due to missingness)
Multiple R-squared:  0.4032,    Adjusted R-squared:  0.3996
F-statistic: 111.9 on 11 and 1823 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      58.4285604  6.0329903   9.685  < 2e-16 ***
TEAM_BATTING_H   -0.0005582  0.0106585  -0.052  0.95824
TEAM_BATTING_2B  -0.0500707  0.0088869  -5.634 2.03e-08 ***
TEAM_BATTING_3B   0.1811378  0.0190036   9.532  < 2e-16 ***
TEAM_BATTING_HR   0.1015750  0.0091610  11.088  < 2e-16 ***
TEAM_BATTING_SO   0.0506646  0.0164782   3.075  0.00214 **
TEAM_BASERUN_SB   0.0701629  0.0055398  12.665  < 2e-16 ***
TEAM_PITCHING_H   0.0284838  0.0092462   3.081  0.00210 **
TEAM_PITCHING_BB  0.0317941  0.0029789  10.673  < 2e-16 ***
TEAM_PITCHING_SO -0.0699744  0.0155588  -4.497 7.31e-06 ***
TEAM_FIELDING_E  -0.1156166  0.0070447 -16.412  < 2e-16 ***
TEAM_FIELDING_DP -0.1127450  0.0122943  -9.170  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.2 on 1823 degrees of freedom
  (441 observations deleted due to missingness)
Multiple R-squared:  0.4032,    Adjusted R-squared:  0.3996
F-statistic: 111.9 on 11 and 1823 DF,  p-value: < 2.2e-16
```

Since missing values were not added, 441 observations were deleted from the analysis. Note, from this model the more hits a team makes the less games they win and also this variable is not significant. Likewise, the more doubles a team hits the less wins and the more double plays the team turns the less wins. To make a common sense model Team_Batting_2B and Team_Fielding_DP were removed from the model. Team_Batting_H was removed due to this variable not being significant.

```
> summary (lm (TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCH
ING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E))
```

```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_3B + TEAM_BATTING_HR +
    TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_BB +
    TEAM_PITCHING_SO + TEAM_FIELDING_E)

Residuals:
    Min      1Q  Median      3Q     Max
-48.895  -7.867   0.428   7.593  61.633

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.3208111  2.3602973  28.946  < 2e-16 ***
TEAM_BATTING_3B   0.1444595  0.0164628   8.775  < 2e-16 ***
TEAM_BATTING_HR   0.1337949  0.0078855  16.967  < 2e-16 ***
TEAM_BATTING_SO  -0.0074868  0.0042148  -1.776   0.0758 .
TEAM_BASERUN_SB   0.0649836  0.0043739  14.857  < 2e-16 ***
TEAM_PITCHING_H   0.0026842  0.0003694   7.265 5.27e-13 ***
TEAM_PITCHING_BB  0.0137980  0.0025848   5.338 1.04e-07 ***
TEAM_PITCHING_SO -0.0174511  0.0033955  -5.139 3.02e-07 ***
TEAM_FIELDING_E  -0.0498982  0.0032184 -15.504  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.97 on 2034 degrees of freedom
  (233 observations deleted due to missingness)
Multiple R-squared:  0.3225,    Adjusted R-squared:  0.3199
F-statistic:   121 on 8 and 2034 DF,  p-value: < 2.2e-16
```

However, this model has much lower adjusted R-square making it a worst model that the first model. A third model was made only removing the TEAM_BATTING_H that was not significant.

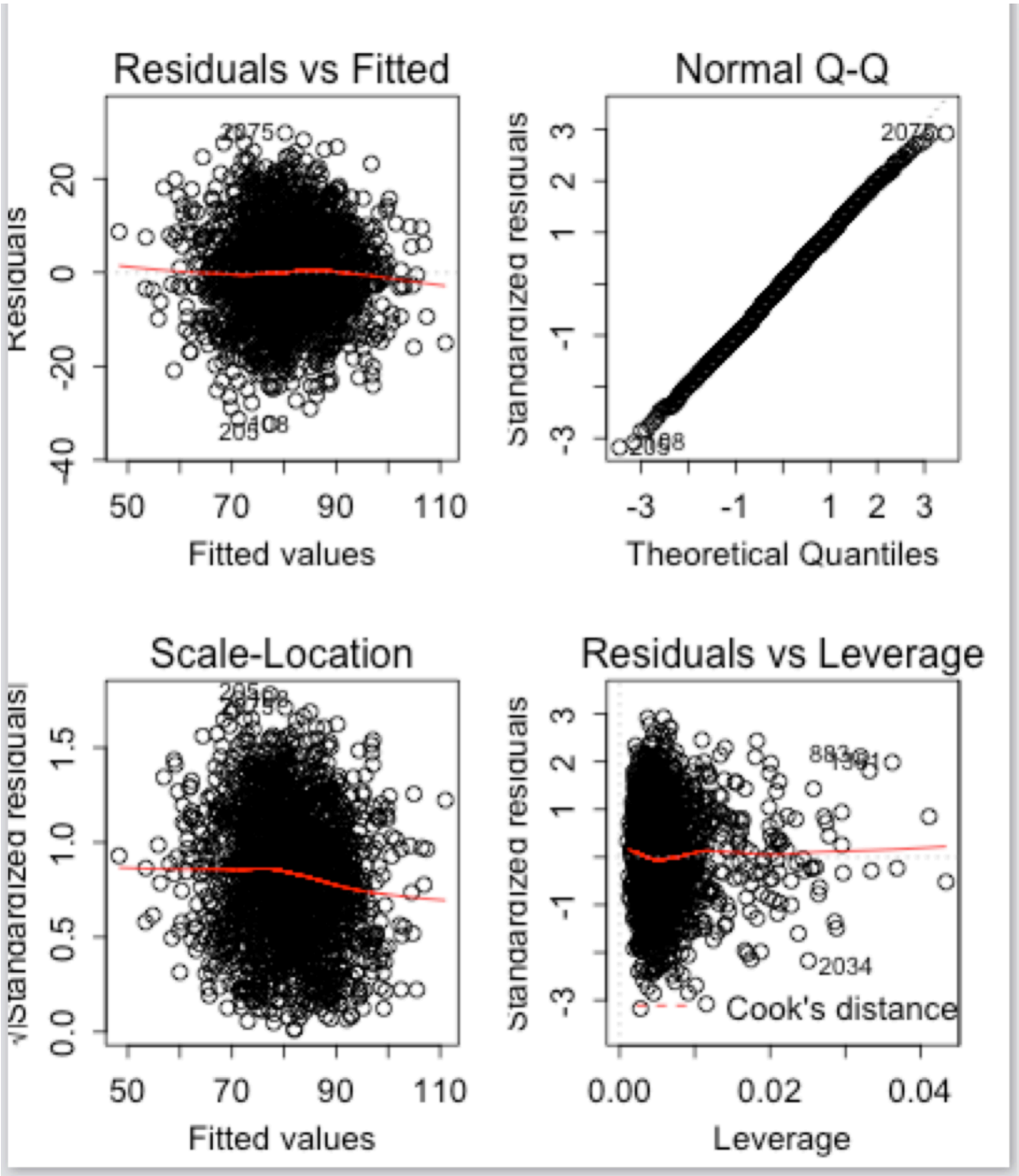
```
> summary (lm (TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BASERU
N_SB + TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP))
```

```
Call:
lm(formula = TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B +
    TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H +
    TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP)

Residuals:
    Min      1Q  Median      3Q     Max
-32.333  -7.254   0.130   7.049  29.770

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      58.321082   5.671656  10.283  < 2e-16 ***
TEAM_BATTING_2B  -0.050202   0.008525  -5.889 4.61e-09 ***
TEAM_BATTING_3B   0.180959   0.018689   9.682  < 2e-16 ***
TEAM_BATTING_HR   0.101474   0.008952  11.336  < 2e-16 ***
TEAM_BATTING_SO   0.049937   0.008847   5.645 1.92e-08 ***
TEAM_BASERUN_SB   0.070169   0.005537  12.672  < 2e-16 ***
TEAM_PITCHING_H   0.028041   0.003760   7.458 1.35e-13 ***
TEAM_PITCHING_BB  0.031806   0.002969  10.713  < 2e-16 ***
TEAM_PITCHING_SO -0.069257   0.007362  -9.408  < 2e-16 ***
TEAM_FIELDING_E  -0.115554   0.006941 -16.648  < 2e-16 ***
TEAM_FIELDING_DP -0.112813   0.012222  -9.230  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.19 on 1824 degrees of freedom
  (441 observations deleted due to missingness)
Multiple R-squared:  0.4032,    Adjusted R-squared:  0.3999
F-statistic: 123.2 on 10 and 1824 DF,  p-value: < 2.2e-16
```

This model is as follows;
Target_Wins = 58.32 + -0.0502 * Team_Batting_2B + 0.181 * Team_Batting_3B + 0.101 * Team_Batting_HR + 0.05 * Team_Batting_SO + 0.07 * Team_Baserun_SB + 0.028 * Team_Pitching_H + 0.032 * Team_Pitching_BB + -0.069 * Team_Pitching_SO + -0.115 * Team_Fielding_E + -0.113 * Team_Fielding_DP

This model was selected and model regression diagnostic curves were generated.

```
> options(na.action="na.exclude")
> lm.velo <- lm (TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BASE
RUN_SB + TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP)
> par (mfrow=c(2,2),mex=0.6)
> plot (lm.velo)
> par(mfrow=c(1,1), mex =1)
```



The residuals and standardized residual to the fitted values do not reveal a pattern. The Normal Q-Q Plot reveals a normal distribution of the residuals. The leverage plot revealed values with some leverage but none greater than 1.

In conclusion, this model was developed excluding all observations with missing values and no transformation of variables. The adjusted R-Square was 0.3999 and the no linear assumptions were violated.