MONEYBALL OLS REGRESSION PROJECT

Created By : Evangelos Giakoumakis, Vishal Ahir

Introduction

Stats and data have always been a huge part of the game of baseball dating back to the 1970's. However, in the last few years the emergence of Sabermetrics Michael Lewis' Moneyball stand as baseball's toe in the data science waters. Essentially, sabermetrics looks at a whole bunch of nontraditional baseball stats and uses them to make player comparisons and, to a degree, predict player performance. In the short 15 years or so since Billy Beane brought the book of Bill James to baseball, data collection and analytics capabilities have grown exponentially and are being used in all industries, with baseball arguably chief among them.

Scope of Project: The reason for this project is to perform OLS (Linear) Regression Analysis on the baseball dataset provided to predict number of wins for each team.

Input: Few pointers about the provided dataset:

- There is a total of 2276 records in the Moneyball dataset with 1 target variable and 15 explanatory variables.
- Each record represents a professional baseball team from the years 1871 to 2006 inclusive. All the statistics provided are for these years combined.
- Each record has the performance of the team for the given year, with all the statistics adjusted to match the performance of a 162 games season.
- Below is the schema of dataset and the effects.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_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

Data Exploration

We will initiate this project by performing exploratory data analysis on the provided dataset. The following steps were performed:

1) Import the data into SAS and ensure all the records (2276) were imported.

The CONTENTS Procedure

Data Set Name	WORK.IMPORT	Observations	2276
Member Type	DATA	Variables	17
Engine	V9	Indexes	0
Created	10/14/2017 08:29:46	Observation Length	136
Last Modified	10/14/2017 08:29:46	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

2) Perform visual analysis of sample records to check formatting and actual values stored.
3) Identify what variables are missing values and at what extent.
4) Check for normality of data for explanatory variables.
5) Check for outliers in all the explanatory variables.
6) See what kind of transformations may be needed to normalize data and remove any outliers.

Let’s deep dive into individual steps for data exploration.

Missing Variables

For every variable, we will mark the missing values with string “Missing” and get a count of those missing values via “PROC FREQ” to get a better idea if we need to impute values for those variables.

Variable	Not Missing	Missing	% Missing
TEAM_BATTING_HBP	191	2085	92
TEAM_BATTING_H	2276	0	0
TEAM_BATTING_2B	2276	0	0
TEAM_BATTING_3B	2276	0	0
TEAM_BATTING_HR	2276	0	0
TEAM_BATTING_BB	2276	0	0
TEAM_BATTING_SO	2174	102	4
TEAM_BASERUN_SB	2145	131	6
TEAM_BASERUN_CS	1504	772	34
TEAM_PITCHING_H	2276	0	0
TEAM_PITCHING_HR	2276	0	0
TEAM_PITCHING_BB	2276	0	0
TEAM_PITCHING_SO	2174	102	4
TEAM_FIELDING_E	2276		0
TEAM_FIELDING_DP	1990	286	13

As we can see above, TEAM_BATTING_HBP has highest and maximum percentage of values missing. We will decide what to do with it Data Prep section.

For rest of the variables that are missing values, it would be useful to try and impute the missing values. We will take care of imputing in Data Prep section.

Outliers

We ran proc univariate on the predictor variables to identify which ones have extreme outliers and see if we can identify the reason behind those obs. Some of these variables may or may not end up in the final model selected but we just wanted to ensure we have this data available as and when needed. Variables with outliers are listed below:

TEAM_BATTING_H

Quantiles (Definition 5)	
Level	Quantile
100% Max	2554.0
99%	1950.0
95%	1696.0
90%	1636.0
75% Q3	1537.5
50% Median	1454.0
25% Q1	1383.0
10%	1315.0
5%	1280.0
1%	1188.0
0% Min	891.0

TEAM_BATTING_3B

Quantiles (Definition 5)	
Level	Quantile
100% Max	223
99%	134
95%	108
90%	96
75% Q3	72
50% Median	47
25% Q1	34
10%	27
5%	23
1%	17
0% Min	0

TEAM_BASERUN_SB

Quantiles (Definition 5)	
Level	Quantile
100% Max	697
99%	439
95%	302
90%	231
75% Q3	156
50% Median	101
25% Q1	66
10%	44
5%	35
1%	23
0% Min	0

TEAM_PITCHING_H

Quantiles (Definition 5)	
Level	Quantile
100% Max	30132
99%	7093
95%	2563
90%	2059
75% Q3	1683
50% Median	1518
25% Q1	1419
10%	1356
5%	1316
1%	1244
0% Min	1137

TEAM_PITCHING_BB

Quantiles (Definition 5)	
Level	Quantile
100% Max	3645.0
99%	924.0
95%	757.0
90%	694.0
75% Q3	611.0
50% Median	536.5
25% Q1	476.0
10%	417.0
5%	377.0
1%	237.0
0% Min	0.0

TEAM_PITCHING_SO

Quantiles (Definition 5)	
Level	Quantile
100% Max	19278.0
99%	1474.0
95%	1173.0
90%	1095.0
75% Q3	968.0
50% Median	813.5
25% Q1	615.0
10%	490.0
5%	420.0
1%	205.0
0% Min	0.0

TEAM_FIELDING_E

Quantiles (Definition 5)	
Level	Quantile
100% Max	1898.0
99%	1237.0
95%	716.0
90%	542.0
75% Q3	249.5
50% Median	159.0
25% Q1	127.0
10%	109.0
5%	100.0
1%	86.0
0% Min	65.0

Data Preparation

After completing the exploratory data analysis, it is now time to start prepping the data so we can use it for fitting any model. Before we model this data, we need to ensure our basic criterion such as normality, independence, constant variance, linearity, etc are met. We will try any possible transformation to ensure the dataset meets these criterion before proceeding with fitting a model.

TEAM_BATTING_HBP: Given that the vast majority of data (91.6%) is missing from Batters hit by pitch, *we decided to remove it* from our analysis. The reason for doing so is that by imputing a set value (mean or median) would homogenize the entire set, causing it to be of little to no significance predicting the target variable.

TEAM_BATTING_SO: Given the fact that only (4.5%) of the data is missing from Strikeouts by batters, *we decided to keep variable and impute* with the mean value (735) rounded down. The reason for doing so is that only a small percentage of data is missing so imputing those missing values with the mean would allow us to use regression analysis while not significantly altering the data.

TEAM_BASERUN_SB: Given the fact that only (5.8%) of the data is missing from Stolen bases, *we decided to keep variable and impute* with the mean value (124) rounded down. The reason for doing so is that only a small percentage of data is missing so imputing those missing values with the mean would allow us to use regression analysis while not significantly altering the data.

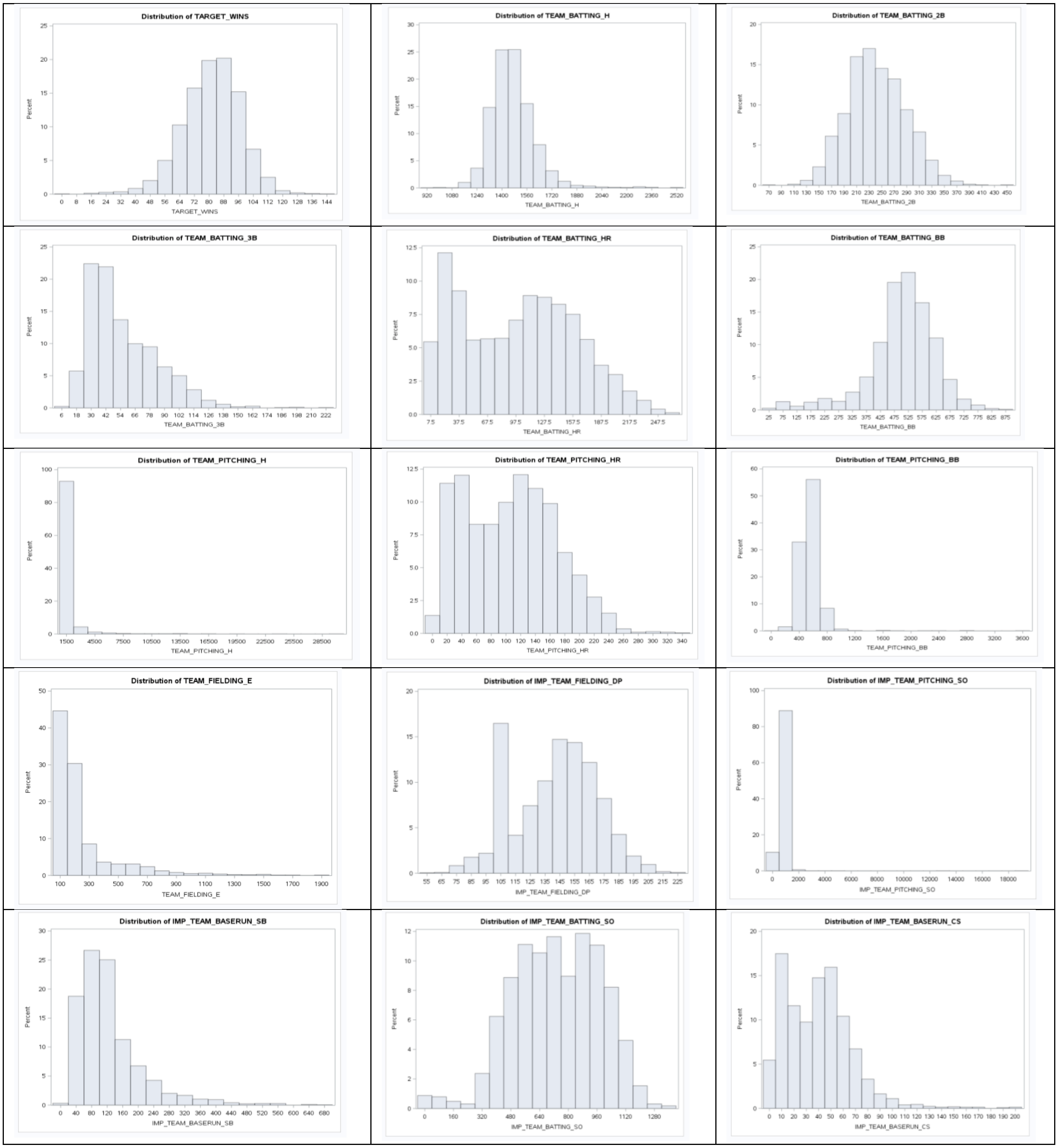
TEAM_BASERUN_CS: Given the fact that a third of the data (34%) is missing from Caught stealing, *we decided to keep variable and impute* with the analogy between average stolen bases and average caught stealing divided. The reason for doing so is that by imputing those missing values would allow us to use regression analysis with this variable. Using the above methodology would give us a more accurate picture of the data than the mean or median would.

TEAM_PITCHING_SO: Given the fact that only (4.5%) of the data is missing from Strikeouts by pitchers, *we decided to keep variable and impute* with mean value (817) rounded down. The reason for doing so is that only a small percentage of data is missing so imputing those missing values with the mean would allow us to use regression analysis while not significantly altering the data.

TEAM_FIELDING_DP: Given the fact that an eighth (12.6%) of the data is missing from Double Plays, *we decided to keep variable and impute* with the mean value (146) rounded down. The reason for doing so is that only a small percentage of data is missing so imputing those missing values with the mean would allow us to use regression analysis while not significantly altering the data.

Transformations

After imputing the missing values, it is time to inspect the distribution of predictor variables to identify the variable that will require transformations. Below are the histograms for predictor variables that will help decide what kind of transformations will be needed that will best benefit overall model fit.



Looking at the various histograms, we have several variables with **heavily skewed** data as well as extreme outliers. The following treatments were applied to deal with these issues:

TEAM_PITCHING_H: We decided to do a reciprocal transformation (-1/x) and place results in a new variable (TRSF_TEAM_PITCHING_H) so that we can get rid of its skewness.

TTEAM_PITCHING_BB: We decided to do a log transformation (e) and place results in a new variable (TRSF_TEAM_PITCHING_BB) so that we can get rid of its skewness.

TEAM_FIELDING_E: We decided to do a log transformation (e) and place results in a new variable (TRSF_TEAM_FIELDING_E) so that we can get rid of its skewness.

IMP_TEAM_PITCHING_SO: We decided to do a log transformation (e) and place results in a new variable (TRSF_IMP_TEAM_PITCHING_SO) so that we can get rid of its skewness.

TEAM_BATTING_3B: We decided to do a log transformation (e) and place results in a new variable (TRSF_ TEAM_BATTING_3B) so that we can get rid of its skewness.

IMP_TEAM_BASERUN_SB: We decided to do a log transformation (e) and place results in a new variable (TRSF_ IMP_TEAM_BASERUN_SB) so that we can get rid of its skewness.

IMP_TEAM_BASERUN_CS: We decided to do a log transformation (e) and place results in a new variable (TRSF_ IMP_TEAM_BASERUN_CS) so that we can get rid of its skewness.

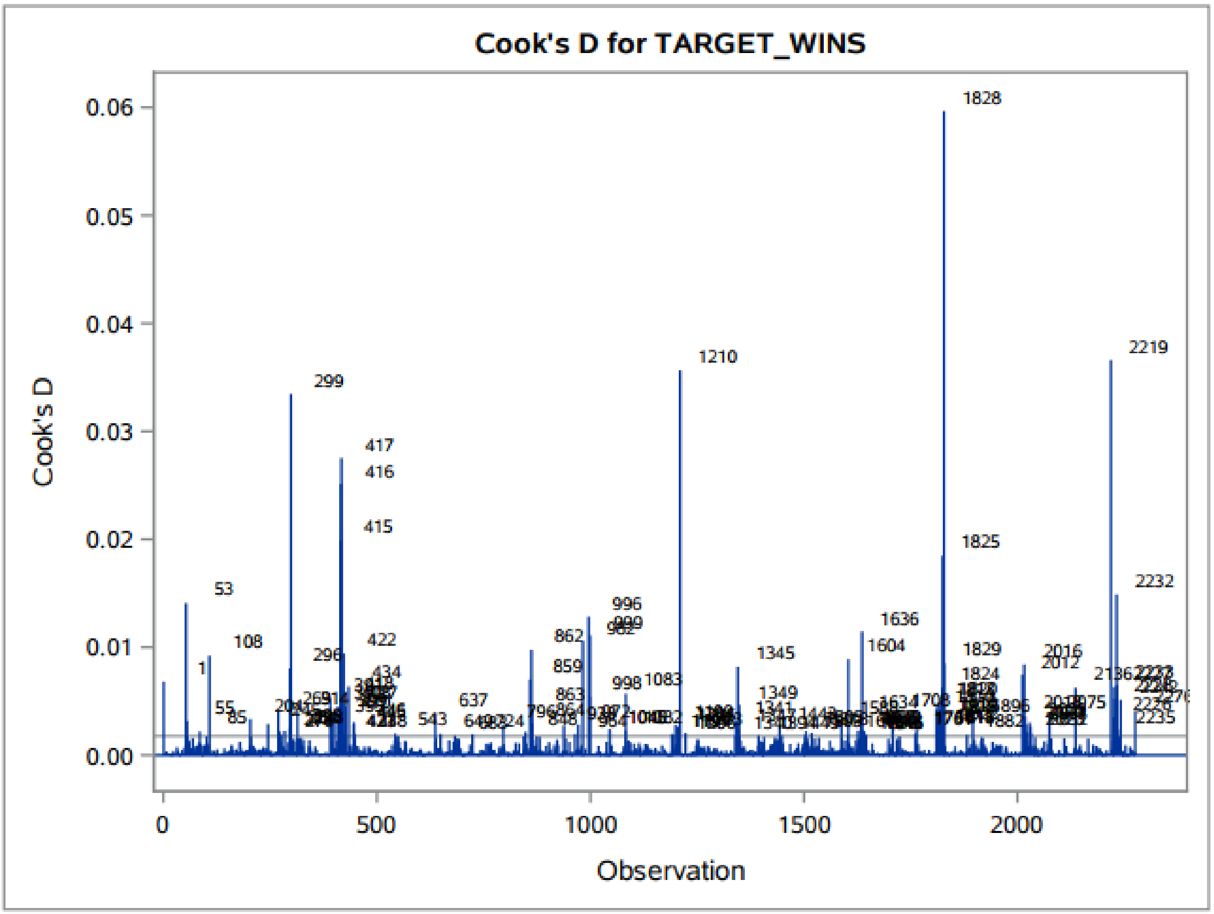
Correlation : Independence among explantory variables is an important metric to ensure there is no collinearity affecting the model fit. As you can see from below tables, none of the variables seem to be highly correlated to each other.

Correlation			
Variable	TEAM_BATTING_H	TEAM_PITCHING_HR	IMP_TEAM_FIELDING_DP
TEAM_BATTING_H	1.0000	0.0699	-0.0466
TEAM_PITCHING_HR	0.0699	1.0000	0.5140
IMP_TEAM_FIELDING_DP	-0.0466	0.5140	1.0000
TRSF_TEAM_PITCHING_H	0.6493	-0.2136	-0.3693
TRSF_TEAM_PITCHING_BB	0.0762	0.3295	0.2150
TRSF_TEAM_BATTING_3B	0.3669	-0.6184	-0.4442
TRSF_IMP_TEAM_BASERUN_SB	0.0716	-0.3714	-0.5264
TRSF_IMP_TEAM_BASERUN_CS	-0.0846	0.3531	0.3226
TARGET_WINS	0.3880	0.1862	-0.0094

Correlation			
Variable	TRSF_TEAM_PITCHING_H	TRSF_TEAM_PITCHING_BB	TRSF_TEAM_BATTING_3B
TEAM_BATTING_H	0.6493	0.0762	0.3669
TEAM_PITCHING_HR	-0.2136	0.3295	-0.6184
IMP_TEAM_FIELDING_DP	-0.3693	0.2150	-0.4442
TRSF_TEAM_PITCHING_H	1.0000	-0.0207	0.4580
TRSF_TEAM_PITCHING_BB	-0.0207	1.0000	-0.0850
TRSF_TEAM_BATTING_3B	0.4580	-0.0850	1.0000
TRSF_IMP_TEAM_BASERUN_SB	0.1614	0.0415	0.3671
TRSF_IMP_TEAM_BASERUN_CS	-0.3365	0.2046	-0.3296
TARGET_WINS	0.0900	0.1619	0.1164

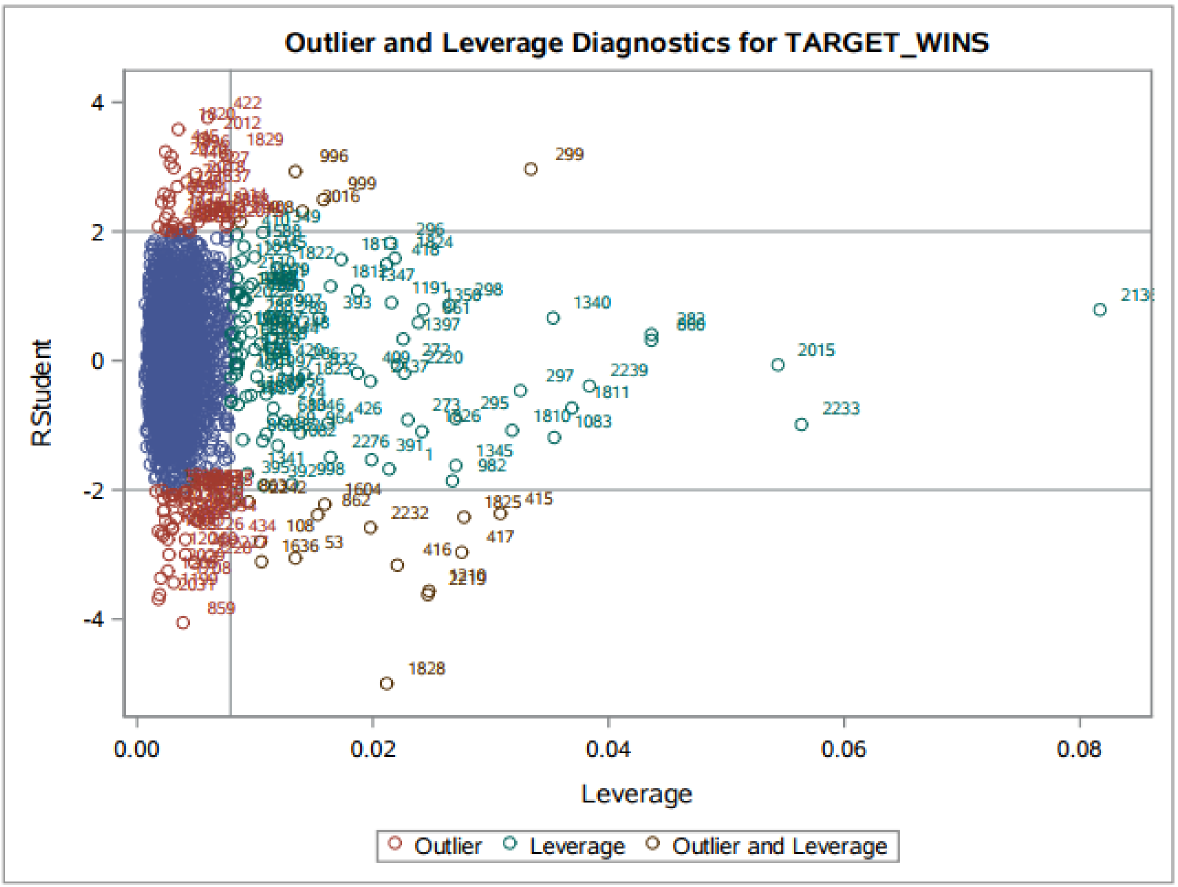
Correlation			
Variable	TRSF_IMP_TEAM_BASERUN_SB	TRSF_IMP_TEAM_BASERUN_CS	TARGET_WINS
TEAM_BATTING_H	0.0716	-0.0846	0.3880
TEAM_PITCHING_HR	-0.3714	0.3531	0.1862
IMP_TEAM_FIELDING_DP	-0.5264	0.3226	-0.0094
TRSF_TEAM_PITCHING_H	0.1614	-0.3365	0.0900
TRSF_TEAM_PITCHING_BB	0.0415	0.2046	0.1619
TRSF_TEAM_BATTING_3B	0.3671	-0.3296	0.1164
TRSF_IMP_TEAM_BASERUN_SB	1.0000	0.1404	0.1134
TRSF_IMP_TEAM_BASERUN_CS	0.1404	1.0000	0.0479
TARGET_WINS	0.1134	0.0479	1.0000

Outliers: Below chart for Cook's D – Target Wins shows many observations may qualify as outliers that could be removed. We removed the top 4 of these records (299, 1210, 1828, 2219) and tried to recalculate the fit however there was no significant difference hence we chose to keep them.



Leverage vs Outliers: We can see from plot below that the highest leverage point is at ~0.08 which is not too high, hence we do not have any heavy leverage points that we should try and remove.

The REG Procedure
Model: MODEL1
Dependent Variable: TARGET_WINS



Build Models

Based on the all the previous analysis done, we will now work towards selecting the best variables that have the most influence over our target variable. We will run several models to understand the variations in parameter estimates and adjusted R2 vs AICs.

In our analysis, we found that Stepwise and Backward both came up with a set of same 11 variables. This set of variables provided an adjusted R2 of 0.3281. In case of Forward selection, we got a set of 13 explanatory variables with an adjusted R2 of 0.3291.

Stepwise Selection

Variable TRSF_TEAM_PITCHING_BB Entered: R-Square = 0.3281 and C(p) = 12.6430

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	182190	16563	100.39	<.0001
Error	2261	373029	164.98389		
Corrected Total	2272	555218			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	152.12250	19.69783	9839.92109	59.64	<.0001
TEAM_BATTING_H	0.04134	0.00365	21182	128.39	<.0001
IMP_TEAM_FIELDING_DP	-0.12623	0.01432	12828	77.75	<.0001
TRSF_TEAM_PITCHING_H	30819	5898.60664	4503.74557	27.30	<.0001
TRSF_TEAM_PITCHING_BB	-10.95391	2.36418	3541.76013	21.47	<.0001
TRSF_TEAM_BATTING_3B	6.63160	0.89933	8970.93294	54.37	<.0001
TRSF_IMP_TEAM_BASERUN_SB	4.29571	0.59945	8472.33647	51.35	<.0001
TRSF_IMP_TEAM_BASERUN_CS	-1.75747	0.37617	3601.17500	21.83	<.0001
TEAM_BATTING_HR	0.03374	0.00825	2759.10565	16.72	<.0001
TEAM_BATTING_BB	0.04132	0.00595	7961.39951	48.26	<.0001
TRSF_TEAM_FIELDING_E	-15.27082	1.12636	30326	183.81	<.0001
TEAM_BATTING_2B	-0.03938	0.00882	3287.57544	19.93	<.0001

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TEAM_BATTING_H		1	0.1506	0.1506	590.368	402.58	<.0001
2	TEAM_BATTING_BB		2	0.0654	0.2160	372.251	189.32	<.0001
3	IMP_TEAM_FIELDING_DP		3	0.0137	0.2297	328.007	40.46	<.0001
4	TRSF_TEAM_FIELDING_E		4	0.0508	0.2805	159.093	160.04	<.0001
5	TRSF_TEAM_BATTING_3B		5	0.0165	0.2970	105.496	53.26	<.0001
6	TRSF_IMP_TEAM_BASERUN_SB		6	0.0057	0.3027	88.3746	18.46	<.0001
7	TRSF_IMP_TEAM_BASERUN_CS		7	0.0059	0.3085	70.6653	19.18	<.0001
8	TEAM_BATTING_HR		8	0.0054	0.3139	54.4530	17.85	<.0001
9	TEAM_BATTING_2B		9	0.0051	0.3190	39.4389	16.80	<.0001
10	TRSF_TEAM_PITCHING_H		10	0.0028	0.3218	32.1164	9.24	0.0024
11	TRSF_TEAM_PITCHING_BB		11	0.0064	0.3281	12.6430	21.47	<.0001

Backward Selection

Backward Elimination: Step 3

Variable IMP_TEAM_BATTING_SO Removed: R-Square = 0.3281 and C(p) = 12.6430

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	182190	16563	100.39	<.0001
Error	2261	373029	164.98389		
Corrected Total	2272	555218			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	152.12250	19.69783	9839.92109	59.64	<.0001
TEAM_BATTING_H	0.04134	0.00365	21182	128.39	<.0001
IMP_TEAM_FIELDING_DP	-0.12623	0.01432	12828	77.75	<.0001
TRSF_TEAM_PITCHING_H	30819	5898.60664	4503.74557	27.30	<.0001
TRSF_TEAM_PITCHING_BB	-10.95391	2.36418	3541.76013	21.47	<.0001
TRSF_TEAM_BATTING_3B	6.63160	0.89933	8970.93294	54.37	<.0001
TRSF_IMP_TEAM_BASERUN_SB	4.29571	0.59945	8472.33647	51.35	<.0001
TRSF_IMP_TEAM_BASERUN_CS	-1.75747	0.37617	3601.17500	21.83	<.0001
TEAM_BATTING_HR	0.03374	0.00825	2759.10565	16.72	<.0001
TEAM_BATTING_BB	0.04132	0.00595	7961.39951	48.26	<.0001
TRSF_TEAM_FIELDING_E	-15.27082	1.12636	30326	183.81	<.0001
TEAM_BATTING_2B	-0.03938	0.00882	3287.57544	19.93	<.0001

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TEAM_PITCHING_HR	13	0.0001	0.3291	13.4336	0.43	0.5103
2	TRSF_IMP_TEAM_PITCHING_SO	12	0.0004	0.3286	12.9296	1.50	0.2214
3	IMP_TEAM_BATTING_SO	11	0.0005	0.3281	12.6430	1.71	0.1907

GLM Select Output: GLM Select procedure also gave the same 11 variables returned from backward and stepwise selection. The adjusted R2 was very close to earlier value at 0.3249.

GLM Select

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 11).

Effects:	Intercept TEAM_BATTING_H IMP_TEAM_FIELDING_DP TRSF_TEAM_PITCHING_H TRSF_TEAM_PITCHING_B TRSF_TEAM_BATTING_3B TRSF_IMP_TEAM_BASERU TRSF_IMP_TEAM_BASERU TEAM_BATTING_HR TEAM_BATTING_BB TRSF_TEAM_FIELDING_E TEAM_BATTING_2B

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	11	182190	16563	100.39
Error	2261	373029	164.98389	
Corrected Total	2272	555218		

Root MSE	12.84461
Dependent Mean	80.83414
R-Square	0.3281
Adj R-Sq	0.3249
AIC	13893
AICC	13893
SBC	11686

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	152.122499	19.697831	7.72
TEAM_BATTING_H	1	0.041339	0.003648	11.33
IMP_TEAM_FIELDING_DP	1	-0.126232	0.014316	-8.82
TRSF_TEAM_PITCHING_H	1	30819	5898.606639	5.22
TRSF_TEAM_PITCHING_B	1	-10.953915	2.364180	-4.63
TRSF_TEAM_BATTING_3B	1	6.631600	0.899332	7.37
TRSF_IMP_TEAM_BASERU	1	4.295710	0.599452	7.17
TRSF_IMP_TEAM_BASERU	1	-1.757469	0.376172	-4.67
TEAM_BATTING_HR	1	0.033743	0.008251	4.09

Forward Selection

Forward Selection: Step 13

Variable TRSF_IMP_TEAM_PITCHING_SO Entered: R-Square = 0.3291 and C(p) = 13.4336

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	182719	14055	85.24	<.0001
Error	2259	372499	164.89563		
Corrected Total	2272	555218			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	159.30140	20.40874	10047	60.93	<.0001
TEAM_BATTING_H	0.03886	0.00399	15632	94.80	<.0001
IMP_TEAM_FIELDING_DP	-0.12866	0.01444	13082	79.33	<.0001
TRSF_TEAM_PITCHING_H	30804	6172.41013	4106.80539	24.91	<.0001
TRSF_TEAM_PITCHING_BB	-11.91927	2.69565	3223.89743	19.55	<.0001
TRSF_TEAM_BATTING_3B	6.23240	0.93029	7400.86867	44.88	<.0001
TRSF_IMP_TEAM_BASERUN_SB	4.55033	0.61923	8904.15101	54.00	<.0001
TRSF_IMP_TEAM_BASERUN_CS	-1.67848	0.37883	3237.07535	19.63	<.0001
TEAM_BATTING_HR	0.04255	0.01016	2894.63702	17.55	<.0001
TEAM_BATTING_BB	0.04185	0.00647	6909.79721	41.90	<.0001
TRSF_TEAM_FIELDING_E	-15.32257	1.12871	30389	184.29	<.0001
IMP_TEAM_BATTING_SO	-0.00483	0.00277	502.49426	3.05	0.0810
TEAM_BATTING_2B	-0.03616	0.00919	2552.83735	15.48	<.0001
TRSF_IMP_TEAM_PITCHING_SO	0.73968	0.60468	246.73981	1.50	0.2214

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TEAM_BATTING_H	1	0.1506	0.1506	590.368	402.58	<.0001
2	TEAM_BATTING_BB	2	0.0654	0.2160	372.251	189.32	<.0001
3	IMP_TEAM_FIELDING_DP	3	0.0137	0.2297	328.007	40.46	<.0001
4	TRSF_TEAM_FIELDING_E	4	0.0508	0.2805	159.093	160.04	<.0001
5	TRSF_TEAM_BATTING_3B	5	0.0165	0.2970	105.496	53.26	<.0001
6	TRSF_IMP_TEAM_BASERUN_SB	6	0.0057	0.3027	88.3746	18.46	<.0001
7	TRSF_IMP_TEAM_BASERUN_CS	7	0.0059	0.3085	70.6653	19.18	<.0001
8	TEAM_BATTING_HR	8	0.0054	0.3139	54.4530	17.85	<.0001
9	TEAM_BATTING_2B	9	0.0051	0.3190	39.4389	16.80	<.0001
10	TRSF_TEAM_PITCHING_H	10	0.0028	0.3218	32.1164	9.24	0.0024
11	TRSF_TEAM_PITCHING_BB	11	0.0064	0.3281	12.6430	21.47	<.0001
12	IMP_TEAM_BATTING_SO	12	0.0005	0.3286	12.9296	1.71	0.1907
13	TRSF_IMP_TEAM_PITCHING_SO	13	0.0004	0.3291	13.4336	1.50	0.2214

Final stepwise selection for model.

Stepwise Selection: Step 8

Variable TRSF_IMP_TEAM_BASERUN_CS Entered: R-Square = 0.2565 and C(p) = 9.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	142407	17801	97.63	<.0001
Error	2264	412812	182.33743		
Corrected Total	2272	555218			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-81.59433	9.47831	13512	74.11	<.0001
TEAM_BATTING_H	0.05489	0.00297	62207	341.17	<.0001
TEAM_PITCHING_HR	0.06855	0.00730	16079	88.18	<.0001
IMP_TEAM_FIELDING_DP	-0.05991	0.01424	3225.58903	17.69	<.0001
TRSF_TEAM_PITCHING_H	-42193	3622.72243	24733	135.64	<.0001
TRSF_TEAM_PITCHING_BB	4.57854	1.28344	2320.46775	12.73	0.0004
TRSF_TEAM_BATTING_3B	4.23690	0.89394	4095.97183	22.46	<.0001
TRSF_IMP_TEAM_BASERUN_SB	3.22002	0.62366	4860.59891	26.66	<.0001
TRSF_IMP_TEAM_BASERUN_CS	-1.13639	0.39136	1537.35258	8.43	0.0037

All variables left in the model are significant at the 0.1500 level.

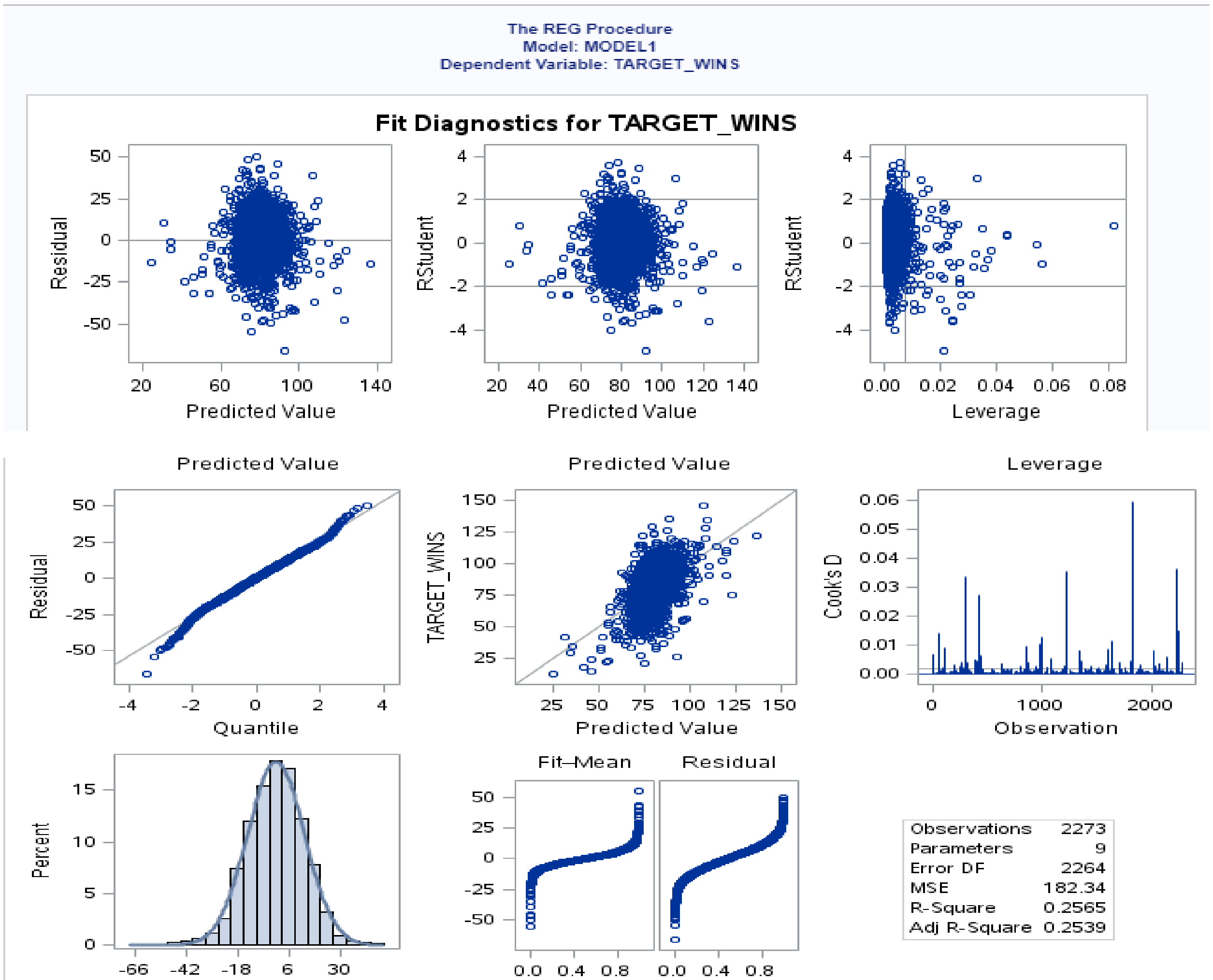
All variables have been entered into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	TEAM_BATTING_H		1	0.1506	0.1506	317.498	402.58	<.0001
2	TRSF_TEAM_PITCHING_H		2	0.0453	0.1959	181.423	128.01	<.0001
3	TRSF_IMP_TEAM_BASERUN_SB		3	0.0143	0.2102	139.985	40.98	<.0001
4	TEAM_PITCHING_HR		4	0.0206	0.2308	79.3389	60.66	<.0001
5	TRSF_TEAM_BATTING_3B		5	0.0122	0.2429	44.3403	36.38	<.0001
6	IMP_TEAM_FIELDING_DP		6	0.0068	0.2498	25.5113	20.66	<.0001
7	TRSF_TEAM_PITCHING_BB		7	0.0040	0.2537	15.4314	12.04	0.0005
8	TRSF_IMP_TEAM_BASERUN_CS		8	0.0028	0.2565	9.0000	8.43	0.0037

Model Selection

With the adjusted R2 and AIC not changing much with addition or removal on any more explanatory variables other than selected in previous step, we attempt to fit a model with the selected variables by Stepwise, Backward and GLMSelect procedures.

Below is the output from Proc Reg with a final **adjusted R2 of 0.2539.**



Data Step

We have a separate SAS code file that lists down the data step. It will be turned in with this paper and the original code.

Conclusion

We tried fitting a model with different approached listed below

- We tried transformations other than already listed in this paper, however the adjusted R2 was not improved enough to consider them.
- We also tried including more explanatory variables into the model, however that did not help either.
- In the model build step, we also provided screen capture of various methods that were tried to select the most influential explanatory variables. The last screen captures show the final 8 variables that were selected.
- With the current selected model, the outlier and normality treatment that was done provided best result based on the adjusted R2 metric. Any attempt to remove more outliers or try and normalize the dataset was not yielding any improvement to fit. Hence, we selected the current model.