Prasanna Rao Sec 58

MONEYBALL OLS REGRESSION PROJECT

INTRODUCTION:

The purpose of the assignment was to analyze 2276 records of baseball team data collected from 1871 to 2006 inclusive, in order to predict the number of wins for the team. The collected data was normalized for 162 games .Ordinary least square linear regression was used to predict the number of wins for the team. OLS regression was performed using forward, backward and stepwise using series of variables deemed fit for the model based on exploratory data analysis. The best model based on goodness of fit was analyzed to determine if the model was adequate to be deemed good enough to predict baseball wins. The best model was run on test data to create a scorecard .The scorecard was

DATA EXPLORATION

Data exploration is a critical component of data analysis. The intent of the data analysis was to understand any relationships in the underlying data, find missing values, find any outliers, influential points.

1) EDA : Using SAS Proc means , the following data was obtained

Variables, Description, Missing Values, Mean, Median, 1st and 99th percentile

Variable	Description	Missing	Mean	Median	1st Pctl	99th Pctl
INDEX		0	1268.46	1270.50	26.0000000	2510.00
TARGET_WINS		0	80.7908612	82.0000000	38.0000000	114.0000000
TEAM_BATTING_H	Base Hits by batters	0	1469.27	1454.00	1188.00	1950.00
TEAM_BATTING_2B	Doubles by batters	0	241.2469244	238.0000000	141.0000000	352.0000000
TEAM_BATTING_3B	Triples by batters	0	55.2500000	47.0000000	17.0000000	134.0000000
TEAM_BATTING_HR	Homeruns by batters	0	99.6120387	102.0000000	4.0000000	235.0000000
TEAM_BATTING_BB	Walks by batters	0	501.5588752	512.0000000	79.0000000	755.0000000
TEAM_BATTING_SO	Strikeouts by batters	102	735.6053358	750.0000000	67.0000000	1193.00
TEAM_BASERUN_SB	Stolen bases	131	124.7617716	101.0000000	23.0000000	439.0000000
TEAM_BASERUN_CS	Caught stealing	772	52.8038564	49.0000000	16.0000000	143.0000000
TEAM_BATTING_HBP	Batters hit by pitch	2085	59.3560209	58.0000000	29.0000000	90.0000000
TEAM_PITCHING_H	Hits allowed	0	1779.21	1518.00	1244.00	7093.00
TEAM_PITCHING_HR	Homeruns allowed	0	105.6985940	107.0000000	8.0000000	244.0000000
TEAM_PITCHING_BB	Walks allowed	0	553.0079086	536.5000000	237.0000000	924.0000000
TEAM_PITCHING_SO	Strikeouts by pitchers	102	817.7304508	813.5000000	205.0000000	1474.00
TEAM_FIELDING_E	Errors	0	246.4806678	159.0000000	86.0000000	1237.00
TEAM_FIELDING_DP	Double Plays	286	146.3879397	149.0000000	79.0000000	204.0000000

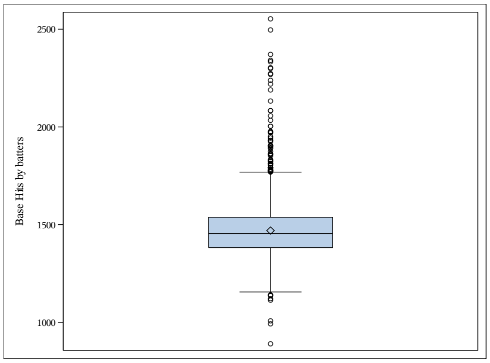
a) The following variables TEAM_PITCHING_SO, TEAM_BATTING_HBP, TEAM_FIELDING_DP, TEAM_BATTING_SO, and TEAM_BASERUN_CS & TEAM_BASERUN_SB were found to have missing value.

b) TEAM_BASERUN_CS was found to have 772 records or 34% of records missing whereas TEAM_BATTING_HBP to have 2085 records or 91.6% of records missing. As 92.6% of records is a big number it would not make sense to impute missing records for these missing records.

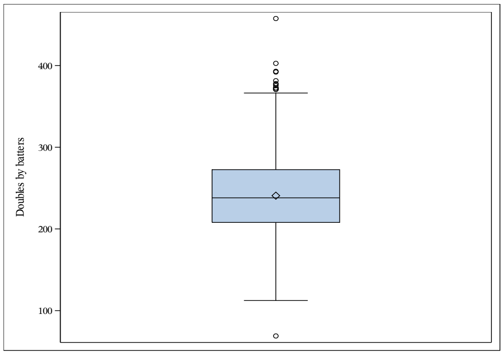
c) Missing values would be imputed in Data Preparation step.

2) **Outliers**

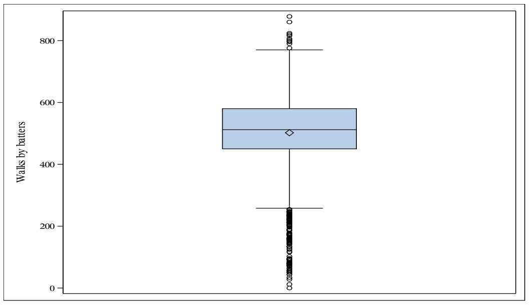
TEAM_BATTING_H



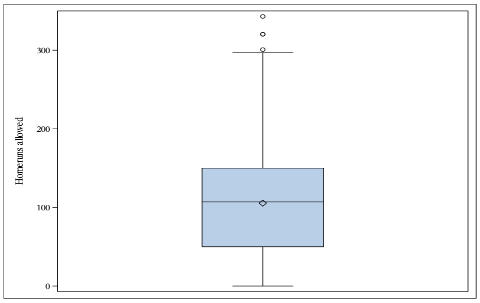
TEAM_BATTING_2B



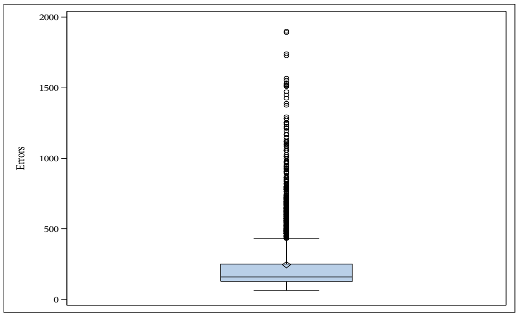
TEAM_BATTING_BB



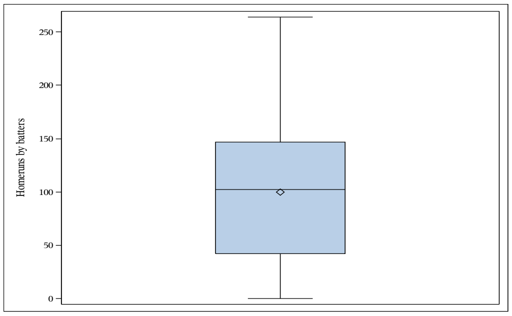
TEAM_PITCHING_HR



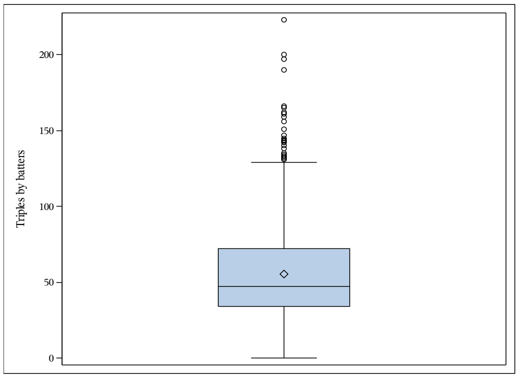
TEAM_FIELDING_E



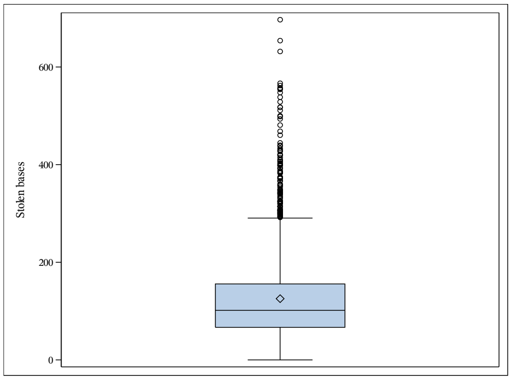
TEAM_BATTING_HR



TEAM_BATTING_3B

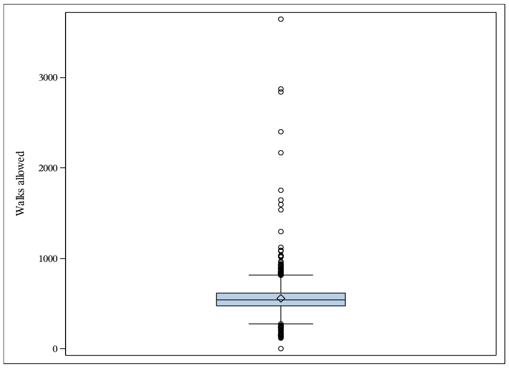


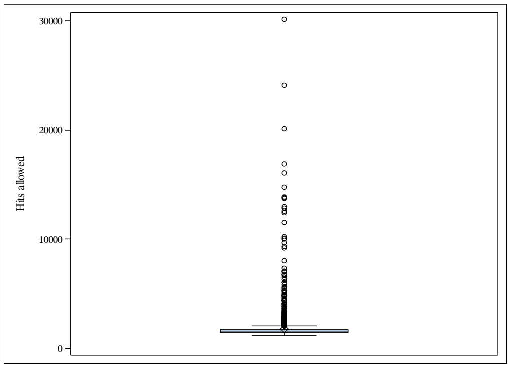
TEAM_BASERUN_SB



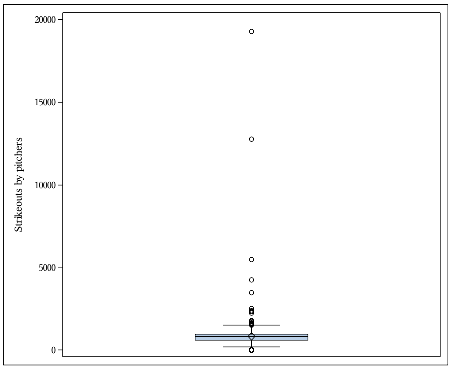
TEAM_PITCHING_BB

TEAM_PITCHING_H

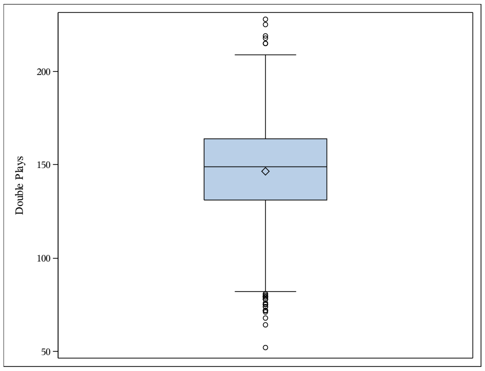




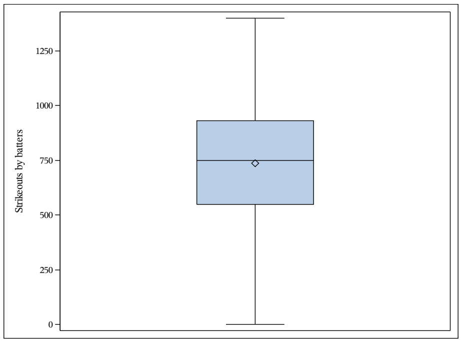
TEAM_PITCHING_SO



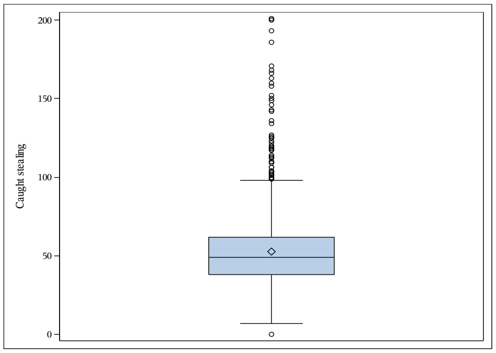
TEAM_FIELDING_DP



TEAM_BATTING_SO



TEAM_BASERUN_CS



Based on the above data, the following are the tabulated results of the outliers

Variable	Description	Missing Values	Outliers
TEAM_BATTING_H	Base hit batters	0	Yes
TEAM_BATTING_2B	Doubles by batters (2B)	0	Yes
TEAM_BATTING_BB	Walks by batters	0	Yes
TEAM_PITCHING_HR	Homeruns Allowed	0	Yes

TEAM_FIELDING_E	ERRORS	0	Yes
TEAM_BATTING_HR	Batters Homerun	0	None
TEAM_BATTING_3B	Batters Triple	0	Yes
TEAM_BASERUN_SB	Stolen Bases	131	Yes
TEAM_PITCHING_BB	Walks allowed	0	Yes
TEAM_PITCHING_H	Hits allowed	0	Yes
TEAM_PITCHING_SO	Strikeout by Pitchers	102	Yes
TEAM_BATTING_HBP	Batters Hit by pitch	2085	Yes
TEAM_FIELDING_DP	Double Plays	286	Yes
TEAM_BATTING_SO	Strikeout by batters	102	None
TEAM_BASERUN_CS	Caught Stealing	772	Yes

Except for TEAM_BATTING_HR & TEAM_BATTING_SO, rest of the variables has outliers. These outliers would be fixed in Data Preparation step .

3) Correlation between target and predictors and among predictors

a) Correlation between the TARGET_WINS and The predictor variables.(Top 9)

	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_BB	TEAM_PITCHING_HR	TEAM_FIELDING_E	TEAM_BATTING_HR	TEAM_BATTING_3b
Target_ Wins	.38	.28	.23	.18	-.176	.176	.142

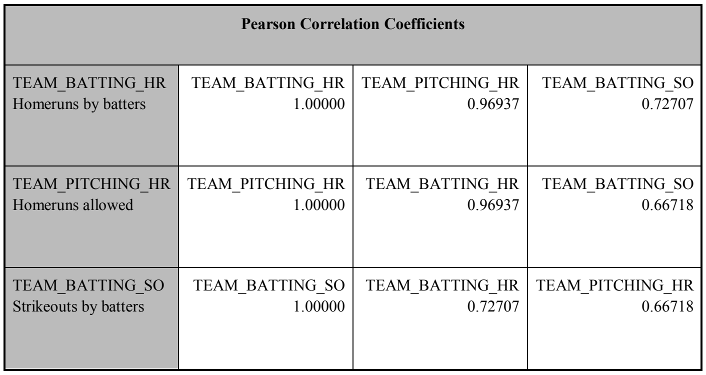
Based on the correlation results, Target Wins had the maximum positive correlation ship with TEAM_BATTING_H with a value of 38.8 %. This is consistent with the view that base hit perhaps the most important aspect of making runs in baseball .This was followed by Doubles and walks by batters.

b) Correlations hip among the predictors

Highest of correlations were found between the following predictors. These could exhibit multicollinearity.

1) TEAM_BATTING_HR & TEAM_PITCHING_HR with a +ve correlation of 97 %
2) TEAM_BATTING_SO & TEAM_BATTING_HR with a +ve correlation of 73%

Correlations O/P from SAS



Pearson Correlation Coefficients			
TEAM_BATTING_HR Homeruns by batters	TEAM_BATTING_HR 1.00000	TEAM_PITCHING_HR 0.96937	TEAM_BATTING_SO 0.72707
TEAM_PITCHING_HR Homeruns allowed	TEAM_PITCHING_HR 1.00000	TEAM_BATTING_HR 0.96937	TEAM_BATTING_SO 0.66718
TEAM_BATTING_SO Strikeouts by batters	TEAM_BATTING_SO 1.00000	TEAM_BATTING_HR 0.72707	TEAM_PITCHING_HR 0.66718

Thus based on EDA, The following is the summary.

Variable	**Description**	**Missing Values**	**Outliers**	**Top 7 Correlations(With Target_Wins)**

TEAM_BATTING_H	Base hit batters	0	Yes	0.38
TEAM_BATTING_2B	Doubles by batters (2B)	0	Yes	0.28
TEAM_BATTING_BB	Walks by batters	0	Yes	0.23
TEAM_PITCHING_HR	Homeruns Allowed	0	Yes	0.18
TEAM_FIELDING_E	ERRORS	0	Yes	-0.176
TEAM_BATTING_HR	Batters Homerun	0	None	0.176
TEAM_BATTING_3B	Batters Triple	0	Yes	0.142
TEAM_BASERUN_SB	Stolen Bases	131	Yes	
TEAM_PITCHING_BB	Walks allowed	0	Yes	
TEAM_PITCHING_H	Hits allowed	0	Yes	
TEAM_PITCHING_SO	Strikeout by Pitchers	102	Yes	
TEAM_BATTING_HBP	Batters Hit by pitch	2085	Yes	
TEAM_FIELDING_DP	Double Plays	286	Yes	
TEAM_BATTING_SO	Strikeout by batters	102	None	
TEAM_BASERUN_CS	Caught Stealing	772	Yes	

DATA PREPARATION

1) Outlier Handling

Based on the outliers detected in the EDA step

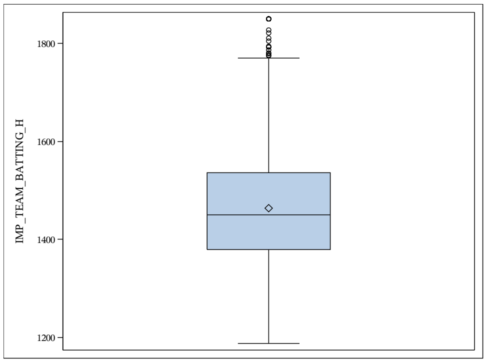
a) Values less than 1 % were forced to 1 % and those above 99% were forced to 99 % 1% and the 99 % of the data were detected using proc means with p1 and p95 and P99 respectively. I did not prune to 95% for I did not want to over fit the data. Example: For TEAM_BATTING_H , 1 % is 1189 and 99 % is 1850 .

```
IF TEAM_BATTING_H  < 1188 THEN
   IMP_TEAM_BATTING_H=1188;
IF TEAM_BATTING_H  > 1850 THEN
   IMP_TEAM_BATTING_H=1850;
```

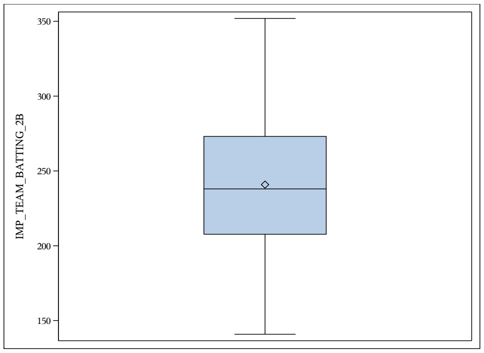
By executing this code MIN value was made equal to 1188 and Maximum value to 1850.

Fixed outlier Boxplots

TEAM_BATTING_H

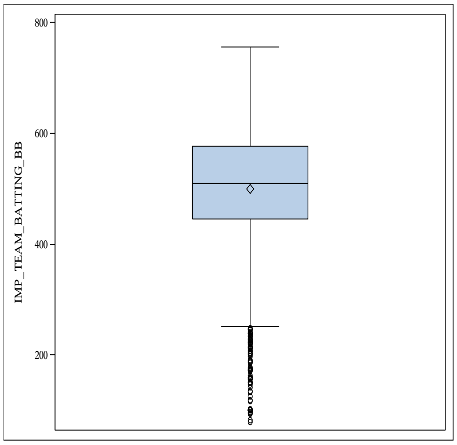


TEAM_BATTING_2B

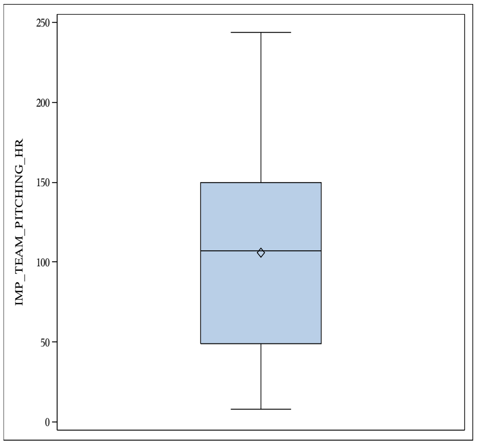


Prasanna Rao Sec 58

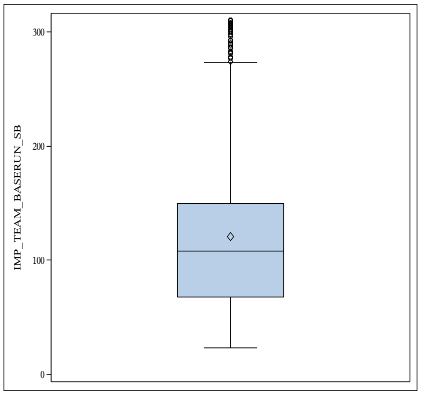
TEAM_BATTING_BB



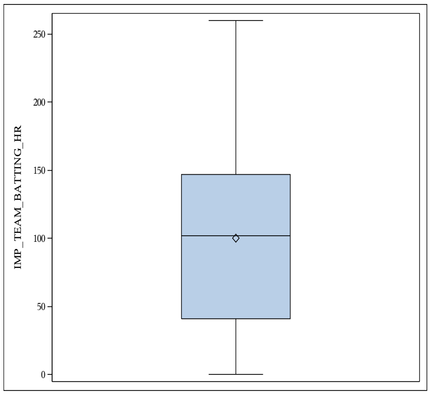
TEAM_PITCHING_HR



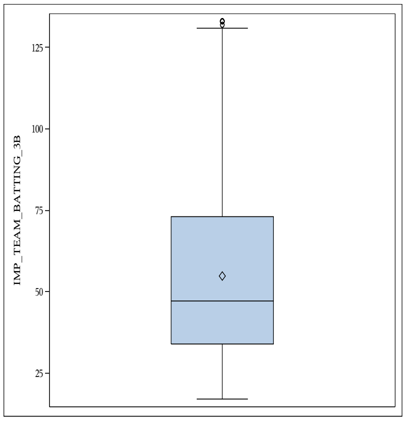
TEAM_FIELDING_E



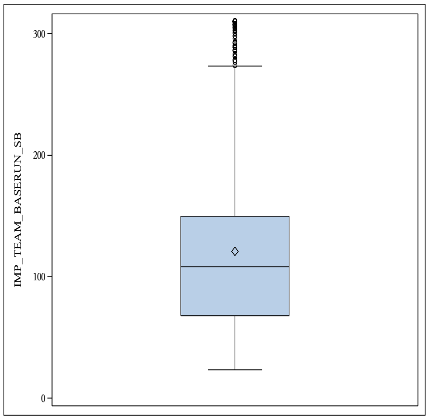
TEAM_BATTING_HR



TEAM_BATTING_3B



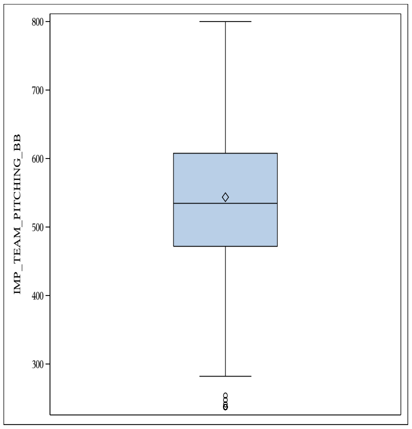
TEAM_BASERUN_SB

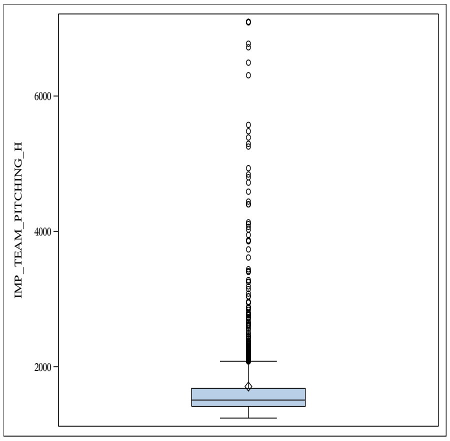


TEAM_PITCHING_BB

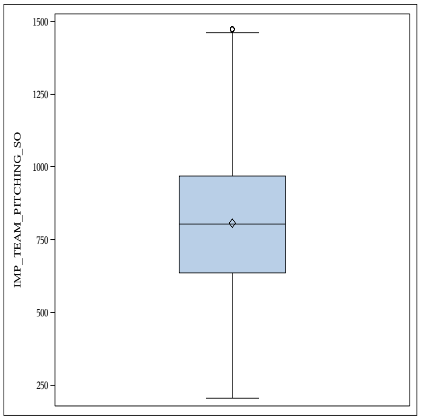
TEAM_PITCHING_H

Prasanna Rao Sec 58

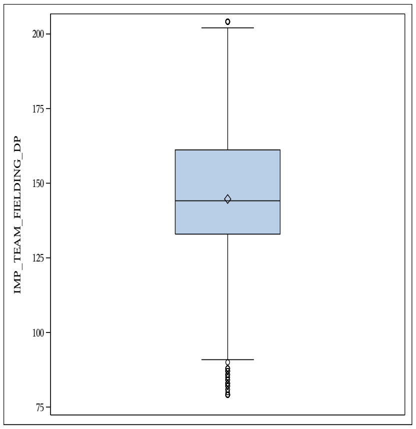




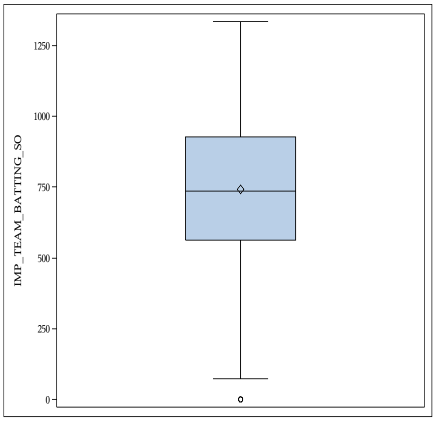
TEAM_PITCHING_SO



TEAM_FIELDING_DP



TEAM_BATTING_SO



2) Missing Data Imputation

a) Missing value is a serious threat as regression program skips missing data. Hence this was fixed by

   a. Creating a Missing flag for missing data if the data is missing

   b. Impute the missing data to their respective means

   The following code handled the missing data. M_TEAM_PITCHING_SO is the flag for TEAM_PITCHING_SO.

```
M_TEAM_PITCHING_SO=0;
 IF missing (TEAM_PITCHING_SO) then
    Do;
        IMP_TEAM_PITCHING_SO=773;
        M_TEAM_PITCHING_SO=1;
 End;
```

Missing Value Variable	Imputed Value	Missing value Flag
IMP_TEAM_BASERUN_SB	116	M_TEAM_BASERUN_SB
IMP_TEAM_PITCHING_SO	773	M_TEAM_PITCHING_SO
IMP_TEAM_FIELDING_DP	137	M_TEAM_FIELDING_DP
IMP_TEAM_BATTING_SO	736	M_TEAM_BATTING_SO
IMP_TEAM_BASERUN_CS	41	M_TEAM_BASERUN_CS

3) Interactions:

When OLS regression was run multiple times

a) It was found that IMP_TEAM_BATTING_2B changed sign with introduction of IMP_TEAM_FIELDING_E . Hence an interaction term interact1 was used to capture the interaction between IMP_TEAM_BATTING_2B & IMP_TEAM_FIELDING_E .
   Interact1 = IMP_TEAM_BATTING_2B * IMP_TEAM_FIELDING_E

b) Likewise an interaction was seen between IMP_TEAM_PITCHING_H & IMP_TEAM_BATTING_H. This was captured using
   interact2 = IMP_TEAM_PITCHING_H * IMP_TEAM_BATTING_H

MODELS

Having cleansed the data, studied the data for outliers and correlations, the next step was regression. Stepwise, Backward and Forward regression techniques were used for regression. R square, Adjusted Rsquare, AIC, SBC, BIC and VIF were captured for these models.
As TEAM_BATTING_HR & TEAM_PITCHING_HR had a very correlation of 96 % , Each of these variables were independently used and then the variable which impacted Adjusted R square ,AIC ,BIC and SBC was used. TEAM_BATTING_HR had the better of the impact and hence retained.

Prasanna Rao Sec 58

Model 1: Selection = STEPWISE

All the parameters IMP_TEAM_BATTING_H, IMP_TEAM_BATTING_2B, IMP_TEAM_BATTING_3B, IMP_TEAM_BATTING_HR, IMP_TEAM_BATTING_BB
IMP_TEAM_BATTING_SO, IMP_TEAM_BASERUN_CS , IMP_TEAM_BASERUN_SB
IMP_TEAM_PITCHING_H, IMP_TEAM_PITCHING_HR , IMP_TEAM_PITCHING_BB
IMP_TEAM_PITCHING_SO, IMP_TEAM_FIELDING_E, IMP_TEAM_FIELDING_DP

Null Indicators: M_TEAM_BASERUN_SB , M_TEAM_PITCHING_SO , M_TEAM_FIELDING_DP
M_TEAM_BATTING_SO & M_TEAM_BASERUN_CS

Results:

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	225364	15024	100.12	<.0001
Error	2260	339132	150.05860		
Corrected Total	2275	564496			

Root MSE	12.24984	R-Square	0.3992
Dependent Mean	80.79086	Adj R-Sq	0.3952
Coeff Var	15.16241		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	27.39060	5.45848	5.02	<.0001	0
IMP_TEAM_BATTING_H	1	0.03386	0.00397	8.54	<.0001	3.89271
IMP_TEAM_BATTING_2B	1	-0.02062	0.00887	-2.33	0.0202	2.50378
IMP_TEAM_BATTING_3B	1	0.10953	0.01709	6.41	<.0001	3.22405
IMP_TEAM_BATTING_HR	1	-0.04941	0.03017	-1.64	0.1016	50.60194
IMP_TEAM_BATTING_BB	1	0.05140	0.00809	6.35	<.0001	13.79783
IMP_TEAM_BASERUN_SB	1	0.06048	0.00500	12.09	<.0001	2.48808
IMP_TEAM_PITCHING_H	1	0.00722	0.00084549	8.54	<.0001	6.64710
IMP_TEAM_PITCHING_HR	1	0.12166	0.02812	4.33	<.0001	43.83695
IMP_TEAM_PITCHING_BB	1	-0.02356	0.00648	-3.64	0.0003	8.59108
IMP_TEAM_PITCHING_SO	1	-0.01210	0.00192	-6.31	<.0001	3.11644
IMP_TEAM_FIELDING_E	1	-0.06869	0.00402	-17.09	<.0001	11.30925

IMP_TEAM_FIELDING_DP	1	-0.10121	0.01385	-7.31	<.0001	1.73373
M_TEAM_BASERUN_SB	1	37.43003	1.88599	19.85	<.0001	2.92647
M_TEAM_PITCHING_SO	1	8.56929	1.46103	5.87	<.0001	1.38594
M_TEAM_FIELDING_DP	1	4.84860	1.47370	3.29	0.0010	3.61913

Obs	_MODEL_	_AIC_	_SBC_	_BIC_	_CP_	_ADJRSQ_
1	MODEL_1	11421.04	11512.72	11423.27	15.5761	**0.39524**

Analysis: Model1

1. BATTING_2B replaced by INTERACT1 (refer: Data Cleaning) since Batting_2B changed signs whenever IMP_TEAM_FIELDING_E was introduced.
2. Also as PICTHING_HR was highly correlated with IMP_BATTING_HR, as evident very high VIF and hence need to be removed;
3. TEAM_PITCHING_H is positive .But to me it does not make sense as it should impact the wins negatively
4. Adjusted R square was 39.52 .

Model 2: Selection = STEPWISE, BATTING_2B replaced by INTERACT1 & PICTHING_HR was highly correlated with IMP_BATTING_HR, as evident very high VIF and hence removed;

Results:

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	232759	16626	113.31	<.0001
Error	2261	331737	146.72148		
Corrected Total	2275	564496			

Root MSE	12.11286	R-Square	0.4123
Dependent Mean	80.79086	Adj R-Sq	0.4087
Coeff Var	14.99286		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	53.39117	5.86077	9.11	<.0001	0

IMP_TEAM_BATTING_H	1	0.01410	0.00351	4.01	<.0001	3.12020
Interact1	1	0.00016644	0.00001902	8.75	<.0001	13.57017
IMP_TEAM_BATTING_3B	1	0.11774	0.01663	7.08	<.0001	3.12068
IMP_TEAM_BATTING_HR	1	0.08859	0.00942	9.40	<.0001	5.04771
IMP_TEAM_BATTING_BB	1	0.05242	0.00694	7.55	<.0001	10.38872
IMP_TEAM_BATTING_SO	1	-0.01788	0.00218	-8.20	<.0001	4.35027
IMP_TEAM_BASERUN_SB	1	0.06136	0.00497	12.35	<.0001	2.51213
IMP_TEAM_PITCHING_H	1	- 0.0069	0.00082619	8.39	<.0001	6.49137
IMP_TEAM_PITCHING_BB	1	-0.02605	0.00537	-4.85	<.0001	6.03874
IMP_TEAM_FIELDING_E	1	-0.10197	0.00538	-18.96	<.0001	20.70786
IMP_TEAM_FIELDING_DP	1	-0.10065	0.01369	-7.35	<.0001	1.73177
M_TEAM_BASERUN_SB	1	34.21821	1.79306	19.08	<.0001	2.70535
M_TEAM_PITCHING_SO	1	10.26114	1.45456	7.05	<.0001	1.40494
M_TEAM_FIELDING_DP	1	5.80274	1.45559	3.99	<.0001	3.61103

Analysis: Model 2

1) Fielding double play must help the fielding team with wins .This is not in conclusion with the results as its negative .This is the only concern. And the correlation between Double play and target is also negative..
2) Adjusted R square was 40.86

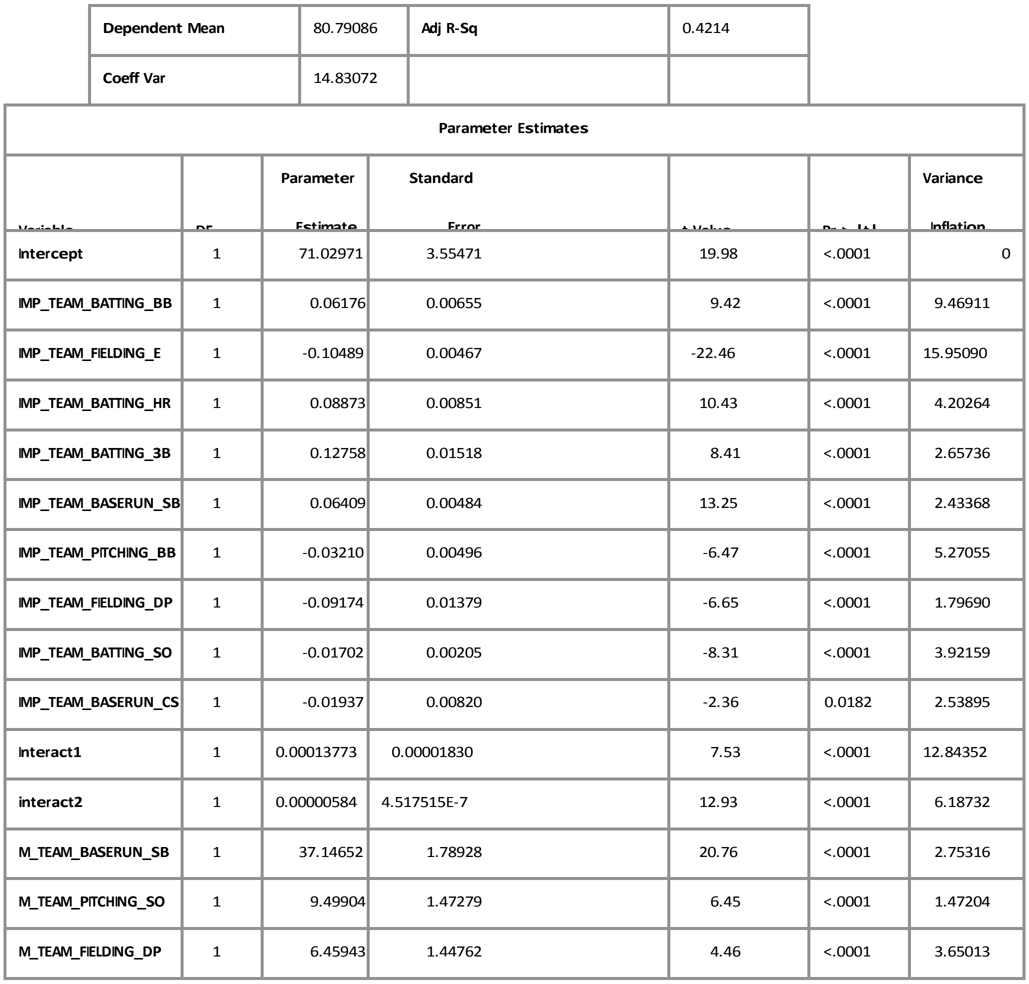
Obs	_MODEL_	_AIC_	_SBC_	_BIC_	_CP_	_ADJRSQ_
2	MODEL_2	11368.86	11454.81	11371.07	14.3502	0.40869

Model3: Selection = STEPWISE, Used interact2 = IMP_TEAM_PITCHING_H * IMP_TEAM_BATTING_H to see the interaction between pitching and the base run.

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	239896	17135	119.36	<.0001
Error	2261	324601	143.56503		
Corrected Total	2275	564496			

Root MSE	11.98186	R-Square	0.4250



Dependent Mean	80.79086	Adj R-Sq	0.4214
Coeff Var	14.83072		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	71.02971	3.55471	19.98	<.0001	0
IMP_TEAM_BATTING_BB	1	0.06176	0.00655	9.42	<.0001	9.46911
IMP_TEAM_FIELDING_E	1	-0.10489	0.00467	-22.46	<.0001	15.95090
IMP_TEAM_BATTING_HR	1	0.08873	0.00851	10.43	<.0001	4.20264
IMP_TEAM_BATTING_3B	1	0.12758	0.01518	8.41	<.0001	2.65736
IMP_TEAM_BASERUN_SB	1	0.06409	0.00484	13.25	<.0001	2.43368
IMP_TEAM_PITCHING_BB	1	-0.03210	0.00496	-6.47	<.0001	5.27055
IMP_TEAM_FIELDING_DP	1	-0.09174	0.01379	-6.65	<.0001	1.79690
IMP_TEAM_BATTING_SO	1	-0.01702	0.00205	-8.31	<.0001	3.92159
IMP_TEAM_BASERUN_CS	1	-0.01937	0.00820	-2.36	0.0182	2.53895
Interact1	1	0.00013773	0.00001830	7.53	<.0001	12.84352
interact2	1	0.00000584	4.517515E-7	12.93	<.0001	6.18732
M_TEAM_BASERUN_SB	1	37.14652	1.78928	20.76	<.0001	2.75316
M_TEAM_PITCHING_SO	1	9.49904	1.47279	6.45	<.0001	1.47204
M_TEAM_FIELDING_DP	1	6.45943	1.44762	4.46	<.0001	3.65013

Analysis: Model 3

1. Intercept value is very high in the range 71. This translates to say that with no match being played any team would have minimum average of 71 and I would not go with this value
2. Adjusted R square was 40.86

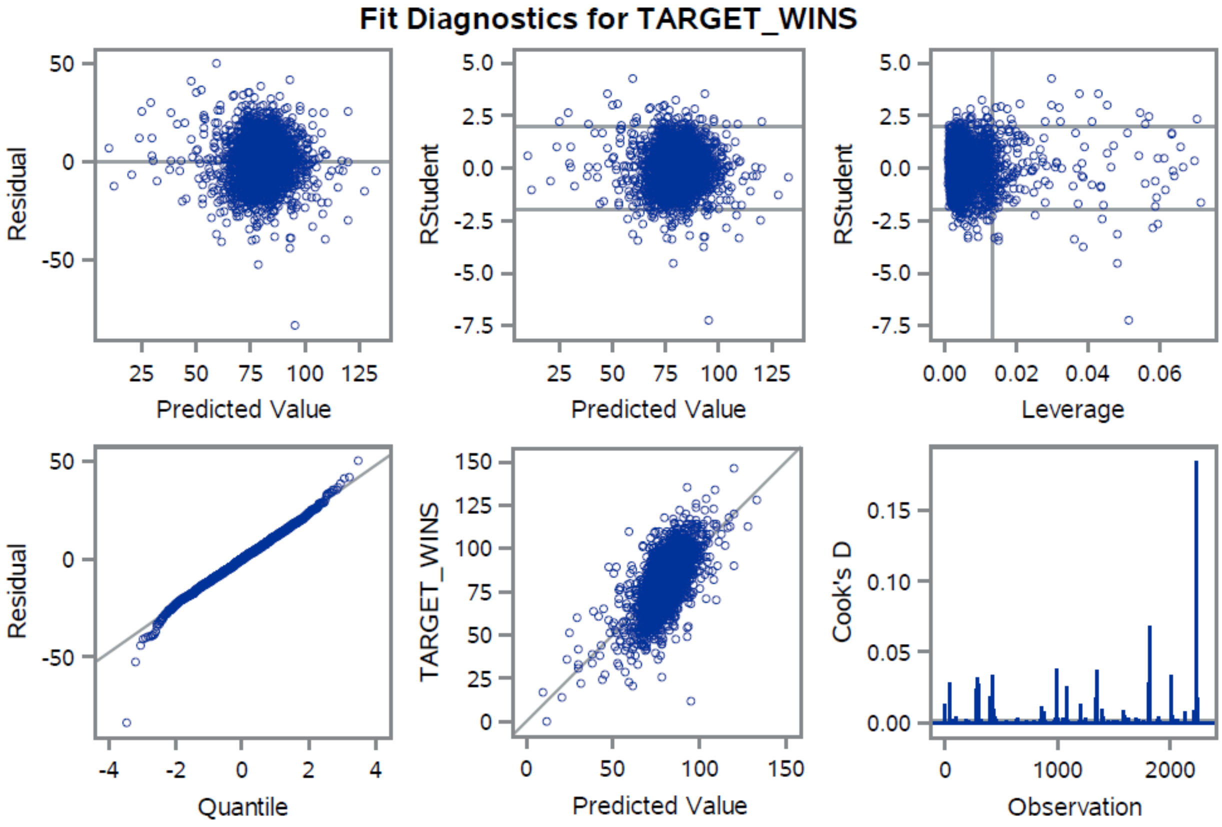
Obs	_MODEL_	_AIC_	_SBC_	_BIC_	_CP_	_ADJRSQ_
1	MODEL3	11319.36	11405.31	11321.58	13.2975	0.42141

MODEL SELECTION

Criteria for model selection:

1) Parsimonious model which clearly explains and make sense the parameters
2) AIC, BIC, SBC & Adjusted R square values for a model would be considered. As model 3 has the least of AIC, BIC & SBC along with highest of adjusted R square, it would have been a good choice compared to MODEL2 and MODEL 1. But model 2 was chosen as model 3 has a very high value of intercept which translates to 72 games if the team does not play and it's a very high value
3) I would go with a more parsimonious model as parsimonious models make more sense to understand what's happening. But it should be complemented with desired level of AIC, BIC, SBC & Adjusted R square

4) Model Diagnostics (Model 2)



Assumption Analysis :

1) Residuals vs Predicted: The residuals of the analysis were then graphed to ascertain whether a pattern exists in the residuals. From the figure, there does not appear to be any obvious pattern in the data and hence there should be no correlation between residuals and predicted values which in turn displays linearity of the variables.

2) Normal probability of standardized residuals : The plot does resembles a straight line and normality assumption is satisfied except for the beginning part of the line
3) Homoscedacity: Standardized residuals vs predicted values (Middle Graph in first row) shows no pattern and thus reflects homoscedacity.

Conclusion

```
TARGET_WINS =
    P_TARGET_WINS = 53.72 +   0.01247*IMP_TEAM_BATTING_H (BASEHIT BY BATTERS)
                        +  0.05354*IMP_TEAM_BATTING_BB   (WALK BY BATTERS)
                        -  0.10153*IMP_TEAM_FIELDING_E    (ERRORS BY FIELDING)
                        +  0.07989*IMP_TEAM_BATTING_HR   (HOMERUNS BY BATTERS)
                        +  0.14568*IMP_TEAM_BATTING_3B   (TRIPPLES BY BATTERS)
                        +   .08503*IMP_TEAM_BASERUN_SB   (STOLEN BASES)
                        -  0.02479*IMP_TEAM_PITCHING_BB   (WALKS ALLOWED)
                        - 0.08134*IMP_TEAM_FIELDING_DP    (FIELDING DOUBLE PLAYS)
                        - 0.01487*IMP_TEAM_BATTING_SO     (STRIKEOUT BY BATTERS)
                        - 0.05897*IMP_TEAM_BASERUN_CS     (CAUGHT STEALING)
                        + 37.43317*M_TEAM_BASERUN_SB      (STOLEN BASE INDICATOR)
                        + 6.19735*M_TEAM_PITCHING_SO      (PITCHING STRIKEOUT IND)
                        + 0.0001764*INTERACT1              (IMP_TEAM_BATTING_2B *
IMP_TEAM_FIELDING_E      )
                         + 0.00000263*INTERACT2             (IMP_TEAM_PITCHING_H *
IMP_TEAM_BATTING_H)
```

1) Base Hits by batters, Doubles by batters (2B), Triples by batters (3B), Homeruns by batters, Walks by Batters and Stolen bases do impact positively wins and this is in same conclusion with the results .
2) Walks Allowed , Strike out by batters and caught stealing , Errors do impact the wins and this in conclusion with the findings as they are all negative
3) Fielding double play must help the fielding team with wins .This is not in conclusion with the results as its negative .This is the only concern. And the correlation between Double play and target is also negative..
4) Adjusted R square was 40.86. It had better AIC, BIC, SBC & Adjusted R square than model 1
5) Model 3 which had the best of AIC, BIC, SBC & Adjusted was not chosen as the intercept value of 72 is high. This translates to 72 wins if no play.

STAND ALONE SCORING PROGRAM

MoneyBall_Scorecard_Code.txt

SCORED DATA FILE

Prasanna_predictions .xlsx