PAVEL DUKEL

NORTHWESTERN UNIVERSITY

PREDICT 411 SEC 58

ASSIGNMENT 1: MONEYBALL

NOTES:

Disclaimer:

I don't understand baseball except for basic rules of the game therefore I will not be able to comment on the models and perform in-depth analysis of whether models make sense.

File names:

1. **SAS Code used in the analysis: Pavel_Dukel_MoneyBall_Code.sas**
2. **SAS Data Step: Pavel_Dukel_MoneyBall_DataStep.sas**
3. **CSV file with scored records: Pavel_Dukel_MoneyBall_Index.csv**

Kaggle Name: Pavel_D

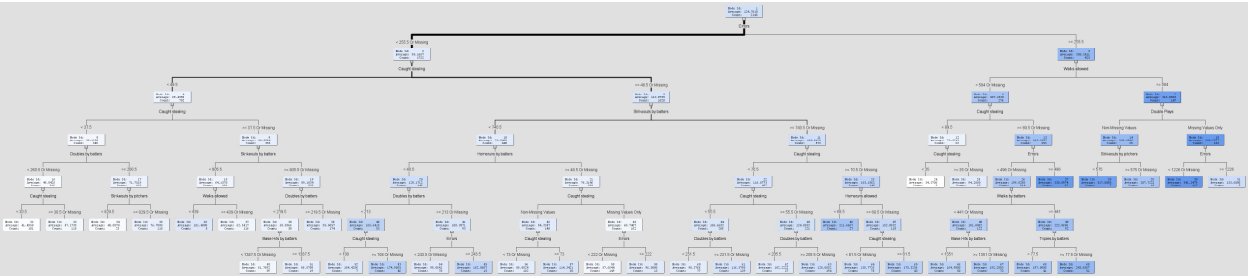
Bingo Bonus Claimed:

1. **Decision Tree from SAS EM (20 points)**
2. **Application of SAS Macros in the code (10 points)**

1. Decision Trees (20 points):

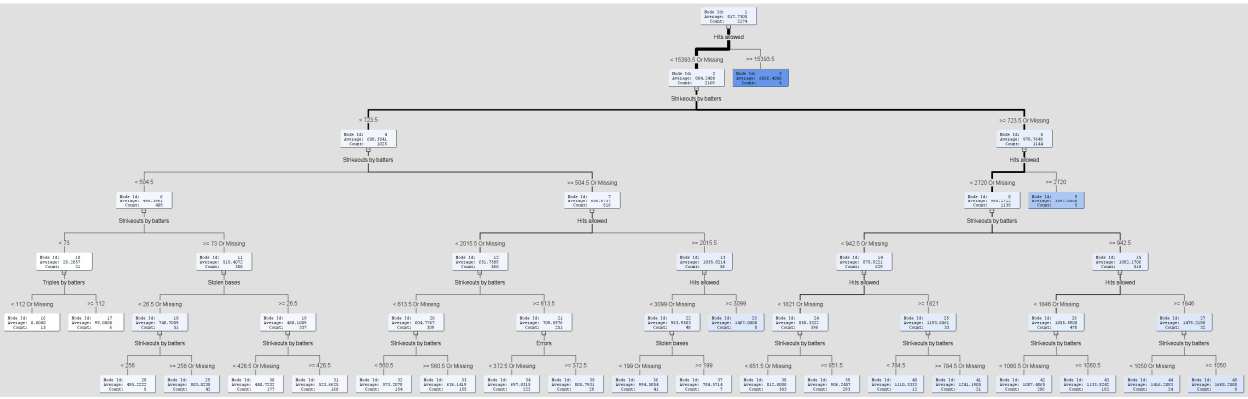
Decision trees have been used to impute missing values for the following variables:

a) TEAM_BASERUN_SB (3-level decision tree developed using Enterprise Miner)



```
IMP_TEAM_BASERUN_SB =TEAM_BASERUN_SB;
if missing( IMP_TEAM_BASERUN_SB ) then do;
        if TEAM_FIELDING_E >= 255.5 then
                IMP_TEAM_BASERUN_SB = 245;
        else do;
                if TEAM_BASERUN_CS < 48.5 then
                        IMP_TEAM_BASERUN_SB = 70;
                else
                        IMP_TEAM_BASERUN_SB = 113;
        end;
end;
```

b) TEAM_PITCHING_SO (3-level decision tree developed using Enterprise Miner)



```
IMP_TEAM_PITCHING_SO = TEAM_PITCHING_SO;
if missing( IMP_TEAM_PITCHING_SO ) then do;
        if TEAM_BATTING_H > 15393 then
                IMP_TEAM_PITCHING_SO = 6666;
        else do;
                if TEAM_BATTING_SO < 723.5 then
                        IMP_TEAM_PITCHING_SO = 608;
                else
                        IMP_TEAM_PITCHING_SO = 979;
        end;
end;
```

2. Macros (10 points):

The following SAS macros have been used in the exploratory data analysis;

```
*///   Macro used in data exploration to look at some boxplots   ///;
%MACRO box_plot ( TEMP , param= );
        proc sgplot data=&TEMP.;
          vbox &param. / group=wins grouporder=ascending;
          xaxis label="Wins";
          keylegend / title="BoxPlot";
        run;
%MEND box_plot;

*///   Macro used to test which out of transformed variables stay in the model  ///;
%MACRO Variable_Test_6 ( TEMP , param1=, param2=,
        param3=, param4=, param5=, param6=);
        proc reg data=&TEMP.;
                model TARGET_WINS =
                                &param1.
                                &param2.
                                &param3.
                                &param4.
                                &param5.
                                &param6.
                                /selection=stepwise VIF;
        run;
%MEND Variable_Test_6;

*///   Macro used to check variable correlation   ///;
%MACRO data_corr ( TEMP , param=);
        proc corr data=&TEMP. outp=CORFILE;
        run;
        data CORFILE1;
                set CORFILE;
                if _TYPE_ in ("CORR");
                keep _NAME_ &param.;
        run;
        proc sort data=CORFILE1;
                by descending &param.;
        run;
        proc print data=CORFILE1;
        run;
%MEND data_corr;
```

MoneyBall OLS Regression Project

Introduction

The objective of this analysis is to build a multivariate linear regression model designed to predict the number of wins by a professional baseball team in a regular season. The data set used in the analysis contains performance indicators such as batting, base-running, pitching, and fielding statistics which are commonly used in baseball. The set contains 2276 records and is based on the data collected from years 1871 to 2006 inclusive. There are total of 15 predictor variables (performance statistics) in this set.
The analysis utilizes the following three variable selection techniques: manual (arbitrarily selected by the analyst), forward, and stepwise. Results obtained (models developed) using the aforementioned techniques are compared based on the values of the commonly used metrics (Adj. R-Squared, AIC, SBC) as well as whether a model is easy to understand and implement in order to identify the optimal one.

Data Exploration

First, data set is examined using MEANS procedure (see table_1 below). There is a number of predictor variables that have missing values that need to be fixed/imputed before the data can be used in regression analysis. One of these variables, namely TEAM_BATTING_HBP, has over 90% values missing and is going to be dropped. The remaining predictor variables with missing values are going to be imputed using average (in our case MEDIAN) or decision tree obtained from the SAS Enterprise Miner.

The MEANS Procedure

Variable	Label	N Miss	Mean	Median	Std Dev	Maximum	Minimum
TARGET_WINS		0	81	82	16	146	0
TEAM_BATTING_H	Base Hits by batters	0	1469	1454	145	2554	891
TEAM_BATTING_2B	Doubles by batters	0	241	238	47	458	69
TEAM_BATTING_3B	Triples by batters	0	55	47	28	223	0
TEAM_BATTING_HR	Homeruns by batters	0	100	102	61	264	0
TEAM_BATTING_BB	Walks by batters	0	502	512	123	878	0
TEAM_BATTING_SO	Strikeouts by batters	102	736	750	249	1399	0
TEAM_BASERUN_SB	Stolen bases	131	125	101	88	697	0
TEAM_BASERUN_CS	Caught stealing	772	53	49	23	201	0
TEAM_BATTING_HBP	Batters hit by pitch	2085	59	58	13	95	29
TEAM_PITCHING_H	Hits allowed	0	1779	1518	1407	30132	1137
TEAM_PITCHING_HR	Homeruns allowed	0	106	107	61	343	0
TEAM_PITCHING_BB	Walks allowed	0	553	537	166	3645	0
TEAM_PITCHING_SO	Strikeouts by pitchers	102	818	814	553	19278	0
TEAM_FIELDING_E	Errors	0	246	159	228	1898	65
TEAM_FIELDING_DP	Double Plays	286	146	149	26	228	52

Table 1: MEANS procedure output

Table 2 below contains Pearson correlation coefficients of independent variables with the target variable ranked from highest to lowest (left-to-right, top-to-bottom). Based on the data in the table, neither independent variable is strongly correlated with TARGET_WINS variable. The variables with the strongest correlation are TEAM_BATTING_H, TEAM_BATTING_2B, and TEAM_BATTING_BB. Also, majority of the variables have positive correlation with target variable except for TEAM_FIELDING_E,

TEAM_PITCHING_H, TEAM_PITCHING_SO, TEAM_FIELDING_DP and TEAM_BATTING_SO that display a moderately negative correlation. Out of 15 analyzed variables, 5 have correlation that is contrary to the expected effect of the variable on the target, that is TEAM_BASERUN_CS, TEAM_PITCHING_BB, and TEAM_PITCHING_HR are supposed to have negative impact on Wins but have positive correlation whereas TEAM_FIELDING_DP and TEAM_PITCHING_SO are supposed to have positive impact but at the same time display negative correlation with the TARGET_WINS variable. The latter five variables will require additional attention and scrutiny when evaluating developed linear regression models.

TARGET_WINS	TEAM_BATTING_H 0.38877 <.0001 2276	TEAM_BATTING_2B 0.28910 <.0001 2276	TEAM_BATTING_BB 0.23256 <.0001 2276	TEAM_PITCHING_HR 0.18901 <.0001 2276	TEAM_FIELDING_E -0.17648 <.0001 2276
TARGET_WINS	TEAM_BATTING_HR 0.17615 <.0001 2276	TEAM_BATTING_3B 0.14261 <.0001 2276	TEAM_BASERUN_SB 0.13514 <.0001 2145	TEAM_PITCHING_BB 0.12417 <.0001 2276	TEAM_PITCHING_H -0.10994 <.0001 2276
TARGET_WINS	TEAM_PITCHING_SO -0.07844 0.0003 2174	TEAM_BATTING_HBP 0.07350 0.3122 191	TEAM_FIELDING_DP -0.03485 0.1201 1990	TEAM_BATTING_SO -0.03175 0.1389 2174	TEAM_BASERUN_CS 0.02240 0.3853 1504

Table 2: Pearson Correlation Coefficients

The following graphs are scatterplots of the 3 variables (that have strongest correlation – **see table 2**) displayed with the "loess smoothing" lines super-imposed on the data (see figure 1 below). Since the loess lines are not perfectly straight, neither of 3 variables appears to have a strong linear relationship. Similarly, the remaining variables (the graphs are not displayed here) don't have straight loess lines.

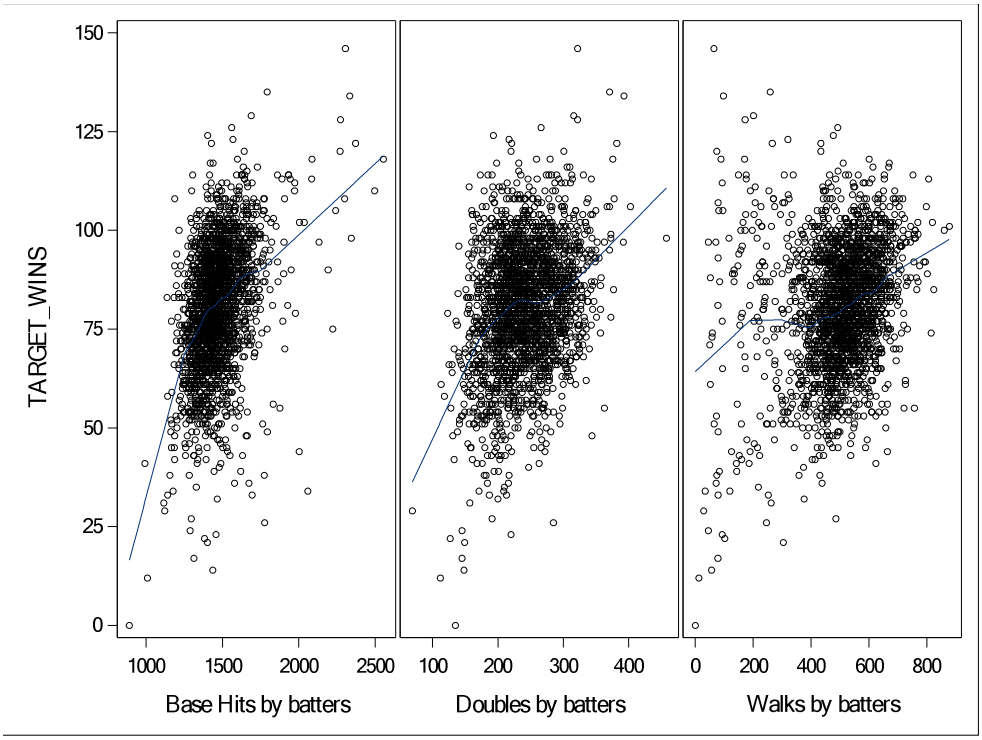


Figure 3: Loess Smoothing of variables TEAM_BATTING_H, TEAM_BATTING_2B, and TEAM_BATTING_BB

Next, let's explore the dependent variable TARGET_WINS box plot (see figure_2 and figure_3 below). Distribution of the variable is approximately normal, moderately skewed to the left (for smaller values). Most of the values are within 3 standard deviations from the mean (from 33 to 129 based on mean of 81 and standard deviation of 16; see table 1). Only 13 values are below and only 3 are above 3 standard deviations from the mean. Based on the box plot (see figure 2), there are two obvious outliers that (values are 0 and 146) that require additional consideration (most likely these values are not realistic).

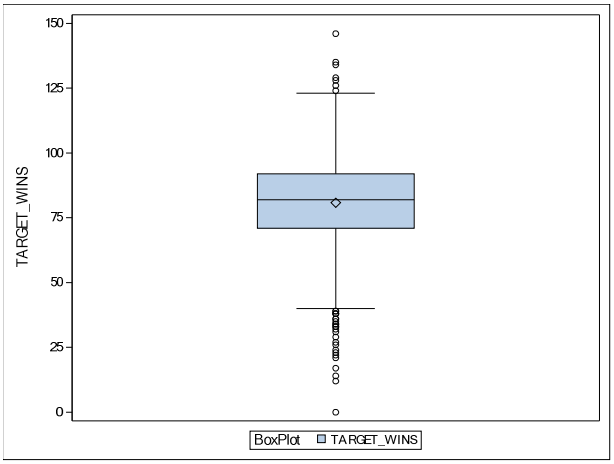


Figure 2: Dependent variable box plot

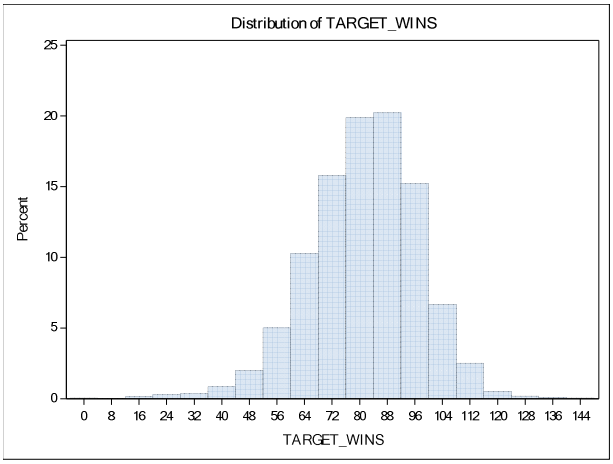


Figure 3: TARGET_WINS histogram

Two new (synthetic) variables are created (TEAM_BATTING_COMB_1 and TEAM_ BATTING_COMB_1) in order to check whether various variable mathematical combinations are going to improve correlation with the response variable TEAM_WINS. These variables are plotted against the dependent variable that has been grouped into five clusters (based on 2 and 3 standard deviation cutoff limits) created for this purpose. The groups are as follows: less than 33, 33-48, 49-112, 113-128, 129 and greater.

Obs	_NAME_	WINS
1	TARGET_WINS	1.00000
2	TEAM_BATTING_COMB_1	0.41066
3	TEAM_BATTING_H	0.38877
4	TEAM_BATTING_2B	0.28910
5	TEAM_BATTING_BB	0.23256
6	TEAM_BATTING_COMB_2	0.21743
7	TEAM_PITCHING_HR	0.18901
8	TEAM_BATTING_HR	0.17615
9	TEAM_BATTING_3B	0.14261
10	TEAM_PITCHING_BB	0.12417
11	IMP_TEAM_BASERUN_SB	0.11250
12	IMP_TEAM_BASERUN_CS	0.01596
13	M_TEAM_BATTING_SO	0.00773
14	M_TEAM_PITCHING_SO	0.00773

Table 3: Pearson Correlation Coefficients

The variables have been developed as follow:

```
TEAM_BATTING_COMB_1 =   TEAM_BATTING_H
                      + 2 * TEAM_BATTING_2B
                      + 3 * TEAM_BATTING_3B
                      + 4 * TEAM_BATTING_HR;
TEAM_BATTING_COMB_2 = TEAM_BATTING_H
                      - TEAM_BATTING_2B
                      - TEAM_BATTING_3B
                      - TEAM_BATTING_HR;
```

TEAM_BATTING_COMB_1 variable shows improved correlation of 0.41066 with the response variable compared to the initial independent variables in the data (see table 3). TEAM_ BATTING_COMB_2 variable doesn't render a significant improvement (see table 3) although both variables fit the TARGET_WINS clusters fairly well (see sets of boxplots in the figure 4 below).

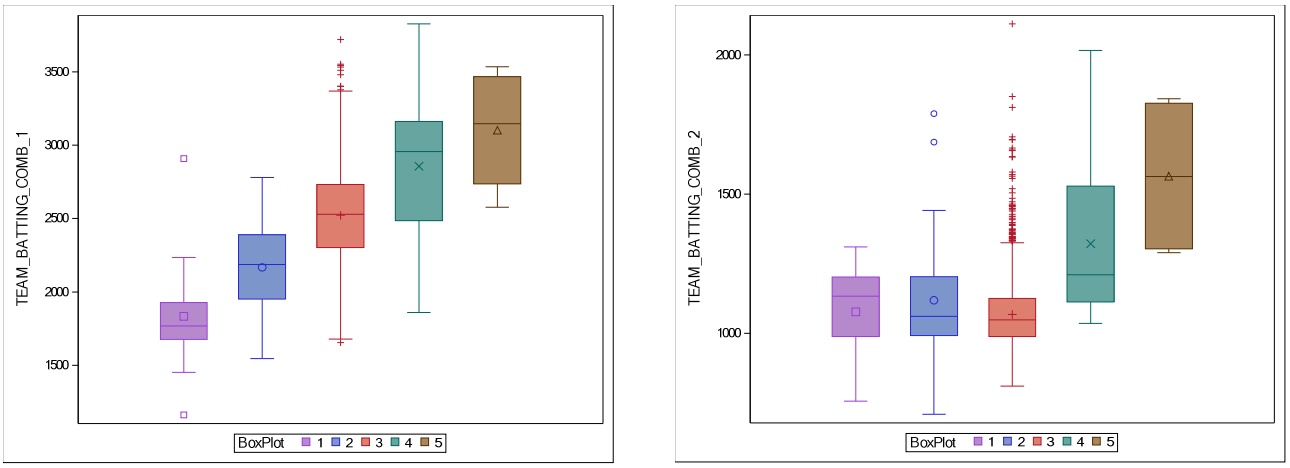


Figure 4: New Variables Box Plots

Data Preparation

Missing Values (see table_1 on page 1 for the variables and data used in fixing missing values)
There are six independent variables in the given data set that contain missing values. Variable TEAM_BATTING_HBP is dropped because over 90% of its values are missing.
Missing values for the following variables are imputed with the average value (variable medians are used as average values): TEAM_ BATTING_SO, TEAM_BASERUN_CS and TEAM_FIELDING_DP. Respective medians of the aforementioned variables are as follows: 750, 49, and 149.
Variables TEAM_BASERUN_SB and TEAM_PITCHING_SO have been imputed based on the decision trees developed in SAS Enterprise Mines setting up the above variables as Target (see Figure 5 and Figure 6 below). For simplicity of application (to hard-code the decision trees) only first 3 levels have been used.

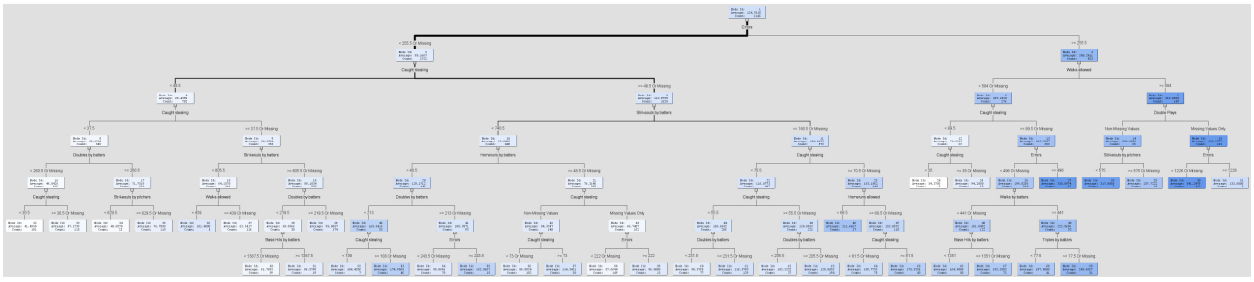


Figure 5: TEAM_BASERUN_SB decision tree

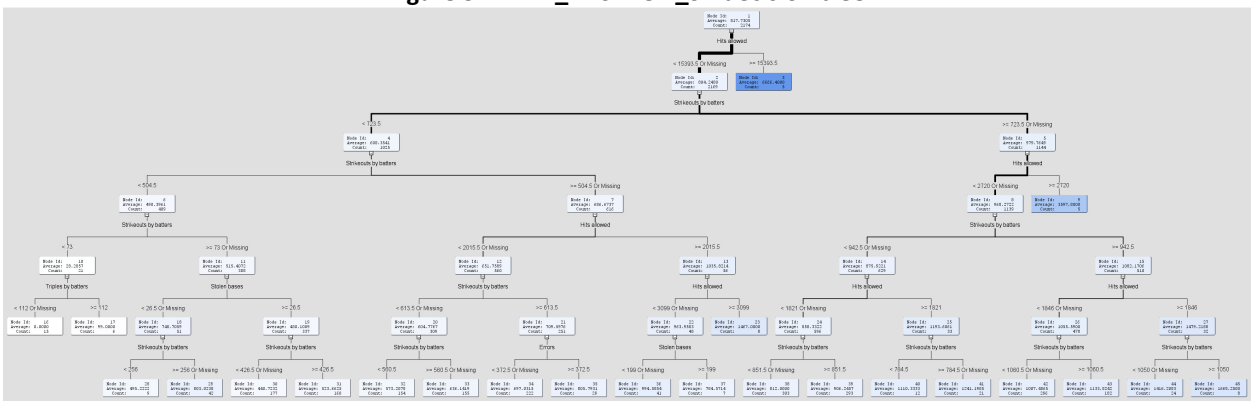


Figure 6: TEAM_PITCHING_SO decision tree

All Imputed variables have been renamed by adding IMP_ prefix to the variable name. Original variables were dropped from the data set after imputation had been completed. Also, flag variables were created for each of the imputed variables (with the prefix M_) in order to indicate that imputation has been performed (flag variable equals 1 if value was missing and 0 otherwise) and to check whether the fact that the value was missing can be predictive and therefore used in a linear regression model.

Variable Transformation

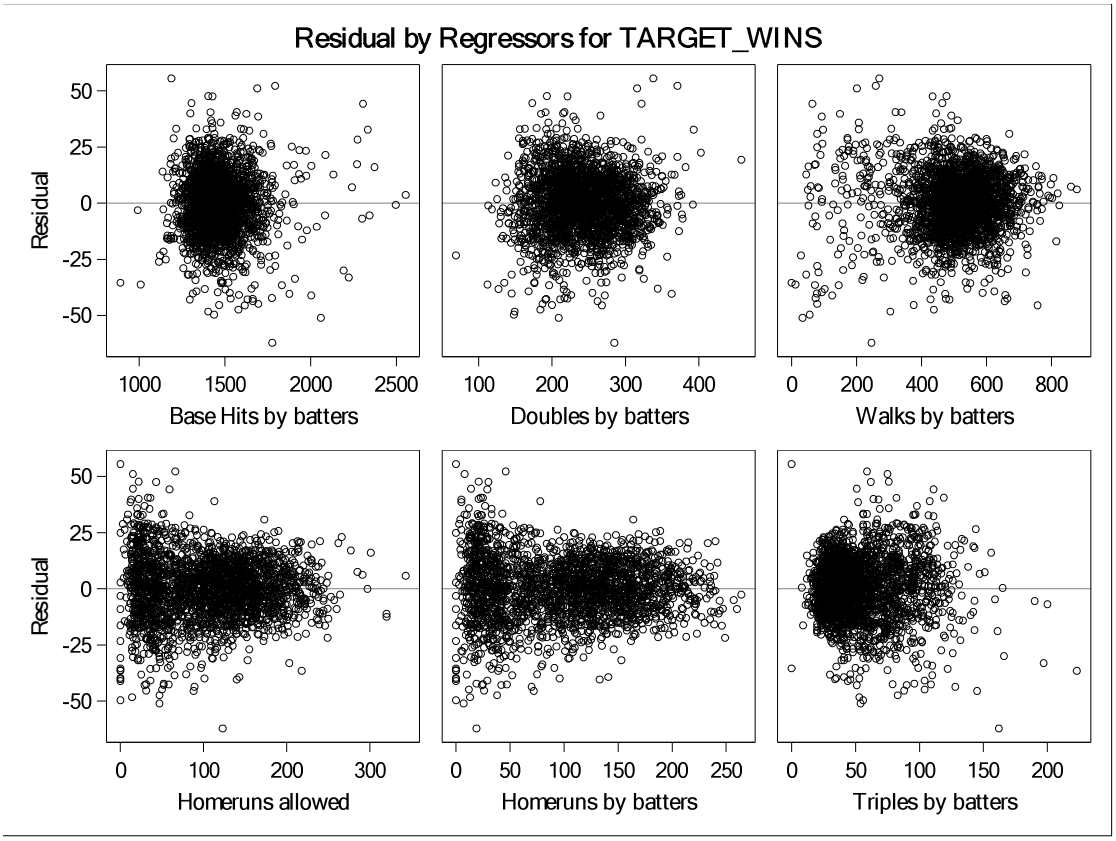


Figure 7: Scatterplots of residuals

One of the approaches used to determine method of variable transformation is the analysis of residuals. The residuals of predictor variables are analyzed in order to determine whether there exists an identifiable pattern. Four out of six variables in figure 7 above are distributed in relatively uniform way, whereas residuals of TEAM_PITCHIG_HR and TEAM_BATTING_HR variables have fan-shaped distribution. In order to eliminate (or at least reduce) this unfavorable pattern that could affect the accuracy of the final linear regression model, these two variables have to be transformed.
Another approach used during variable transformation was to apply set of transformation methods to each variable; namely, each variable has been subject to the following transformations (TEAM_BATTING_H variable code has been used in examples below; stat. parameters were hard-coded):

1) Natural logarithm transformation (after adding 1) with the initial sign retention;
   ex: LN_TEAM_BATTING_H = sign(TEAM_BATTING_H) * log(abs(TEAM_BATTING_H)+1);
2) Square root transformation with the initial sign retention;
   ex: SQRT_TEAM_BATTING_H = sign(TEAM_BATTING_H) * sqrt(abs(TEAM_BATTING_H));
3) Truncation of variable data (trimming level used is 99%);
   ex: T99_TEAM_BATTING_H = max(min(TEAM_BATTING_H,1950),1188);

4) Standardization of variable using Z-Transformation (trimming to 3 standard deviations)
   ex: STD_TEAM_BATTING_H = (TEAM_BATTING_H - 1469.27)/144.59;
   T_TEAM_BATTING_H = max(min(STD_TEAM_BATTING_H,3),-3);
5) Binning variable into buckets (4 equally-spaced buckets are used);
   ex: NORM_TEAM_BATTING_H = (TEAM_BATTING_H - 891) / (2554 - 891);
   BUCKET_TEAM_BATTING_H = min(int(&BIN.* NORM_TEAM_BATTING_H),&BIN.-1);

Additionally, each set of transformed variables has been put into OLS regression (using PROC REG) with TARGET_WINS as response variable (stepwise selection method has been used) in order to check which transformation will be retained in the model and which variables have greater absolute t-value.
Finally, all the original and transformed variables have been "dumped" into a stepwise regression model in order to check which transformation are going to be kept in the model and have predictive power. As the result of this analysis, the following set of transformations is applied in model building (see table 4).

#	Original Variable	Transformed Variable	Transformation Method	Details
1	TEAM_BATTING_H	T99_TEAM_BATTING_H	Trimming (99% level)	To constrain data outliers
2	TEAM_BATTING_2B	SQRT_TEAM_BATTING_2B	Square Root	To constrain data outliers
3	TEAM_BATTING_3B	T99_TEAM_BATTING_3B	Trimming (99% level)	To constrain data outliers
4	TEAM_BATTING_HR	LN_TEAM_BATTING_HR	Natural Logarithm	To constrain data outliers
5	TEAM_BATTING_BB	TEAM_BATTING_BB	No Transformation	Original Variable Retained
6	TEAM_BATTING_SO	IMP_TEAM_BATTING_SO	Missing Value Imputation	Median value of 750 used
		M_TEAM_BATTING_SO	New Binary Variable	Flag variable created to check whether there is any significance to the fact that the variable is missing
		BUCKET_IMP_TEAM_BATTING_SO	Binning	4 equally-spaced bins created
7	TEAM_BASERUN_SB	IMP_TEAM_BASERUN_SB	Missing Value Imputation	3-level decision tree is used
		M_TEAM_BASERUN_SB	New Binary Variable	Flag variable created to check whether there is any significance to the fact that the variable is missing
		SQRT_IMP_TEAM_BASERUN_SB	Square Root	To constrain data outliers
8	TEAM_BASERUN_CS	IMP_TEAM_BASERUN_CS	Missing Value Imputation	Median value of 49 used
		M_TEAM_BASERUN_CS	New Binary Variable	Flag variable created to check whether there is any significance to the fact that the variable is missing
		LN_IMP_TEAM_BASERUN_CS	Natural Logarithm	To constrain data outliers
9	TEAM_BATTING_HBP	Variable Dropped	No Transformation	Over 90% of variable values are missing
10	TEAM_PITCHING_H	BUCKET_TEAM_PITCHING_H	Binning	4 equally-spaced bins created
11	TEAM_PITCHING_HR	BUCKET_TEAM_PITCHING_HR	Binning	4 equally-spaced bins created

#	Original Variable	Transformed Variable	Transformation Method	Details
12	TEAM_PITCHING_BB	T_TEAM_PITCHING_BB	Standardizing (Z-Transform)	To constrain data outliers
13	TEAM_PITCHING_SO	IMP_TEAM_PITCHING_SO	Missing Value Imputation	Median value of 49 used
		M_TEAM_PITCHING_SO	New Binary Variable	Flag variable created to check whether there is any significance to the fact that the variable is missing
		T_IMP_TEAM_PITCHING_SO	Standardizing (Z-Transform)	To constrain data outliers
14	TEAM_FIELDING_E	SQRT_TEAM_FIELDING_E	Square Root	To constrain data outliers
15	TEAM_FIELDING_DP	IMP_TEAM_FIELDING_DP	Missing Value Imputation	Median value of 149 used
		M_TEAM_FIELDING_DP	New Binary Variable	Flag variable created to check whether there is any significance to the fact that the variable is missing
		LN_IMP_TEAM_FIELDING_DP	Natural Logarithm	To constrain data outliers
16	N/A	TEAM_BATTING_COMB_1	Newly created variable by combining a set of original variables (see details)	TEAM_BATTING_COMB_1 = TEAM_BATTING_H + 2 * TEAM_BATTING_2B + 3 * TEAM_BATTING_3B + 4 * TEAM_BATTING_HR
17	N/A	TEAM_BATTING_COMB_2	Newly created variable by combining a set of original variables (see details)	TEAM_BATTING_COMB_2 = TEAM_BATTING_H - TEAM_BATTING_2B - TEAM_BATTING_3B - TEAM_BATTING_HR

Table 4: Variable Transformation Summary

Transformations listed in the table 4 above will eliminate or reduce the influence of the potential outliers that otherwise could have an undue influence on the parameters in the model as well as improve the normality of distributions of predictor variables. Moreover, the two created variables (see items # 16 and 17 in table 4 above) have been retained to be used in the regression models.

Model Building

The training data was split into parts (66/34 split) into 1474 observations for the training set, in order to develop the models, and 802 observations for the testing set, the check the accuracy of the models. Four multiple linear regression models have been built (using PROC REG). First model (Manual) was built using the set of manually (arbitrarily) selected (based on the analyst's limited knowledge of baseball game) predictor variables. The following three were built using Forward, Backward, and Stepwise regression procedures applied to the set of all variables. Model validation metrics is the same for both Backward and Stepwise selected models therefore only backward regression model will be reviewed.

Manual Variable Selection Model

In order to evaluate selected model, diagnostics presented in tables 5 & 6 below is analyzed. The goal is to determine whether this model is an adequate predictor of the response variable TARGET_WINS.
The analysis of variance in table 5 shows that the null hypothesis that all the betas (parameter estimates) except for the intercept are zero should be rejected
The potential issue is p-score of the intercept which is unreasonably high. Intercept value is not significant at the level of 95% (p-values is 0.4983; see table 6). It should be viewed with caution.
Remaining predictor variables are all significant at 95% and therefore all can be considered as predictive.
Next, variance inflation factors (VIF – last column; table 6) for all the variables are less than 10 indicating that there is no multi-collinearity between the predictive variables. Only two VIFs exceed 5 but by less than 1. It has been shown that values from 5 to 10 might indicate multi-collinearity among variables.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	140370	17546	110.00	<.0001
Error	1465	233688	159.51417		
Corrected Total	1473	374058			

Table 5: Analysis of Variance; Manual Selection

The value signs of model parameters (betas) correspond to the theoretical effect of corresponding predictor variables for all the variables in the model except for SQRT_IMP_TEAM_BASERUN_SB (which is transformed variable for "Stolen Bases"). The parameter has negative value, which is counterintuitive because the above variable should have a positively impact on the number of wins.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	3.19123	4.71182	0.68	0.4983	0
TEAM_BATTING_COMB_1	1	0.01052	0.00233	4.52	<.0001	5.90573
SQRT_TEAM_FIELDING_E	1	-1.51004	0.11480	-13.15	<.0001	3.69812
T99_TEAM_BATTING_H	1	0.02874	0.00596	4.82	<.0001	5.63670
TEAM_BATTING_BB	1	0.02437	0.00391	6.23	<.0001	2.13047
SQRT_IMP_TEAM_BASERUN_SB	1	1.53582	0.12143	12.65	<.0001	1.72122
T_IMP_TEAM_PITCHING_SO	1	-5.81952	0.93717	-6.21	<.0001	1.91435
M_TEAM_BASERUN_SB	1	23.69617	2.06520	11.47	<.0001	2.21245
M_TEAM_PITCHING_SO	1	9.40600	1.67789	5.61	<.0001	1.20868

Table 6: Parameter Estimates; Manual Selection

Adjusted R-Squared is 0.3719 which is reasonably high and the number of predictor variables is only 8.

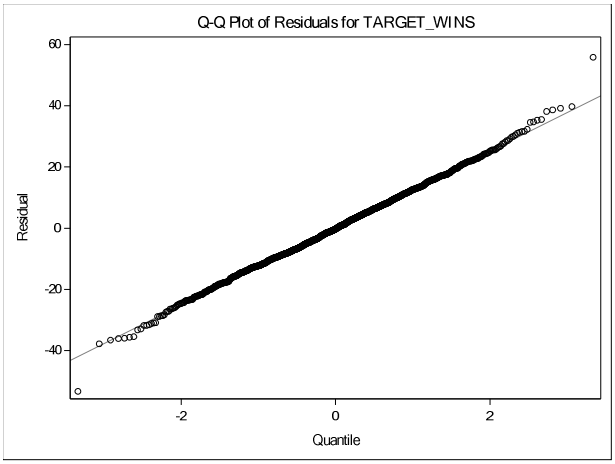


Figure 8: Normality Plot of Residuals

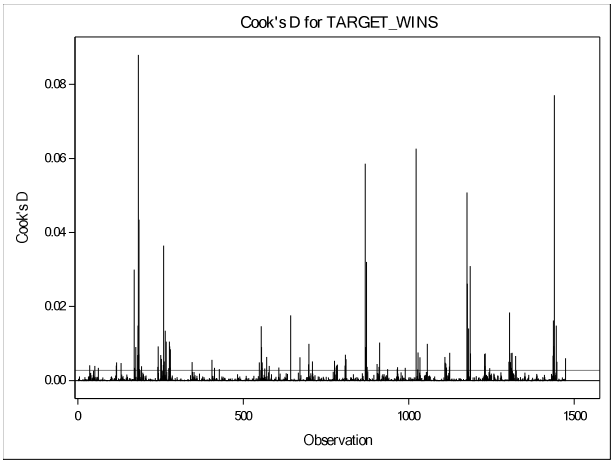


Figure 9: Cook's D influence Plot of Data

Normality of errors displayed in Figure 8 above appears to be linear and not to have any major deviations except for the values at the both ends of distribution. However, this deviation is not significant enough and it is reasonable to assume that the residual are distributed normally.
Examination of the Cook's Distance Values in the figure 9 above indicates that there are potentially significant outliers that might exert influence on the model. There are values that are marginally exceeding the limit. It warrants additional examination in order to determine the possible factors that are not reflected in the data. The possible reason could be that the data is from long period of time, namely that the data comes from the games played during the range of over hundred years.

Forward Variable Selection Model

In order to evaluate selected model, diagnostics presented in tables 7 & 8 below is analyzed. The goal is to determine whether this model is an adequate predictor of the response variable TARGET_WINS.
The analysis of variance in table 7 shows that the null hypothesis that all the betas (parameter estimates) except for the intercept are zero should be rejected and model is predictive.
The potential issues are p-scores of the M_TEAM_BASERUN_SC (flag variable for missing values) and T_IMP_TEAM_PITCHING_SO which are unreasonably high. Also, the intercept value of 105.20 is high: for all the predictor variables equal zero TARGET_WINS is 105.20, which means that the model predicts 105 wins in case a baseball team doesn't play any games and/or has all the performance indicators equal 0.
Remaining predictor variables are all significant at 95% and therefore all can be considered as predictive.
Next, variance inflation factors (VIF – last column; table 8) for some variables are greater than 10 indicating multi-collinearity between the predictive variables. Most likely the multi-collinearity is due to the fact that variable TEAM_BATTING_COMB_1 is created by combining a set of initial predictors.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	19	172124	9059.13684	65.23	<.0001
Error	1454	201935	138.88214		
Corrected Total	1473	374058			

Table 7: Analysis of Variance; Forward Selection

Another model property that requires additional scrutiny is the fact that value signs of 7 model parameters (betas) don't correspond to the theoretical effect of corresponding predictor variables. Two of these variables (mentioned above) are not significant at 95% and potentially can be removed.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	105.19888	14.88163	7.07	<.0001	0
TEAM_BATTING_COMB_1	1	0.02254	0.00416	5.42	<.0001	21.65485
TEAM_BATTING_COMB_2	1	0.08204	0.00850	9.65	<.0001	12.36281
T99_TEAM_BATTING_H	1	-0.06037	0.01209	-4.99	<.0001	26.62241
T99_TEAM_BATTING_3B	1	0.13700	0.02054	6.67	<.0001	3.25202
LN_TEAM_BATTING_HR	1	1.01495	1.07737	0.94	0.3463	9.91688
TEAM_BATTING_BB	1	0.04466	0.00839	5.32	<.0001	11.27429
BUCKET_IMP_TEAM_BATTING_SO	1	-3.82512	1.00626	-3.80	0.0001	5.46060
SQRT_IMP_TEAM_BASERUN_SB	1	1.32698	0.15252	8.70	<.0001	3.11883
M_TEAM_BASERUN_SB	1	31.23367	2.14780	14.54	<.0001	2.74846
LN_IMP_TEAM_BASERUN_CS	1	-3.12739	1.15973	-2.70	0.0071	1.57419
M_TEAM_BASERUN_CS	1	1.21465	1.08896	1.12	0.2649	2.82652
BUCKET_TEAM_PITCHING_H	1	12.83671	3.02171	4.25	<.0001	2.41287
BUCKET_TEAM_PITCHING_HR	1	2.44245	0.96966	2.52	0.0119	5.10901
T_TEAM_PITCHING_BB	1	-2.48121	1.11467	-2.23	0.0262	7.37115
T_IMP_TEAM_PITCHING_SO	1	-0.10465	1.38489	-0.08	0.9398	4.80138
M_TEAM_PITCHING_SO	1	8.12815	2.03559	3.99	<.0001	2.04323
SQRT_TEAM_FIELDING_E	1	-2.68291	0.18114	-14.81	<.0001	10.57504
LN_IMP_TEAM_FIELDING_DP	1	-15.47047	2.34628	-6.59	<.0001	1.92833
M_TEAM_FIELDING_DP	1	7.42227	1.90707	3.89	0.0001	4.03864

Table 8: Parameter Estimates; Forward Selection

Adjusted R-Squared is 0.4531 which is fairly high and the number of predictor variables is 19. Normality of errors displayed in Figure 10 below appears to be linear and not to have any major deviations except for the values at the both ends of distribution. However, this deviation is not significant enough and it is reasonable to assume that the residual are distributed normally. Examination of the Cook's Distance Values in the figure 11 below indicates that there are few potentially significant outliers that might exert influence on the model. There are values that marginally exceed the limit. It warrants additional examination of data in order to determine the possible factors that are not reflected. Another approach that could produce good results would be simply removing these observations (since these are clearly outliers) and rerunning the model.

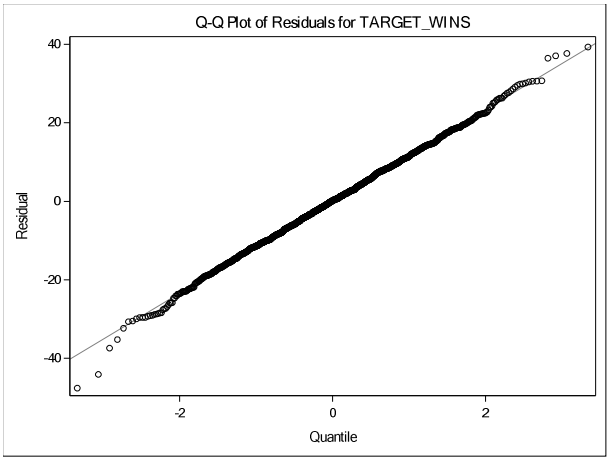


Figure 10: Normality Plot of Residuals

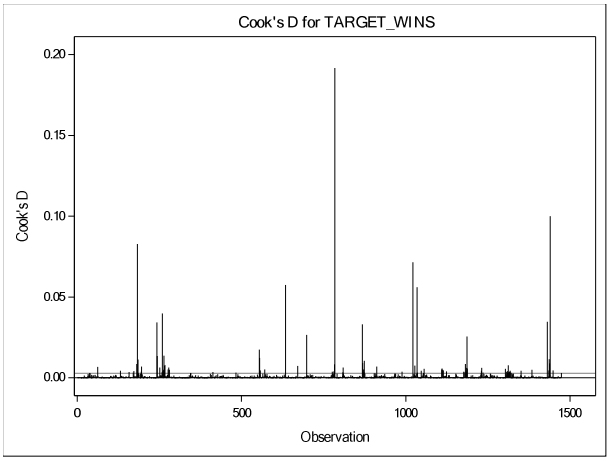


Figure 11: Cook's D influence Plot of Data

Based on the above analysis of the parameters and diagrams of the selected model, it can be assumed that the forward selection model doesn't explicitly violate any of the OLS regression assumptions.

Backward Selection

The analysis of variance in table 9 shows that the null hypothesis that all the betas (parameter estimates) except for the intercept are zero should be rejected and the model is predictive. All predictor variables including intercept are all significant at 95% and therefore all can be considered as predictive. The intercept value of 111.22 is high: for all the predictor variables equal to zero, TARGET_WINS response variable is equal to the intercept value, which means that the model predicts over 111 wins in case a baseball team doesn't play any games and/or has all the performance indicators equal to 0. Variance inflation factors (VIF – last column; table 10 below) for some variables are greater than 10 indicating multi-collinearity between the predictor variables. Most likely the multi-collinearity is due to the fact that variables TEAM_BATTING_COMB_1 and TEAM_BATTING_COMB_2 are created by combining a set of initial predictors and therefore are highly correlated with TEAM_BATTING_H variable. Four betas have coefficient signs that do not correspond to the impact (positive or negative) on the outcome. T99_TEAM_BATTIN_H and LN_IMP_TEAM_FIELDING_DP have negative sings but positive impact whereas BUCKET_TEAM_PITCHING_H and BUCKET_TEAM_PITCHING_HR have positive signs and supposedly negative impact. Due to analyst's limited understanding of baseball a more extensive analysis in order to determine the reason of the above discrepancy cannot be performed at this stage.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	16	171831	10739	77.38	<.0001
Error	1457	202227	138.79694		
Corrected Total	1473	374058			

Table 9: Analysis of Variance; Backward Selection

Adjusted R-Squared is 0.4534 which is reasonably high and the number of predictor variables is only 16. This model has slightly greater Adj. R-Squared than Forward Selection model and fewer coefficients.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	111.22428	14.01846	7.93	<.0001	0
TEAM_BATTING_COMB_1	1	0.02382	0.00364	6.54	<.0001	16.61689
TEAM_BATTING_COMB_2	1	0.08170	0.00815	10.03	<.0001	11.37171
T99_TEAM_BATTING_H	1	-0.06171	0.01154	-5.35	<.0001	24.25320
T99_TEAM_BATTING_3B	1	0.13718	0.02005	6.84	<.0001	3.10076
TEAM_BATTING_BB	1	0.04531	0.00709	6.39	<.0001	8.05666
BUCKET_IMP_TEAM_BATTING_SO	1	-3.98804	0.76792	-5.19	<.0001	3.18213
M_TEAM_BATTING_SO	1	8.15999	1.85223	4.41	<.0001	1.69276
SQRT_IMP_TEAM_BASERUN_SB	1	1.33348	0.14943	8.92	<.0001	2.99554
M_TEAM_BASERUN_SB	1	31.00240	2.11954	14.63	<.0001	2.67824
LN_IMP_TEAM_BASERUN_CS	1	-3.38315	1.12903	-3.00	0.0028	1.49286
BUCKET_TEAM_PITCHING_H	1	11.51203	2.77223	4.15	<.0001	2.03214
BUCKET_TEAM_PITCHING_HR	1	2.53265	0.95845	2.64	0.0083	4.99468
T_TEAM_PITCHING_BB	1	-2.54951	0.88957	-2.87	0.0042	4.69751
SQRT_TEAM_FIELDING_E	1	-2.70119	0.17676	-15.28	<.0001	10.07636
LN_IMP_TEAM_FIELDING_DP	1	-15.68397	2.30640	-6.80	<.0001	1.86449
M_TEAM_FIELDING_DP	1	7.77966	1.88976	4.12	<.0001	3.96812

Table 10: Parameter Estimates; Backward Selection

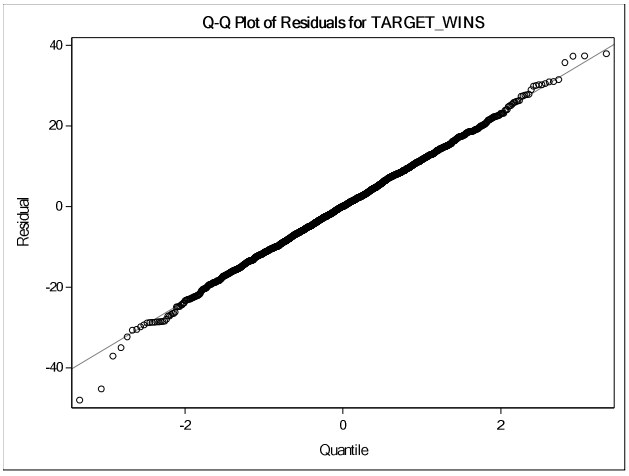


Figure 12: Normality Plot of Residuals

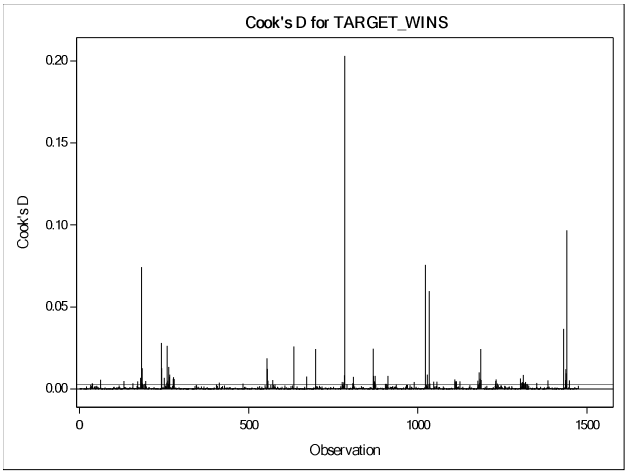


Figure 13: Cook's D influence Plot of Data

Normality of errors displayed in Figure 12 above appears to be linear and not to have any major deviations except for the values at the both ends of distribution. However, this deviation is not significant enough and it is reasonable to assume that the residual are distributed normally.
Examination of the Cook's Distance Values in the figure 13 above indicates that there are few potentially significant outliers that might exert influence on the model. There are values that marginally exceed the limit. It warrants additional examination of data in order to determine the possible factors that are not reflected in the available. The possible reason could be that the data is from a long period of time, namely that the data comes from the games played during the range of over hundred years.
Based on the above analysis of the parameters and diagrams of the selected model, it can be assumed that the forward selection model doesn't explicitly violate any of the OLS regression assumptions.

Model Selection

In order to select the best model out of 3 available, a list of diagnostic measures has been generated for each selection during the model building process. These metrics are displayed in table 11 (see below). The model validation measures in this table are based on the training data set containing 1474 observations. This is the training set that all four regression models have been built on.
Forward, Backward, and Stepwise selection approaches differ from the optimal model search algorithms because they are based on heuristic approach and only subsets of possible models are considered. Forward, Backward, and Stepwise selection techniques produce similar outputs with Backward and Stepwise selections having exactly the same metrics and manual model performing slightly worse.

Obs	_MODEL_	RMSE	IN	P	EDF	RSQ	ADJRSQ	CP	AIC	BIC	SBC
1	Manual	12.63	8	9	1465	0.37526	0.37185	9	7485.30	7487.41	7532.96
2	Forward	11.78	19	20	1454	0.46015	0.45310	19.0039	7292.03	7294.60	7397.94
3	Backward	11.78	16	17	1457	0.45937	0.45343	15.1086	7288.16	7290.60	7378.19
4	Stepwise	11.78	16	17	1457	0.45937	0.45343	15.1086	7288.16	7290.60	7378.19

Table 11: Diagnostic Measures for Regression Models (produced by SAS)

Based on CP metrics, better model is the one that has CP value close to P with smaller P value. AIC, SBC, and BIC try to balance the conflicting demands of accuracy (fit) and simplicity (small number of variable). Models with smaller values are preferred but, for example in case of AIC, the rule of thumb is to consider models with AIC values different by less than two as equal. Based on the metrics discussed so far, Backward and Stepwise selections perform better: adjusted R-Squared value is greater and AIC, BIC, and SBC metrics values are smaller. Moreover, regression models derived by Backward and Stepwise selection techniques have fewer independent variables and therefore are easier to interpret and deploy. It follows then that the model derived through Backward or Stepwise selection should be used.

In addition, all four models have been tested against the test data set (contains 802 observations) that was set aside for testing purposes and was not used in model building process. The results are very similar to the ones in the table above: Backward and Stepwise selection models perform better. Moreover, the validation metrics in the table below have been computed manually (based on SSE). It is worth noticing that the adjusted R-Squared values in the table 12 below are much lower compared to those in the table above indicating that the regression models don't perform as good on the test set as on the training set. This indicates that adjusted R- Squared value produced in the SAS output is not always a very reliable indicator of how a particular regression model will perform on a set of new data.

Model	N	SSE	K	SST	D	ADJ RSQ	AIC	SBC	AICC
Manual	802	129164	9	190258.3	161.0524	0.3143	4093.55	4135.73	4093.77
Forward	802	117912	20	190258.3	147.0224	0.3652	4042.45	4136.19	4043.52
Backward	802	118112	17	190258.3	147.2718	0.3665	4037.81	4117.49	4038.59
Stepwise	802	118112	17	190258.3	147.2718	0.3665	4037.81	4117.49	4038.59

Table 12: Diagnostic Measures for Regression Models Compute Manually

In addition, the training set of transformed variables has been uploaded into SAS Enterprise Miner. The four models have been built using the following techniques: Regression (backward selection), Neural Network, Decision Tree, and Gradient Boosting. Based on the EM output, the best performing model is the one built using Neural Network (see figure below). Nevertheless, the Backward Regression model is second-best which confirms that linear regression is very robust and can be powerful despite simplicity.

Fit Statistics Table
Target: TARGET_WINS

Data Role=Train

Statistics	Neural	Reg	Tree	Boost
Train: Akaike's Information Criterion	7052.72	7288.16	.	.
Train: Average Squared Error	108.83	137.20	143.17	164.99
Train: Average Error Function	108.83	137.20	.	.
Selection Criterion: Train: Average Squared Error	108.83	137.20	143.17	164.99
Train: Degrees of Freedom for Error	1404.00	1457.00	.	.
Train: Model Degrees of Freedom	70.00	17.00	.	.
Train: Total Degrees of Freedom	1474.00	1474.00	1474.00	1474.00
Train: Divisor for ASE	1474.00	1474.00	1474.00	1474.00
Train: Error Function	160412.84	202227.14	.	.
Train: Final Prediction Error	119.68	140.40	.	.
Train: Maximum Absolute Error	43.82	48.01	47.11	60.83
Train: Misclassification Rate	.	.	.	.
Train: Mean Square Error	114.25	138.80	.	.
Train: Sum of Frequencies	1474.00	1474.00	1474.00	1474.00
Train: Number of Estimate Weights	70.00	17.00	.	.
Train: Root Average Sum of Squares	10.43	11.71	11.97	12.84
Train: Root Final Prediction Error	10.94	11.85	.	.
Train: Root Mean Squared Error	10.69	11.78	.	.
Train: Schwarz's Bayesian Criterion	7423.42	7378.19	.	.
Train: Sum of Squared Errors	160412.84	202227.14	211038.75	243192.07
Train: Sum of Case Weights Times Freq	1474.00	1474.00	.	1474.00
Train: Number of Wrong Classifications	.	.	.	.

Figure 14: SAS Enterprise Miner Output

As the result of the above analysis the optimal model selected and used in the model deployment is:

P_TARGET_WINS =	111.22428
+ TEAM_BATTING_COMB_1	* 0.02382
+ TEAM_BATTING_COMB_2	* 0.08170
+ T99_TEAM_BATTING_H	* -0.06171
+ T99_TEAM_BATTING_3B	* 0.13718
+ TEAM_BATTING_BB	* 0.04531
+ BUCKET_IMP_TEAM_BATTING_SO	* -3.98804
+ M_TEAM_BATTING_SO	* 8.15999
+ SQRT_IMP_TEAM_BASERUN_SB	* 1.33348
+ M_TEAM_BASERUN_SB	* 31.00240
+ LN_IMP_TEAM_BASERUN_CS	* -3.38315
+ BUCKET_TEAM_PITCHING_H	* 11.51203
+ BUCKET_TEAM_PITCHING_HR	* 2.53265
+ T_TEAM_PITCHING_BB	* -2.54951
+ SQRT_TEAM_FIELDING_E	* -2.70119
+ LN_IMP_TEAM_FIELDING_DP	* -15.68397
+ M_TEAM_FIELDING_DP	* 7.77966;

The values of the equation above make sense for the most part. Exception is “Doubles by Batter” and “Double Plays”, which have negative beta values. This is counterintuitive because the above variables should positively impact the number of wins. Also, “Hits allowed” and “Homeruns allowed” variables have positive coefficients and expected negative impact. It is not obvious however, whether these metrics actually respectively increase or reduce the chance of winning. More extensive knowledge of the baseball game and deeper analysis is needed in order to determine whether that is the case.

Conclusion

Several techniques were used to develop a model to predict the number of wins for a baseball team during a regular season. The model chosen for deployment was derived using regression backward selection approach and was determined to be optimal (out of 4 models analyzed) based on the values of number of diagnostics measures and fewer predictor variables employed in the model. The few issues with the model that warrant further investigation are intercept value and several predictor variables coefficients which seems to be counterintuitive based on the sign of coefficient compared with the theoretical effect. Although there are ways to derive a better model (the one that has stronger predictive power and accuracy) like a model developed using Neural Network or by applying more effective variable imputation and transformation techniques (with help of SAS EM), model presented can be useful in predicting the number of wins using standard baseball performance statistics.