Introduction

The objective of the following exercise is to generate an "optimal" linear regression model based on the Moneyball data set to predict the number of wins for a given baseball team.

The subsequent write-up and analysis will detail the steps of data exploration and preparation to ensure the data set is analysis-ready. Trial and error data transformations will be executed to impute missing values as necessary and create new variables for inclusion into the models tested.

Following that will be a review of the model building and selection steps to determine the "best" model. A combination of different variables and variable selection methods will be tested to arrive at the final model.

Results

1 & 2. Data Exploration & Data Preparation

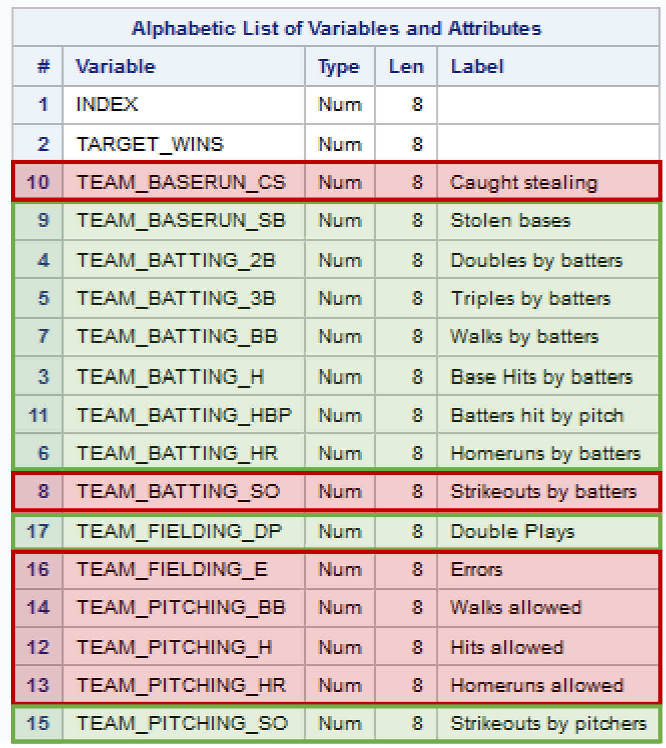
This section of the report reviews the data exploration and data preparation steps that are used to finalize an analysis-ready data set. The **data exploration sections** (headlined in bold black text) to follow provide details about the Moneyball data set, while the **data preparation sections** (headlined in bold blue text) provide details about specific data transformation. Throughout the following discourse, names of data tables that appear are *italicized* for convenience.

The Moneyball data set includes 2276 records/observations, each detailing a set of attributes for a given baseball team's performance, adjusted to match performance across a 162 game season.

Data are collected across 15 different performance variables, as shown below in the *Alphabetic List of Variables and Attributes* table below. Each variable's expected direction of influence on the number of team wins is also indicated. Green shading represents a positive effect on the number of wins, while red shading represents a negative effect on the number of wins. All 15 of the original predictor variables in the Moneyball data set are numeric.

Positive effect on wins

Negative effect on wins



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
1	INDEX	Num	8	
2	TARGET_WINS	Num	8	
10	TEAM_BASERUN_CS	Num	8	Caught stealing
9	TEAM_BASERUN_SB	Num	8	Stolen bases
4	TEAM_BATTING_2B	Num	8	Doubles by batters
5	TEAM_BATTING_3B	Num	8	Triples by batters
7	TEAM_BATTING_BB	Num	8	Walks by batters
3	TEAM_BATTING_H	Num	8	Base Hits by batters
11	TEAM_BATTING_HBP	Num	8	Batters hit by pitch
6	TEAM_BATTING_HR	Num	8	Homeruns by batters
8	TEAM_BATTING_SO	Num	8	Strikeouts by batters
17	TEAM_FIELDING_DP	Num	8	Double Plays
16	TEAM_FIELDING_E	Num	8	Errors
14	TEAM_PITCHING_BB	Num	8	Walks allowed
12	TEAM_PITCHING_H	Num	8	Hits allowed
13	TEAM_PITCHING_HR	Num	8	Homeruns allowed
15	TEAM_PITCHING_SO	Num	8	Strikeouts by pitchers

In the data set, "INDEX" is a numeric variable assigned to reference a specific baseball team (e.g., 8, 9) and "TARGET_WINS" represents the number of wins by each team.

The subsequent data exploration discusses correlations (relationships between the dependent variable and independent variables, relationships among the independent variables), outliers (large values, "0" values), data transformations (log, square root), missing data (imputation, flag creation), and interaction terms.

Data Exploration: Understanding Correlations (Two Parts Discussed Below)
Two sets of correlations are of interest when building a linear regression model. The correlations between the dependent variable (TARGET_WINS) and independent variables can help better understand how the dependent variable is influenced by the independent variables, and the correlations among the independent variables can help better understand how the independent variables are related to and influence each other.

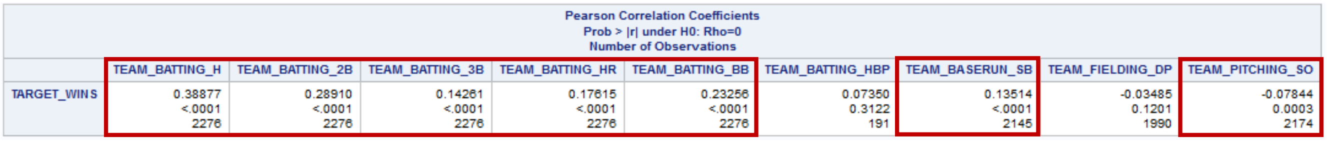
Understanding Correlations, Part 1: Relationships between dependent and independent variables
Let's begin by looking at the correlations between the dependent variable (TARGET_WINS) and the independent variables.

The two *Pearson Correlation Coefficients* tables below represent the correlations between the dependent variable and the independent variables that have a positive effect on wins and the correlations between the dependent variable and the independent variables that have a negative effect on wins, respectively.

From the first *Pearson Correlation Coefficients* table, it appears that the number of times players get on base, either from base hits (TEAM_BATTING_H, TEAM_BATTING_2B, TEAM_BATTING_3B, TEAM_BATTING_HR), walks (TEAM_BATTING_BB), or stolen bases (TEAM_BASERUN_SB) has a significant relationship with the number of wins at the 0.05 level. Pitching strikeouts (TEAM_PITCHING_SO) also appears to have a relationship that is significant at the 0.05 level.

Of these variables, TEAM_BATTING_H/TEAM_BATTING_2B/TEAM_BATTING_BB appear to have relatively stronger relationships with TARGET_WINS, as the associated correlations are relatively larger in magnitude. All variables mentioned above will be important to monitor during model building.

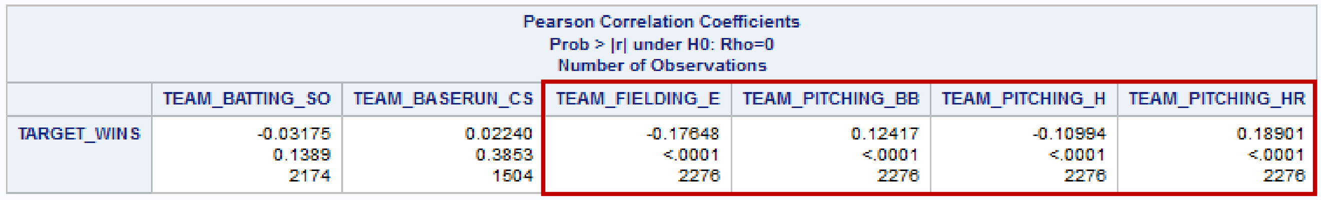
The correlations between TEAM_BATTING_HBP/TEAM_FIELDING_DP and TARGET_WINS are among the weaker relationships in this group. They are also not statistically significant. Alone, these two variables may receive lower consideration for inclusion during model building.



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	TEAM_BATTING_H	TEAM_BATTING_2B	TEAM_BATTING_3B	TEAM_BATTING_HR	TEAM_BATTING_BB	TEAM_BATTING_HBP	TEAM_BASERUN_SB	TEAM_FIELDING_DP	TEAM_PITCHING_SO
TARGET_WINS	0.38877 <.0001 2276	0.28910 <.0001 2276	0.14261 <.0001 2276	0.17615 <.0001 2276	0.23256 <.0001 2276	0.07350 0.3122 191	0.13514 <.0001 2145	-0.03485 0.1201 1990	-0.07844 0.0003 2174

Some of the other variables in the Moneyball data set also appear to have some statistically significant, though slightly smaller influence on the number of team wins. TEAM_FIELDING_E, TEAM_PITCHING_BB, TEAM_PITCHING_H, and TEAM_PITCHING_HR have a significant relationship with TARGET_WINS at the 0.05 level and might be of interest for inclusion during model building.



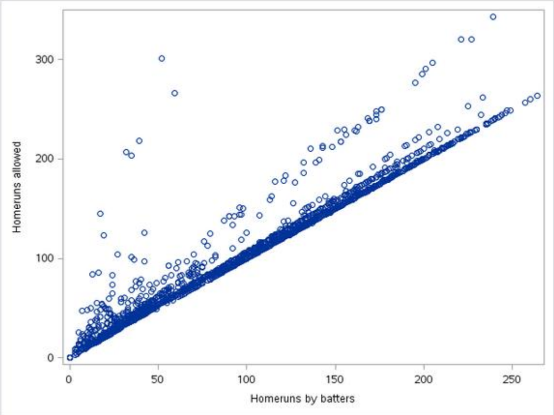
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	TEAM_BATTING_SO	TEAM_BASERUN_CS	TEAM_FIELDING_E	TEAM_PITCHING_BB	TEAM_PITCHING_H	TEAM_PITCHING_HR
TARGET_WINS	-0.03175 0.1389 2174	0.02240 0.3853 1504	-0.17648 <.0001 2276	0.12417 <.0001 2276	-0.10994 <.0001 2276	0.18901 <.0001 2276

Understanding Correlations, Part 2: Relationships among independent variables
Now let's look at the correlations among the independent variables. Strong correlations may indicate strong linear relationships between variables.

The following pairs of variables appear to have some relatively strong (both negative and positive) relationships, which may be of interest for developing interaction terms to test in the subsequent model building phase. The respective correlations are included to the right:

- TEAM_BATTING_H and TEAM_BATTING_2B: 0.56285
- TEAM_BATTING_3B and TEAM_BATTING_HR: -0.63557
- TEAM_BATTING_3B and TEAM_BATTING_SO: -0.66978
- TEAM_BATTING_HR and TEAM_BATTING_SO: 0.72707
- TEAM_BATTING_HR and TEAM_PITCHING_HR: 0.96937
  - This correlation value is extremely high, indicating that these two variables have a nearly perfect linear relationship. In model building, we can further explore this relationship and decide whether it would be reasonable to remove one of the two variables in the final model. Here is the scatterplot showing the relationship between these two variables:



- TEAM_BATTING_HR and TEAM_FIELDING_E: -0.58734
- TEAM_BATTING_BB and TEAM_FIELDING_E: -0.65597
- TEAM_BATTING_SO and TEAM_PITCHING_HR: 0.66718
- TEAM_BATTING_SO and TEAM_FIELDING_E: -0.58466
- TEAM_BASERUN_SB and TEAM_BASERUN_CS: 0.65524
- TEAM_PITCHING_H and TEAM_FIELDING_E: 0.66776

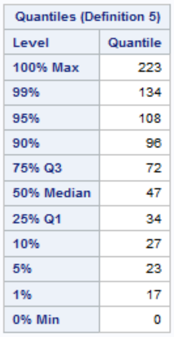
Data Exploration: Reconciling Outliers (Two Part Discussion Below)

Outliers in the data may skew final results, so it is important to identify and reconcile them. One way of identifying outliers is by comparing the mean and the median values for a given variable. If the mean value is very different from the median value, one or more outliers may be the culprit. Another way of identifying outliers is by looking at a variable's distribution. The latter method is used in the following analysis.

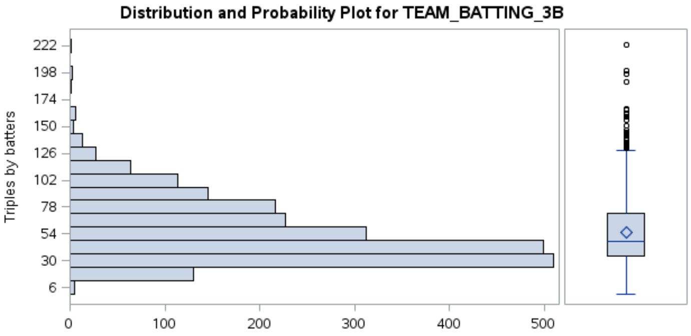
Reconciling Outliers, Part 1: Large Values

Large outliers should be immediately removed or capped from the data before further data preparation and analysis, as these have great potential to influence regression results. The following method removes/caps some values that fall outside the 99+ percentile for variables with moderately/very skewed distribution.

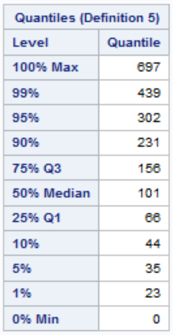
> **Data Preparation: Delete/cap extremely large values for TEAM_BATTING_3B, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_PITCHING_BB, TEAM_PITCHING_SO**
>
> From the distribution plots shown on the following pages, there appears to be a noticeable tail/skew among a handful of variables in the Moneyball data set. The extreme outliers may skew the final results and so are deleted and/or capped at the 99 percentile or higher (if there appear to be only a few extreme data points).



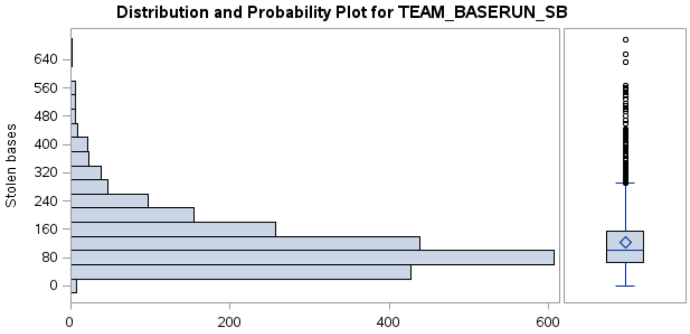
Quantiles (Definition 5)	
Level	Quantile
100% Max	223
99%	134
95%	108
90%	96
75% Q3	72
50% Median	47
25% Q1	34
10%	27
5%	23
1%	17
0% Min	0



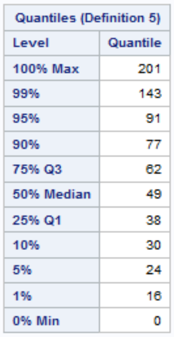
Data preparation for TEAM_BATTING_3B: Cap values > 175 at 175 and/or delete values over 175



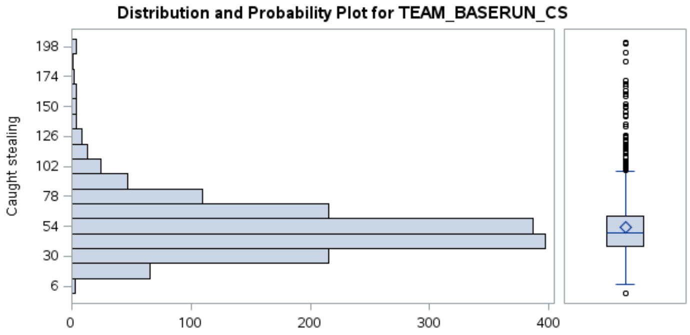
Quantiles (Definition 5)	
Level	Quantile
100% Max	697
99%	439
95%	302
90%	231
75% Q3	156
50% Median	101
25% Q1	66
10%	44
5%	35
1%	23
0% Min	0



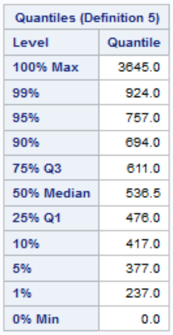
Data preparation for TEAM_BASERUN_SB: Cap values > 600 at 600 and/or delete values over 600



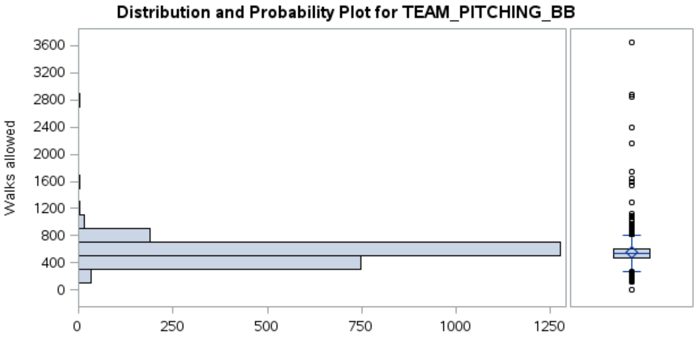
Quantiles (Definition 5)	
Level	Quantile
100% Max	201
99%	143
95%	91
90%	77
75% Q3	62
50% Median	49
25% Q1	38
10%	30
5%	24
1%	16
0% Min	0



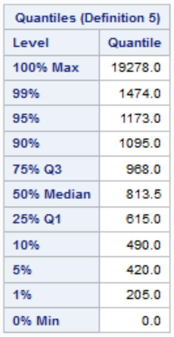
Data preparation for TEAM_BASERUN_CS: Cap values > 175 at 175 and/or delete values over 175



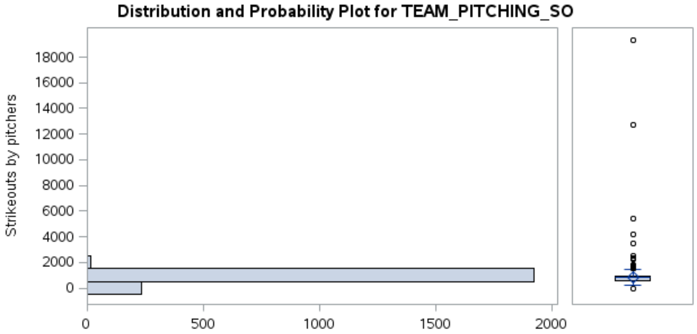
Quantiles (Definition 5)	
Level	Quantile
100% Max	3645.0
99%	924.0
95%	757.0
90%	694.0
75% Q3	611.0
50% Median	536.5
25% Q1	476.0
10%	417.0
5%	377.0
1%	237.0
0% Min	0.0



Data preparation for TEAM_PITCHING_BB: Cap values > 2000 at 2000 and/or delete values over 2000



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



Data preparation for TEAM_PITCHING_SO: Cap values > 2500 at 2500 and/or delete values over 2500

Reconciling Outliers, Part 2: Values of "0"

Values of zero throughout the data set may also merit further examination, as these may also skew final regression results. If there are many zeros for a given variable and the variable's mean is not close to zero, creation and inclusion of a flag variable may be reasonable.

Data preparation: Delete observations where INDEX in (1347, 1494, 1769)

These records have a value of zero for one or more of the following variables: TARGET_WINS, TEAM_BATTING_3B, TEAM_BASERUN_SB. The mean values of these variables are noticeably different from zero, so it is unlikely that zero is a reasonable value. Also, there are only one or two observations with a value of "0" for these variables in the entire data set, so these are likely outliers and can be deleted from the data set.

Data preparation: Create flag variable for TEAM_BATTING_SO and TEAM_PITCHING_HR

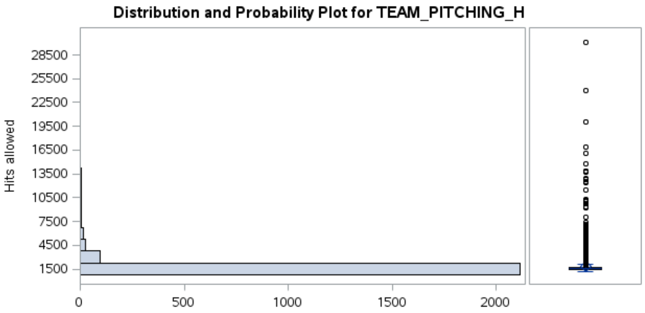
The mean values for these variables are noticeably different from zero, so it is unlikely that zero is a reasonable value in these cases. Since there are a non-insignificant number of "0" values, we should create a flag variable for each of these two variables to capture the presence/absence of zeroes.

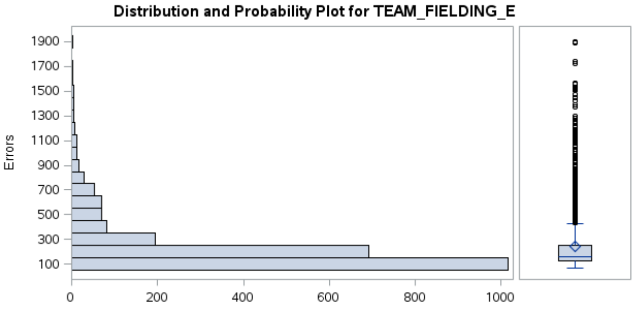
Data Exploration: Logarithmic & Square Root Transformations

Data for some variables may appear to have an extremely long tail of data, with most data appearing as "outliers." These variables can be logarithmically transformed to achieve more linear relationships.

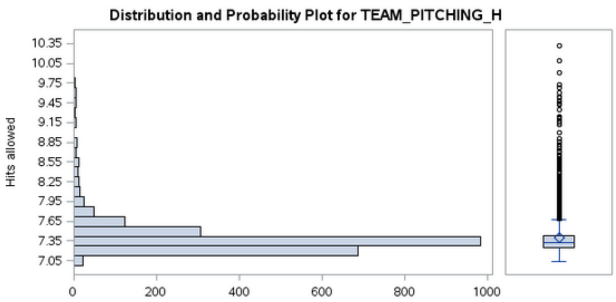
Data preparation: Log transformation of TEAM_PITCHING_H and TEAM_FIELDING_E

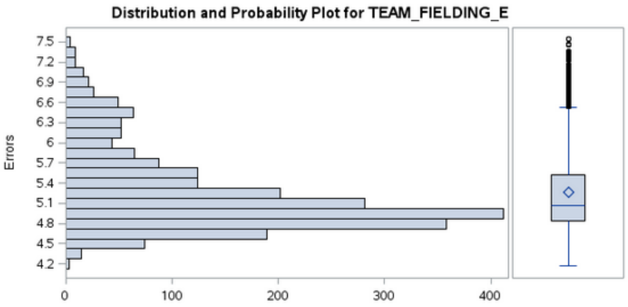
Two variables that fit the description above are TEAM_PITCHING_H and TEAM_FIELDING_E. These variables have extremely skewed distributions with many outliers, as shown by the long tails in the following graphs.





The transformed variables, which have noticeably shorter tails would be used in model building in lieu of the original variables.





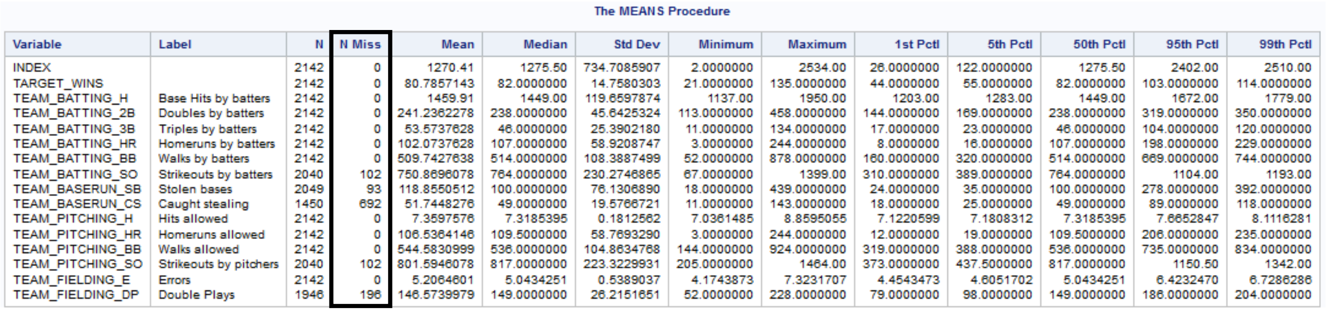
Data Preparation: Square root transformation of TEAM_BASERUN_SB

In an effort to linearize the data, the TEAM_BASERUN_SB variable is transformed via a square root function.

Data Exploration: Missing Data

Six of the variables in the *Moneyball* data set have missing data, as indicated in the "N Miss" column of *The MEANS Procedure* table below. Data could be missing for several reasons, including unavailability during data collection. For linear regression, missing data needs to be imputed before proceeding with model building.

Note: If a variable is missing too many values (i.e. 90-95%), one option could be to remove that variable from the data set and subsequent analyses or to create a flag to represent the presence or absence of data.



The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum	1st Pctl	5th Pctl	50th Pctl	95th Pctl	99th Pctl
INDEX		2142	0	1270.41	1275.50	734.7085907	2.0000000	2534.00	26.0000000	122.0000000	1275.50	2402.00	2510.00
TARGET_WINS		2142	0	80.7857143	82.0000000	14.7580303	21.0000000	135.0000000	44.0000000	55.0000000	82.0000000	103.0000000	114.0000000
TEAM_BATTING_H	Base Hits by batters	2142	0	1459.91	1449.00	119.6597874	1137.00	1950.00	1203.00	1283.00	1449.00	1672.00	1779.00
TEAM_BATTING_2B	Doubles by batters	2142	0	241.2362278	238.0000000	45.6425324	113.0000000	458.0000000	144.0000000	169.0000000	238.0000000	319.0000000	350.0000000
TEAM_BATTING_3B	Triples by batters	2142	0	53.5737628	46.0000000	25.3902180	11.0000000	134.0000000	17.0000000	23.0000000	46.0000000	104.0000000	120.0000000
TEAM_BATTING_HR	Homeruns by batters	2142	0	102.0737628	107.0000000	58.9208747	3.0000000	244.0000000	8.0000000	16.0000000	107.0000000	198.0000000	229.0000000
TEAM_BATTING_BB	Walks by batters	2142	0	509.7427638	514.0000000	108.3887499	52.0000000	878.0000000	160.0000000	320.0000000	514.0000000	669.0000000	744.0000000
TEAM_BATTING_SO	Strikeouts by batters	2040	102	750.8696078	764.0000000	230.2746865	67.0000000	1399.00	310.0000000	389.0000000	764.0000000	1104.00	1193.00
TEAM_BASERUN_SB	Stolen bases	2049	93	118.8550512	100.0000000	76.1306890	18.0000000	439.0000000	24.0000000	35.0000000	100.0000000	278.0000000	392.0000000
TEAM_BASERUN_CS	Caught stealing	1450	692	51.7448276	49.0000000	19.5766721	11.0000000	143.0000000	18.0000000	25.0000000	49.0000000	89.0000000	118.0000000
TEAM_PITCHING_H	Hits allowed	2142	0	7.3597576	7.3185395	0.1812562	7.0361485	8.8595055	7.1220599	7.1808312	7.3185395	7.6652847	8.1116281
TEAM_PITCHING_HR	Homeruns allowed	2142	0	106.5364146	109.5000000	58.7693290	3.0000000	244.0000000	12.0000000	19.0000000	109.5000000	206.0000000	235.0000000
TEAM_PITCHING_BB	Walks allowed	2142	0	544.5830999	536.0000000	104.8634768	144.0000000	924.0000000	319.0000000	388.0000000	536.0000000	735.0000000	834.0000000
TEAM_PITCHING_SO	Strikeouts by pitchers	2040	102	801.5946078	817.0000000	223.3229931	205.0000000	1464.00	373.0000000	437.5000000	817.0000000	1150.50	1342.00
TEAM_FIELDING_E	Errors	2142	0	5.2064601	5.0434251	0.5389037	4.1743873	7.3231707	4.4543473	4.6051702	5.0434251	6.4232470	6.7286286
TEAM_FIELDING_DP	Double Plays	1946	196	146.5739979	149.0000000	26.2151651	52.0000000	228.0000000	79.0000000	98.0000000	149.0000000	186.0000000	204.0000000

Data Preparation: Create flag variable for TEAM_BATTING_HBP and TEAM_BASERUN_CS, drop TEAM_BATTING_HBP

The TEAM_BATTING_HBP variable is missing more than 90% of its data points, so its removal from the data set would be reasonable. First, a flag variable representing availability (or unavailability) of original data for this variable is created, before the variable is removed. This flag variable will be used during model building. Since TEAM_BASERUN_CS is also missing about one-third of data points, a flag can also be created for this variable. The original variable is not deleted, since a significant number of data points are available.

Data Preparation: Impute missing values for TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_PITCHING_SO, TEAM_FIELDING_DP based on mean values

After applying the previously-mentioned data preparation steps (including removal/capping of data), missing values for these variables can be imputed from the mean values of the remaining data points. Imputing with mean values is a more reasonable option than imputing with median values, as the previous data preparation steps would likely have already removed the most extreme/unusual values.

Data Exploration: Interaction Terms

Sometimes, the value of a dependent variable may be influenced by an effect that stems from the relationship between two or more independent variables. Including interaction terms in a model can help increase predictive power. Interaction terms are created by multiplying two or more independent variables.

Data Preparation: Create interaction terms pitchingso_fieldingdp, baserunsb_baseruncs, battingh_batting2b, battingh_batting3b, battingh_battinghr, fieldinge_pitchingbb, fieldinge_pitchingh, fieldinge_pitchinghr, battingbb_battingso, baserunsb_baseruncs

For example, in baseball, the number of double plays (TEAM_FIELDING_DP) may depend on the number of strikeouts by pitchers (TEAM_PITCHING_SO), so an interaction term is created in an attempt to capture that relationship. The four interaction terms above are created from four pairs of independent variables. These interaction terms are included in some of the subsequent model building scenarios.

3. Model Building

When building models, it is important to test many combinations of selection techniques, using different variables and variable transformations, before settling on the “best” outcome. The following models test several different approaches, some with their own “optimal” model, and are sorted from largest number of model parameters to smallest number of parameters. Coincidentally the following models also tend to be ranked from highest to lowest adjusted R^2 and lowest to highest AIC values, though this may not always represent the relationship between number of model parameters and adjusted R^2 / AIC.

Note that in the following analysis analysis, “optimal” refers to a model that is superior to others, while “best” refers to the final model that is selected. Also, the magnitudes of parameter coefficients do not jump out as nonsensical, so no further discussion on that is merited.

This section of the report discusses the models built, including different combinations of original and transformed variable, as well as various variable selection techniques. The following section of the report, Model Selection, further discusses some “optimal” models and finally, how the best-fitting model is selected.

Model #1 (includes 23 parameters)

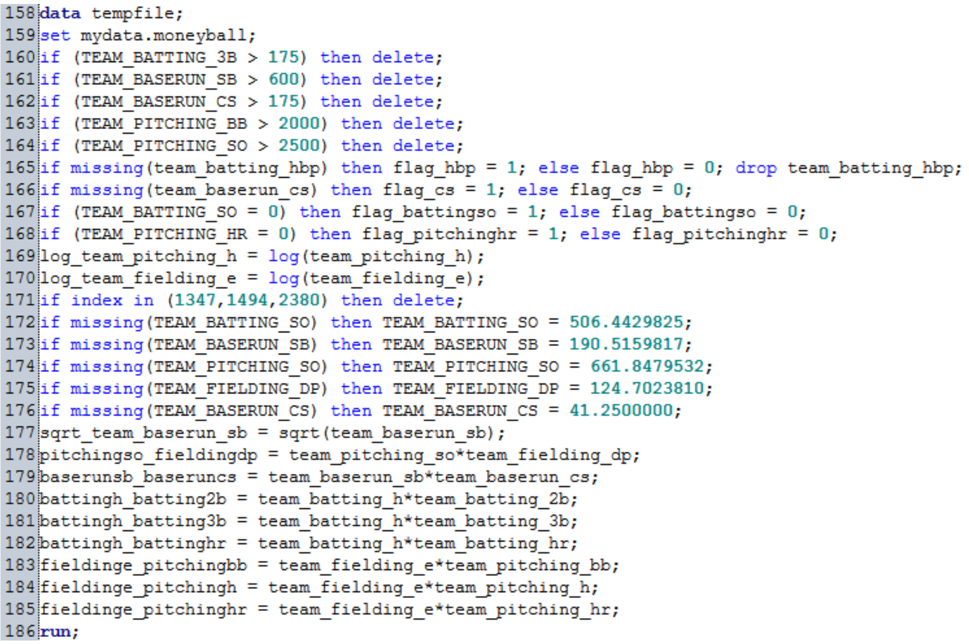
Model selection criteria: Backward
Model adjusted R^2: 0.3906
Model AIC: 11294

Data Transformations

The data transformations in this model include deleted observations, flagged variables, logistic and square root transformations, mean values as replacements for missing values, and interaction terms.

The backward variable selection method is used to generate the optimal model with the given existing and newly created variables. This approach starts with all variables in the model, and removes the least significant variables one at a time, as long as they are not significant at the default critical level, p = 0.10.

Model #1 Code:



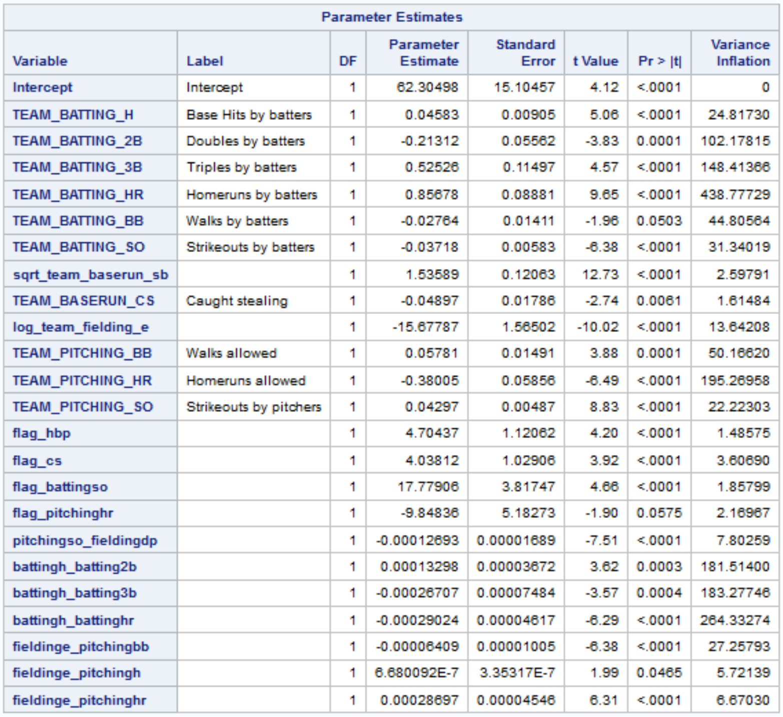
```
data tempfile;
set mydata.moneyball;
if (TEAM_BATTING_3B > 175) then delete;
if (TEAM_BASERUN_SB > 600) then delete;
if (TEAM_BASERUN_CS > 175) then delete;
if (TEAM_PITCHING_BB > 2000) then delete;
if (TEAM_PITCHING_SO > 2500) then delete;
if missing(team_batting_hbp) then flag_hbp = 1; else flag_hbp = 0; drop team_batting_hbp;
if missing(team_baserun_cs) then flag_cs = 1; else flag_cs = 0;
if (TEAM_BATTING_SO = 0) then flag_battingso = 1; else flag_battingso = 0;
if (TEAM_PITCHING_HR = 0) then flag_pitchinghr = 1; else flag_pitchinghr = 0;
log_team_pitching_h = log(team_pitching_h);
log_team_fielding_e = log(team_fielding_e);
if index in (1347,1494,2380) then delete;
if missing(TEAM_BATTING_SO) then TEAM_BATTING_SO = 506.4429825;
if missing(TEAM_BASERUN_SB) then TEAM_BASERUN_SB = 190.5159817;
if missing(TEAM_PITCHING_SO) then TEAM_PITCHING_SO = 661.8479532;
if missing(TEAM_FIELDING_DP) then TEAM_FIELDING_DP = 124.7023810;
if missing(TEAM_BASERUN_CS) then TEAM_BASERUN_CS = 41.2500000;
sqrt_team_baserun_sb = sqrt(team_baserun_sb);
pitchingso_fieldingdp = team_pitching_so*team_fielding_dp;
baserunsb_baseruncs = team_baserun_sb*team_baserun_cs;
battingh_batting2b = team_batting_h*team_batting_2b;
battingh_batting3b = team_batting_h*team_batting_3b;
battingh_battinghr = team_batting_h*team_batting_hr;
fieldinge_pitchingbb = team_fielding_e*team_pitching_bb;
fieldinge_pitchingh = team_fielding_e*team_pitching_h;
fieldinge_pitchinghr = team_fielding_e*team_pitching_hr;
run;
```

Since there are a handful of interaction terms included in this model, it would be unsurprising if some predictor variables have VIFs greater than 10, especially predictor variables from the original data set that are used in the construction of the interaction terms.

Running the backward variable selection method generates a model with some parameters whose sign are counterintuitive. For example, the sign for TEAM_BATTING_2B should be positive, since theoretically, a larger number of doubles hit by batters should result in more team wins because it means that players get on base more often. Another example where the sign of a parameter is opposite what is expected is evident with the TEAM_BATTING_BB variable. The sign for this parameter should be positive, as more walks should also result in more wins because it means that players get on base more often. Other variables whose coefficient should have an opposite sign include TEAM_PITCHING_BB and flag_pitching_hr.

Since some of these variables are (borderline) not significant at the 0.05 level, but they are left in the model anyway, as removing them to generate a 19-parameter model adversely affects both the adjusted R^2 and AIC values, which would change to 0.3697 and 11366, respectively.

Model #1 Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	62.30498	15.10457	4.12	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.04583	0.00905	5.06	<.0001	24.81730
TEAM_BATTING_2B	Doubles by batters	1	-0.21312	0.05562	-3.83	0.0001	102.17815
TEAM_BATTING_3B	Triples by batters	1	0.52526	0.11497	4.57	<.0001	148.41366
TEAM_BATTING_HR	Homeruns by batters	1	0.85678	0.08881	9.65	<.0001	438.77729
TEAM_BATTING_BB	Walks by batters	1	-0.02764	0.01411	-1.96	0.0503	44.80564
TEAM_BATTING_SO	Strikeouts by batters	1	-0.03718	0.00583	-6.38	<.0001	31.34019
sqrt_team_baserun_sb		1	1.53589	0.12063	12.73	<.0001	2.59791
TEAM_BASERUN_CS	Caught stealing	1	-0.04897	0.01786	-2.74	0.0061	1.61484
log_team_fielding_e		1	-15.67787	1.56502	-10.02	<.0001	13.64208
TEAM_PITCHING_BB	Walks allowed	1	0.05781	0.01491	3.88	0.0001	50.16620
TEAM_PITCHING_HR	Homeruns allowed	1	-0.38005	0.05856	-6.49	<.0001	195.26958
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.04297	0.00487	8.83	<.0001	22.22303
flag_hbp		1	4.70437	1.12062	4.20	<.0001	1.48575
flag_cs		1	4.03812	1.02906	3.92	<.0001	3.60690
flag_battingso		1	17.77906	3.81747	4.66	<.0001	1.85799
flag_pitchinghr		1	-9.84836	5.18273	-1.90	0.0575	2.16967
pitchingso_fieldingdp		1	-0.00012693	0.00001689	-7.51	<.0001	7.80259
battingh_batting2b		1	0.00013298	0.00003672	3.62	0.0003	181.51400
battingh_batting3b		1	-0.00026707	0.00007484	-3.57	0.0004	183.27746
battingh_battinghr		1	-0.00029024	0.00004617	-6.29	<.0001	264.33274
fieldinge_pitchingbb		1	-0.00006409	0.00001005	-6.38	<.0001	27.25793
fieldinge_pitchingh		1	6.680092E-7	3.35317E-7	1.99	0.0465	5.72139
fieldinge_pitchinghr		1	0.00028697	0.00004546	6.31	<.0001	6.67030

Model #2 (includes 21 parameters)

Model selection criteria: Backward

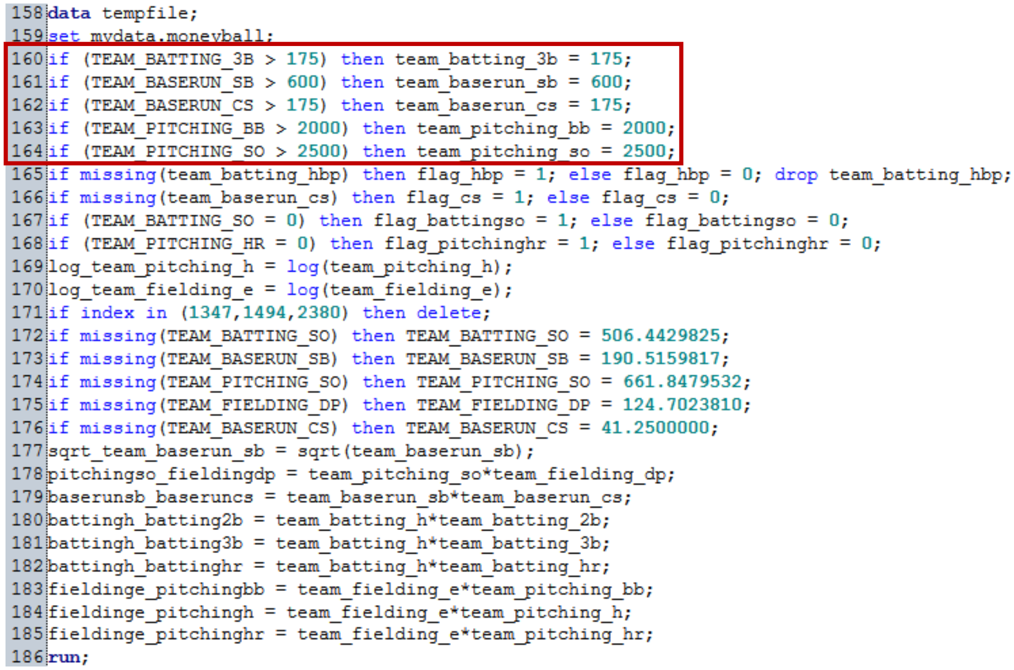
Model adjusted R^2: 0.3845

Model AIC: 11419

Data Transformations

The data transformations in this model include deleted observations, capped variables, flagged variables, logistic and square root transformations, mean values as replacements for missing values, and interaction terms. The only difference between this set of data transformations and the first set of data transformations is capped vs. deleted outliers, as highlighted in red in the code below.

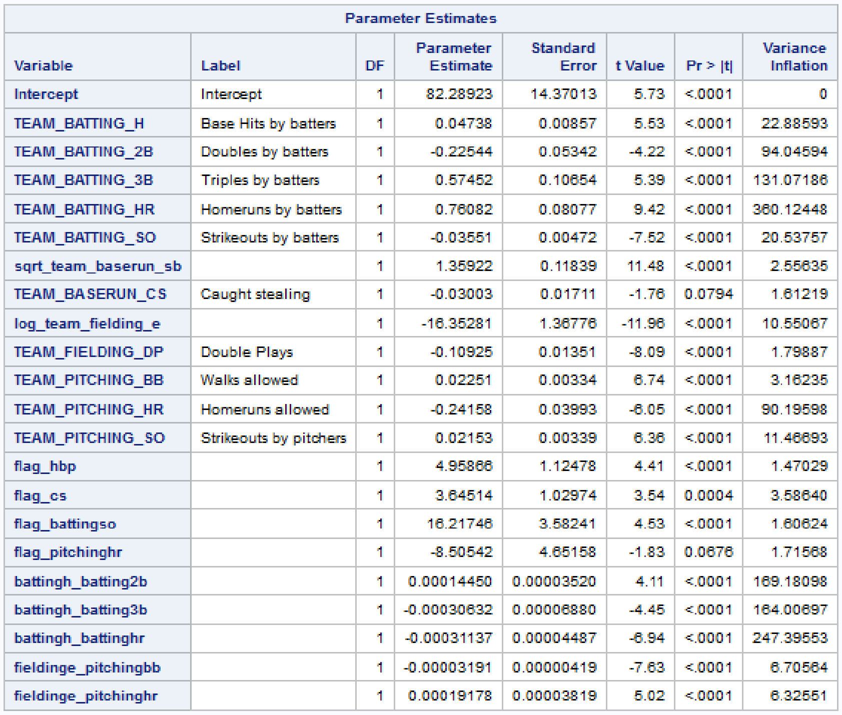
Model #2 code:



```
data tempfile;
set mydata.moneyball;
if (TEAM_BATTING_3B > 175) then team_batting_3b = 175;
if (TEAM_BASERUN_SB > 600) then team_baserun_sb = 600;
if (TEAM_BASERUN_CS > 175) then team_baserun_cs = 175;
if (TEAM_PITCHING_BB > 2000) then team_pitching_bb = 2000;
if (TEAM_PITCHING_SO > 2500) then team_pitching_so = 2500;
if missing(team_batting_hbp) then flag_hbp = 1; else flag_hbp = 0; drop team_batting_hbp;
if missing(team_baserun_cs) then flag_cs = 1; else flag_cs = 0;
if (TEAM_BATTING_SO = 0) then flag_battingso = 1; else flag_battingso = 0;
if (TEAM_PITCHING_HR = 0) then flag_pitchinghr = 1; else flag_pitchinghr = 0;
log_team_pitching_h = log(team_pitching_h);
log_team_fielding_e = log(team_fielding_e);
if index in (1347,1494,2380) then delete;
if missing(TEAM_BATTING_SO) then TEAM_BATTING_SO = 506.4429825;
if missing(TEAM_BASERUN_SB) then TEAM_BASERUN_SB = 190.5159817;
if missing(TEAM_PITCHING_SO) then TEAM_PITCHING_SO = 661.8479532;
if missing(TEAM_FIELDING_DP) then TEAM_FIELDING_DP = 124.7023810;
if missing(TEAM_BASERUN_CS) then TEAM_BASERUN_CS = 41.2500000;
sqrt_team_baserun_sb = sqrt(team_baserun_sb);
pitchingso_fieldingdp = team_pitching_so*team_fielding_dp;
baserunsb_baseruncs = team_baserun_sb*team_baserun_cs;
battingh_batting2b = team_batting_h*team_batting_2b;
battingh_batting3b = team_batting_h*team_batting_3b;
battingh_battinghr = team_batting_h*team_batting_hr;
fieldinge_pitchingbb = team_fielding_e*team_pitching_bb;
fieldinge_pitchingh = team_fielding_e*team_pitching_h;
fieldinge_pitchinghr = team_fielding_e*team_pitching_hr;
run;
```

In the previous scenario, observations are deleted if certain variables had extreme outliers. In this scenario, values are capped for certain variables. The same backward variable selection method is then applied. This scenario results in a 21-parameter optimal model. Like in the 23-parameter model, the 21-parameter model still has some counterintuitive signs in its parameters. Multicollinearity is also still present, though to a lesser extent, as evident from the VIF in the very right of the *Parameter Estimates* table below.

Model #2 Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	82.28923	14.37013	5.73	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.04738	0.00857	5.53	<.0001	22.88593
TEAM_BATTING_2B	Doubles by batters	1	-0.22544	0.05342	-4.22	<.0001	94.04594
TEAM_BATTING_3B	Triples by batters	1	0.57452	0.10654	5.39	<.0001	131.07186
TEAM_BATTING_HR	Homeruns by batters	1	0.76082	0.08077	9.42	<.0001	360.12448
TEAM_BATTING_SO	Strikeouts by batters	1	-0.03551	0.00472	-7.52	<.0001	20.53757
sqrt_team_baserun_sb		1	1.35922	0.11839	11.48	<.0001	2.55635
TEAM_BASERUN_CS	Caught stealing	1	-0.03003	0.01711	-1.76	0.0794	1.61219
log_team_fielding_e		1	-16.35281	1.36776	-11.96	<.0001	10.55067
TEAM_FIELDING_DP	Double Plays	1	-0.10925	0.01351	-8.09	<.0001	1.79887
TEAM_PITCHING_BB	Walks allowed	1	0.02251	0.00334	6.74	<.0001	3.16235
TEAM_PITCHING_HR	Homeruns allowed	1	-0.24158	0.03993	-6.05	<.0001	90.19598
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.02153	0.00339	6.36	<.0001	11.46693
flag_hbp		1	4.95866	1.12478	4.41	<.0001	1.47029
flag_cs		1	3.64514	1.02974	3.54	0.0004	3.58640
flag_battingso		1	16.21746	3.58241	4.53	<.0001	1.60624
flag_pitchinghr		1	-8.50542	4.65158	-1.83	0.0676	1.71568
battingh_batting2b		1	0.00014450	0.00003520	4.11	<.0001	169.18098
battingh_batting3b		1	-0.00030632	0.00006880	-4.45	<.0001	164.00697
battingh_battinghr		1	-0.00031137	0.00004487	-6.94	<.0001	247.39553
fieldinge_pitchingbb		1	-0.00003191	0.00000419	-7.63	<.0001	6.70564
fieldinge_pitchinghr		1	0.00019178	0.00003819	5.02	<.0001	6.32551

Model #3 (including 20 parameters)

Model selection criteria: Backward, with slstay = 0.05

Model adjusted R^2: 0.3891

Model AIC: 11297

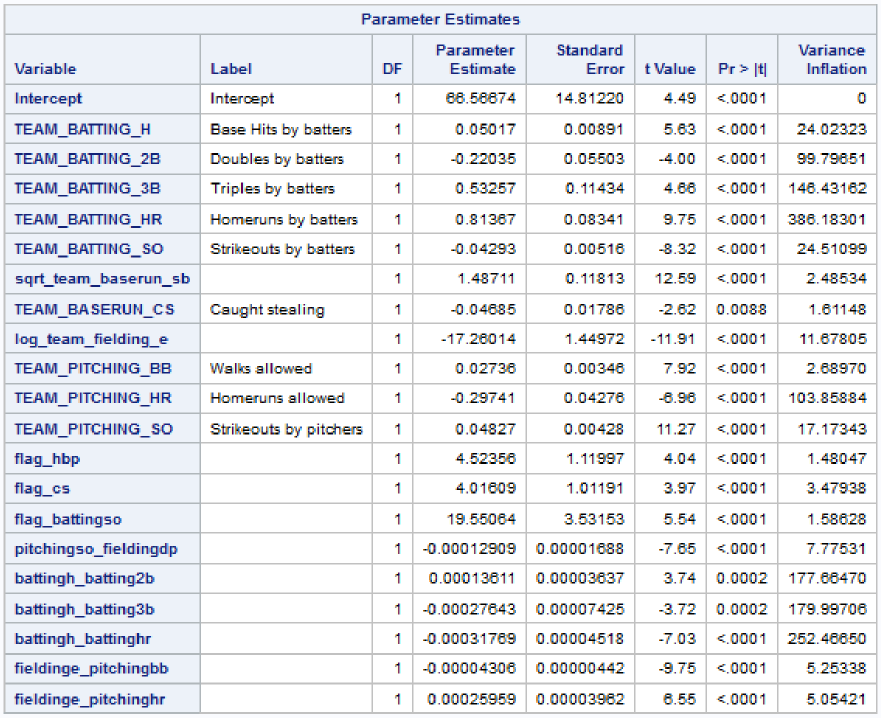
Data Transformations

The data transformations in this model mimic those applied in Model #1 above.

Of the previous two models, since the first one with deleted observations performs better than the second one with capped values, this model further tests the first model, using backward variable selection and a slstay value of 0.05 (vs. the 0.10 default value).

Though fewer variables (20) make it into the model, the adjusted R^2 is higher and the AIC is lower than the model with 21 parameters, indicating that it may be a comparatively superior model. There are only two parameters in which the signs are counterintuitive, but for reasons discussed above, these will both be left in the model.

Model #3 Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	66.56674	14.81220	4.49	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.05017	0.00891	5.63	<.0001	24.02323
TEAM_BATTING_2B	Doubles by batters	1	-0.22035	0.05503	-4.00	<.0001	99.79651
TEAM_BATTING_3B	Triples by batters	1	0.53257	0.11434	4.66	<.0001	146.43162
TEAM_BATTING_HR	Homeruns by batters	1	0.81367	0.08341	9.75	<.0001	386.18301
TEAM_BATTING_SO	Strikeouts by batters	1	-0.04293	0.00516	-8.32	<.0001	24.51099
sqrt_team_baserun_sb		1	1.48711	0.11813	12.59	<.0001	2.48534
TEAM_BASERUN_CS	Caught stealing	1	-0.04685	0.01786	-2.62	0.0088	1.61148
log_team_fielding_e		1	-17.26014	1.44972	-11.91	<.0001	11.67805
TEAM_PITCHING_BB	Walks allowed	1	0.02736	0.00346	7.92	<.0001	2.68970
TEAM_PITCHING_HR	Homeruns allowed	1	-0.29741	0.04276	-6.96	<.0001	103.85884
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.04827	0.00428	11.27	<.0001	17.17343
flag_hbp		1	4.52356	1.11997	4.04	<.0001	1.48047
flag_cs		1	4.01609	1.01191	3.97	<.0001	3.47938
flag_battingso		1	19.55064	3.53153	5.54	<.0001	1.58628
pitchingso_fieldingdp		1	-0.00012909	0.00001688	-7.65	<.0001	7.77531
battingh_batting2b		1	0.00013611	0.00003637	3.74	0.0002	177.66470
battingh_batting3b		1	-0.00027643	0.00007425	-3.72	0.0002	179.99706
battingh_battinghr		1	-0.00031769	0.00004518	-7.03	<.0001	252.46650
fieldinge_pitchingbb		1	-0.00004306	0.00000442	-9.75	<.0001	5.25338
fieldinge_pitchinghr		1	0.00025959	0.00003962	6.55	<.0001	5.05421

Model #4 (including 20 parameters)

Model selection criteria: Adjrsq

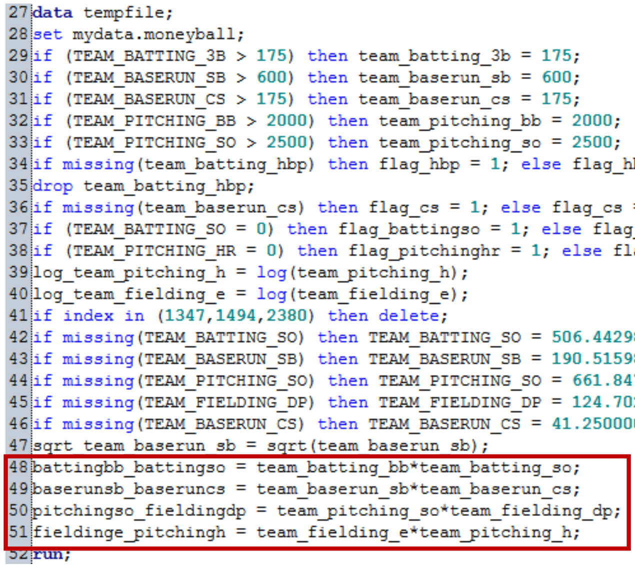
Model adjusted R^2: 0.3702

Model AIC: 11470

Data Transformations

The data transformations in this model include deleted observations, capped variables, flagged variables, logistic and square root transformations, mean values as replacements for missing values, and interaction terms. Compared to the previous models, this model includes a different set of interaction terms, as highlighted in red in the code below.

Model #4 Code:

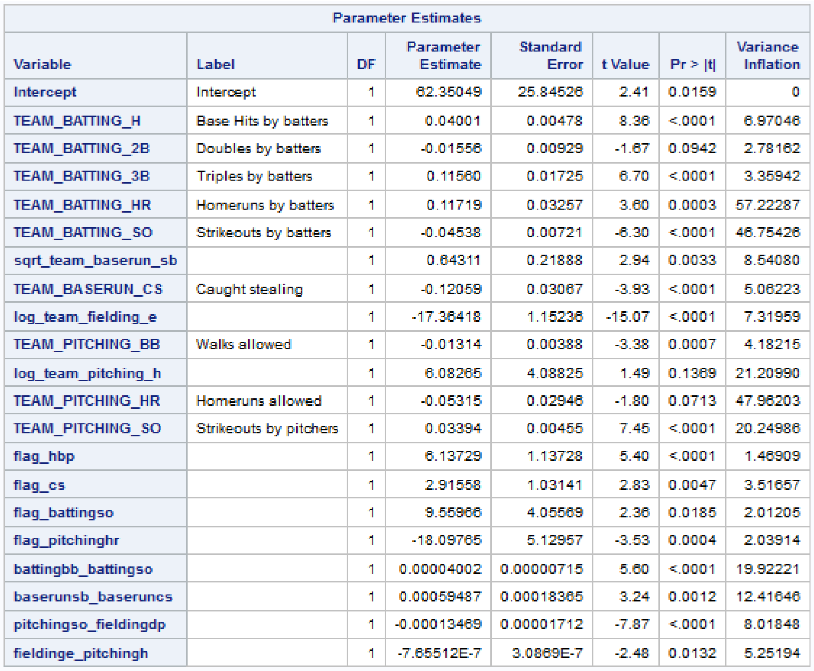


```
data tempfile;
set mydata.moneyball;
if (TEAM_BATTING_3B > 175) then team_batting_3b = 175;
if (TEAM_BASERUN_SB > 600) then team_baserun_sb = 600;
if (TEAM_BASERUN_CS > 175) then team_baserun_cs = 175;
if (TEAM_PITCHING_BB > 2000) then team_pitching_bb = 2000;
if (TEAM_PITCHING_SO > 2500) then team_pitching_so = 2500;
if missing(team_batting_hbp) then flag_hbp = 1; else flag_hbp = 0;
drop team_batting_hbp;
if missing(team_baserun_cs) then flag_cs = 1; else flag_cs = 0;
if (TEAM_BATTING_SO = 0) then flag_battingso = 1; else flag_battingso = 0;
if (TEAM_PITCHING_HR = 0) then flag_pitchinghr = 1; else flag_pitchinghr = 0;
log_team_pitching_h = log(team_pitching_h);
log_team_fielding_e = log(team_fielding_e);
if index in (1347,1494,2380) then delete;
if missing(TEAM_BATTING_SO) then TEAM_BATTING_SO = 506.4429825;
if missing(TEAM_BASERUN_SB) then TEAM_BASERUN_SB = 190.5159817;
if missing(TEAM_PITCHING_SO) then TEAM_PITCHING_SO = 661.8479532;
if missing(TEAM_FIELDING_DP) then TEAM_FIELDING_DP = 124.7023810;
if missing(TEAM_BASERUN_CS) then TEAM_BASERUN_CS = 41.2500000;
sqrt_team_baserun_sb = sqrt(team_baserun_sb);
battingbb_battingso = team_batting_bb*team_batting_so;
baserunsb_baseruncs = team_baserun_sb*team_baserun_cs;
pitchingso_fieldingdp = team_pitching_so*team_fielding_dp;
fieldinge_pitchingh = team_fielding_e*team_pitching_h;
run;
```

The adjrsq selection criteria, which automatically ranks the models in order from largest to smallest adjusted R^2 value, is applied to generate this scenario's optimal model. The optimal regression equation generated from the adjrsq selection criteria has 20 predictor variables, including four flag variables (as indicated by the prefix "flag") and four interaction terms.

Per the "Variance Inflation" column in the *Parameter Estimates* table below, there appears to be multicollinearity, indicating that two or more of the predictor variables are highly correlated. Some parameter estimates have VIFs greater than 10. This may partially be attributable to the presence of interaction terms.

Model #4 Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	62.35049	25.84526	2.41	0.0159	0
TEAM_BATTING_H	Base Hits by batters	1	0.04001	0.00478	8.36	<.0001	6.97046
TEAM_BATTING_2B	Doubles by batters	1	-0.01556	0.00929	-1.67	0.0942	2.78162
TEAM_BATTING_3B	Triples by batters	1	0.11560	0.01725	6.70	<.0001	3.35942
TEAM_BATTING_HR	Homeruns by batters	1	0.11719	0.03257	3.60	0.0003	57.22287
TEAM_BATTING_SO	Strikeouts by batters	1	-0.04538	0.00721	-6.30	<.0001	46.75426
sqrt_team_baserun_sb		1	0.64311	0.21888	2.94	0.0033	8.54080
TEAM_BASERUN_CS	Caught stealing	1	-0.12059	0.03067	-3.93	<.0001	5.06223
log_team_fielding_e		1	-17.36418	1.15236	-15.07	<.0001	7.31959
TEAM_PITCHING_BB	Walks allowed	1	-0.01314	0.00388	-3.38	0.0007	4.18215
log_team_pitching_h		1	6.08265	4.08825	1.49	0.1369	21.20990
TEAM_PITCHING_HR	Homeruns allowed	1	-0.05315	0.02946	-1.80	0.0713	47.96203
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.03394	0.00455	7.45	<.0001	20.24986
flag_hbp		1	6.13729	1.13728	5.40	<.0001	1.46909
flag_cs		1	2.91558	1.03141	2.83	0.0047	3.51657
flag_battingso		1	9.55966	4.05569	2.36	0.0185	2.01205
flag_pitchinghr		1	-18.09765	5.12957	-3.53	0.0004	2.03914
battingbb_battingso		1	0.00004002	0.00000715	5.60	<.0001	19.92221
baserunsb_baseruncs		1	0.00059487	0.00018365	3.24	0.0012	12.41646
pitchingso_fieldingdp		1	-0.00013469	0.00001712	-7.87	<.0001	8.01848
fieldinge_pitchingh		1	-7.65512E-7	3.0869E-7	-2.48	0.0132	5.25194

While the direction of most parameter estimates seems intuitive, there are a few variables for which the sign does not make sense. These parameters include: TEAM_BATTING_2B, log_team_pitching_h, and flag_pitchinghr. I would expected these parameters to have the signs opposite what is indicated in the *Parameter Estimates* table. For example, doubles by batters (TEAM_BATTING_2B) should have a positive impact on the number of team wins.

The TEAM_BATTING_2B and log_team_pitching_h variables aren't statistically significant at the 0.05 level. , but excluding them from the model would result in decreased adjusted R^2 and increased AIC/SBC, so they should remain in the model.

Model #5 (including 18 parameters)

Model selection criteria: Stepwise

Model adjusted R^2: 0.3696

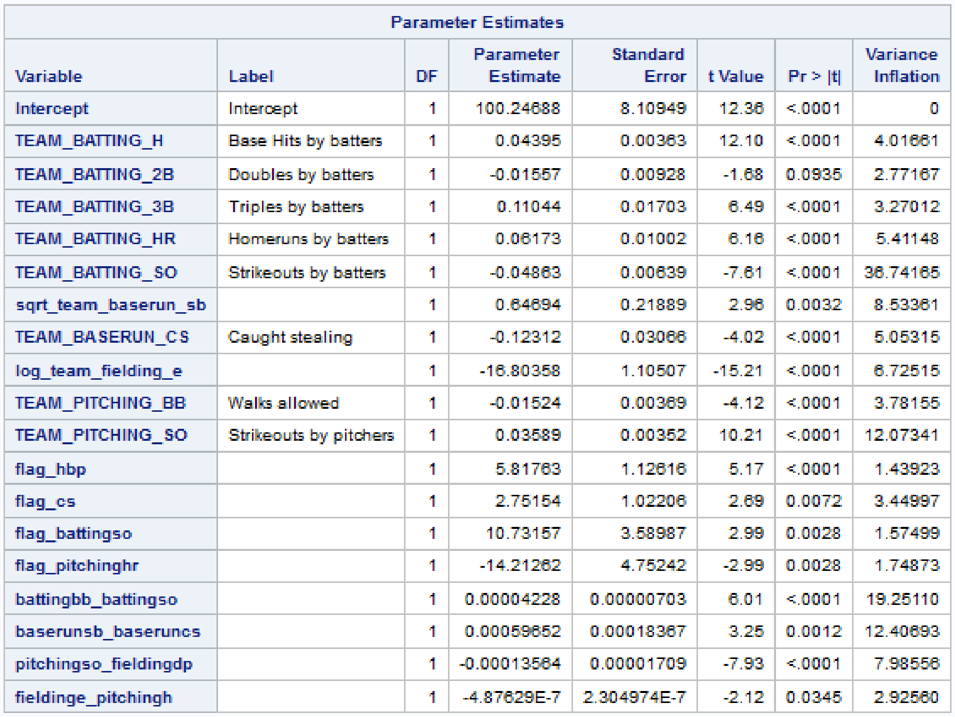
Model AIC: 11470

Data Transformations

The data transformations in this model mimic those applied in Model #4 above.

The optimal regression equation generated from the stepwise selection criteria yields 18 parameters. Compared to the previous model, this model excludes the log_team_pitching_h and TEAM_PITCHING_HR variables. The four flag variables and four interaction terms remain in the model.

Model #5 Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	100.24688	8.10949	12.36	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.04395	0.00363	12.10	<.0001	4.01661
TEAM_BATTING_2B	Doubles by batters	1	-0.01557	0.00928	-1.68	0.0935	2.77167
TEAM_BATTING_3B	Triples by batters	1	0.11044	0.01703	6.49	<.0001	3.27012
TEAM_BATTING_HR	Homeruns by batters	1	0.06173	0.01002	6.16	<.0001	5.41148
TEAM_BATTING_SO	Strikeouts by batters	1	-0.04863	0.00639	-7.61	<.0001	36.74165
sqrt_team_baserun_sb		1	0.64694	0.21889	2.96	0.0032	8.53361
TEAM_BASERUN_CS	Caught stealing	1	-0.12312	0.03066	-4.02	<.0001	5.05315
log_team_fielding_e		1	-16.80358	1.10507	-15.21	<.0001	6.72515
TEAM_PITCHING_BB	Walks allowed	1	-0.01524	0.00369	-4.12	<.0001	3.78155
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.03589	0.00352	10.21	<.0001	12.07341
flag_hbp		1	5.81763	1.12616	5.17	<.0001	1.43923
flag_cs		1	2.75154	1.02206	2.69	0.0072	3.44997
flag_battingso		1	10.73157	3.58987	2.99	0.0028	1.57499
flag_pitchinghr		1	-14.21262	4.75242	-2.99	0.0028	1.74873
battingbb_battingso		1	0.00004228	0.00000703	6.01	<.0001	19.25110
baserunsb_baseruncs		1	0.00059652	0.00018367	3.25	0.0012	12.40693
pitchingso_fieldingdp		1	-0.00013564	0.00001709	-7.93	<.0001	7.98556
fieldinge_pitchingh		1	-4.87629E-7	2.304974E-7	-2.12	0.0345	2.92560

There still appears to be multicollinearity in this model, though to a lesser extent than in the previous model (created using the adjrsq variable selection method). Only four variables have a VIF greater than 10 in this scenario, with one whose VIF value significantly higher than 10.

In this model, the sign of the TEAM_BATTING_2B parameter doesn't make intuitive sense. Since the parameter is also not statistically significant, it can be optionally removed out of the model. Another variable with a counterintuitive sign is flag_pitchinghr. Since this variable is coded with a value of "1" for any values of zero in the original variable and a value of "0" for any non-zero values in the original variable, this variable is expected to have a positive influence, not a negative influence.

Model #6 (including 14 parameters)

Model selection criteria: Adjrsq/stepwise, considering only the original/transformed 15 variables

Model adjusted R^2: 0.3550

Model AIC: 11519

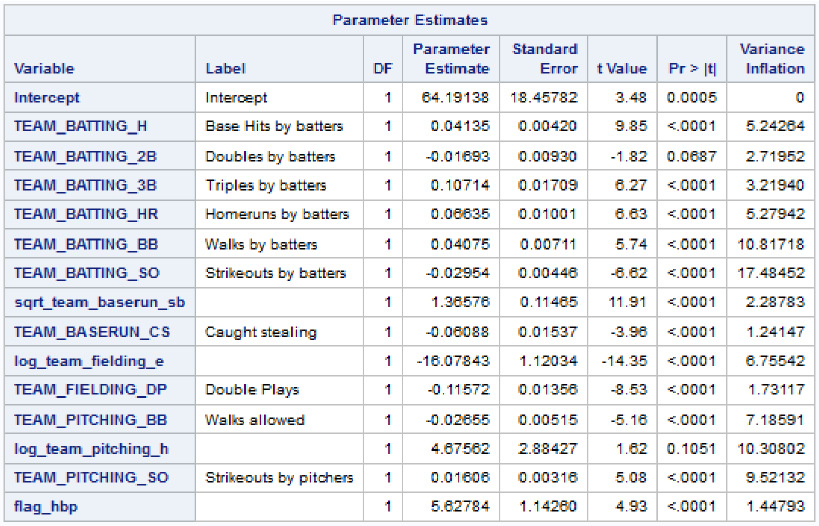
Data Transformations

The data transformations in this model also mimic those applied in Model #4 above.

The variables considered for inclusion in this scenario take on some form of one of the original 15 variables. That is, the variables take on either the form of the original variable or a transformed version of the original variable (log, square root, flag). No additional flag variables or interaction variables are added.

The optimal regression equation generated from applying the adjrsq selection criteria contains 14 parameters. Of the variables for consideration, the only one that does not make it into this model is TEAM_PITCHING_HR. Comparatively, this model contains fewer parameters, so it is in a sense, more parsimonious and easy to understand, but it also has a lower adjusted R^2, indicating that it also has lower predictive power.

Model #6 Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	64.19138	18.45782	3.48	0.0005	0
TEAM_BATTING_H	Base Hits by batters	1	0.04135	0.00420	9.85	<.0001	5.24264
TEAM_BATTING_2B	Doubles by batters	1	-0.01693	0.00930	-1.82	0.0687	2.71952
TEAM_BATTING_3B	Triples by batters	1	0.10714	0.01709	6.27	<.0001	3.21940
TEAM_BATTING_HR	Homeruns by batters	1	0.06635	0.01001	6.63	<.0001	5.27942
TEAM_BATTING_BB	Walks by batters	1	0.04075	0.00711	5.74	<.0001	10.81718
TEAM_BATTING_SO	Strikeouts by batters	1	-0.02954	0.00446	-6.62	<.0001	17.48452
sqrt_team_baserun_sb		1	1.36576	0.11465	11.91	<.0001	2.28783
TEAM_BASERUN_CS	Caught stealing	1	-0.06088	0.01537	-3.96	<.0001	1.24147
log_team_fielding_e		1	-16.07843	1.12034	-14.35	<.0001	6.75542
TEAM_FIELDING_DP	Double Plays	1	-0.11572	0.01356	-8.53	<.0001	1.73117
TEAM_PITCHING_BB	Walks allowed	1	-0.02655	0.00515	-5.16	<.0001	7.18591
log_team_pitching_h		1	4.67562	2.88427	1.62	0.1051	10.30802
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.01606	0.00316	5.08	<.0001	9.52132
flag_hbp		1	5.62784	1.14260	4.93	<.0001	1.44793

This model also possesses multicollinearity, though to a lesser extent than some of the previous models. There are three variable with VIF greater than 10, and only one variable has a VIF significantly greater than 10.

Again, the signs for some of the parameters don't make intuitive sense. The signs for TEAM_BATTING_2B and TEAM_FIELDING_DP in particular, are expected to be positive, and the sign for log_team_pitching_h is expected to be negative. Two of those variables appear don't appear to be significant at the 0.05 level. The model generated after removal of the two non-significant variables is discussed next.

Model #7 (including 12 parameters)

Model selection criteria: Adjrsq, removed two variables from 14-parameter model

Model adjusted R^2: 0.3335

Model AIC: 11591

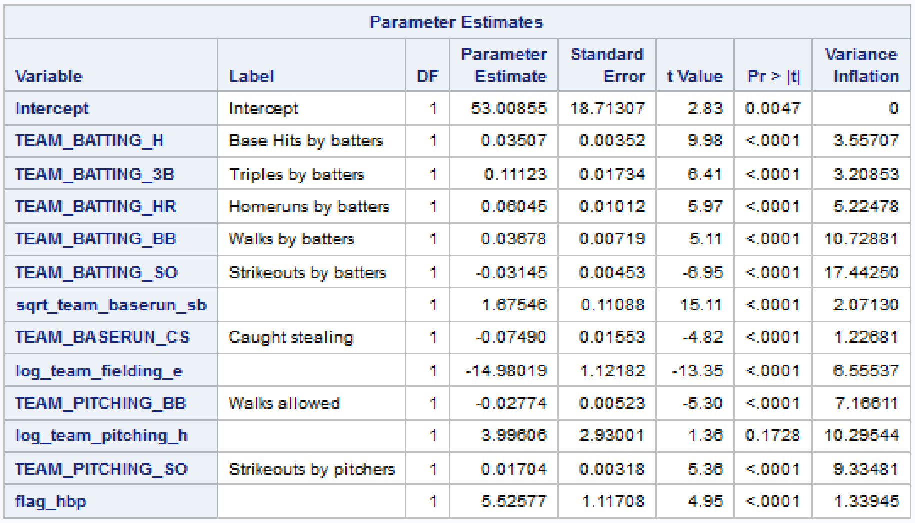
Data Transformations

The data transformations in this model also mimic those applied in Model #4 above.

Additionally, two variables that are not significant at the 0.05 level are manually removed from this model. Removing the TEAM_BATTING_2B and TEAM_FIELIDNG_DP variables from the 14-parameter model above and applying the adjrsq selection criteria results in a model with a lower adjusted R^2 and higher AIC, both less desirable qualities in terms of model fit. Also, in this model, there is not much difference (decrease from the previous model) in the multicollinearity present, as there are still three variables with a VIF greater than 10.

This 12-parameter model is comparatively worse than the 14-parameter model.

Model #7 Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	53.00855	18.71307	2.83	0.0047	0
TEAM_BATTING_H	Base Hits by batters	1	0.03507	0.00352	9.98	<.0001	3.55707
TEAM_BATTING_3B	Triples by batters	1	0.11123	0.01734	6.41	<.0001	3.20853
TEAM_BATTING_HR	Homeruns by batters	1	0.06045	0.01012	5.97	<.0001	5.22478
TEAM_BATTING_BB	Walks by batters	1	0.03678	0.00719	5.11	<.0001	10.72881
TEAM_BATTING_SO	Strikeouts by batters	1	-0.03145	0.00453	-6.95	<.0001	17.44250
sqrt_team_baserun_sb		1	1.67546	0.11088	15.11	<.0001	2.07130
TEAM_BASERUN_CS	Caught stealing	1	-0.07490	0.01553	-4.82	<.0001	1.22681
log_team_fielding_e		1	-14.98019	1.12182	-13.35	<.0001	6.55537
TEAM_PITCHING_BB	Walks allowed	1	-0.02774	0.00523	-5.30	<.0001	7.16611
log_team_pitching_h		1	3.99606	2.93001	1.36	0.1728	10.29544
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.01704	0.00318	5.36	<.0001	9.33481
flag_hbp		1	5.52577	1.11708	4.95	<.0001	1.33945

4. Model Selection

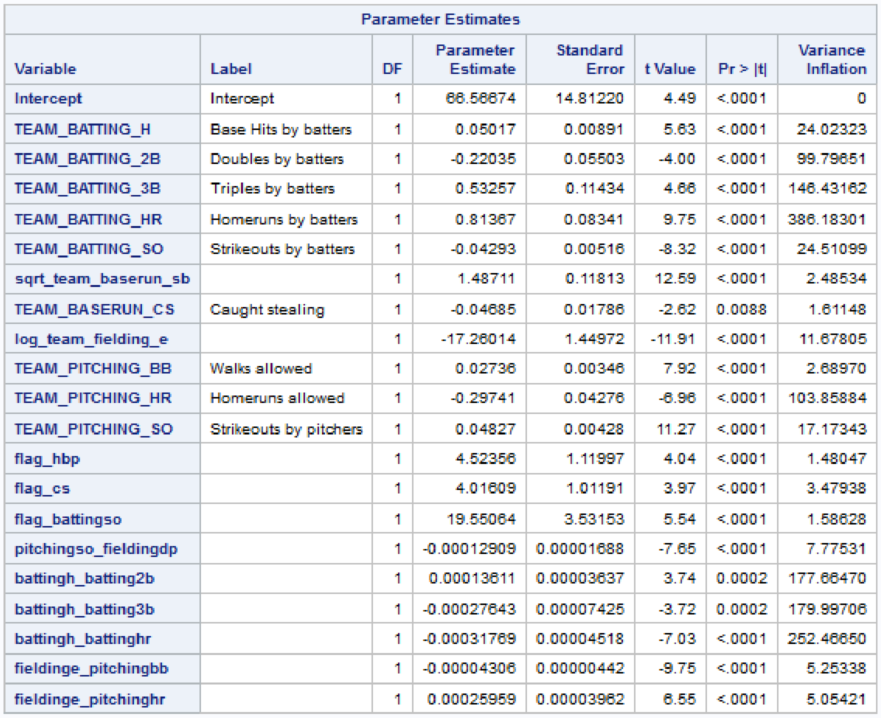
When comparing and selecting models, the adjusted R^2 metric can be compared across models to determine which has more predictive power. Models with higher adjusted R^2 values generally have greater predictive power. Compared to the standard R^2 metric, *adjusted* R^2 penalizes models with a larger number of predictor variables, and thus can be useful to compare model fit across models with different numbers of parameters.

Alternative metrics for model comparison and selection are AIC/SBC/BIC, which tend to penalize for over-fitting and model complexity. Lower values of AIC/SBC/BIC are preferable and tend to correspond to models that are more parsimonious. Parsimonious models capture a certain level of prediction using as few predictor variables as possible.

My preference of adoption leans more toward models with more predictive power rather than models with more parsimony. As magnitude of predictive power is measured by the adjusted R^2 statistic, I would select a model with a relatively higher adjusted R^2. That said, if the parsimony-predictive power tradeoff (i.e. parsimony gained for predictive power given up) is reasonable, I would be willing to consider a model with lower adjusted R^2.

Based on the trials I've run, as shown in the prior Model Building section of this report, I would elect to proceed with Model #3 that includes 20 parameters. Compared to Model #1 with 23 parameters, the adjusted R^2 is lower by only 0.0015 and the AIC is higher only by 3 points. These differences are relatively minimal. The tradeoff is that Model #3 has 3 fewer variables than Model #1, making it a more parsimonious model.

All the variables in Model #3 are statistically significant at a 0.05 level, indicating that they should remain in the model, despite the occasional unintuitive signs of the parameter estimates (see mention above). Model #3 is thus my "best" model.



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	66.56674	14.81220	4.49	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.05017	0.00891	5.63	<.0001	24.02323
TEAM_BATTING_2B	Doubles by batters	1	-0.22035	0.05503	-4.00	<.0001	99.79651
TEAM_BATTING_3B	Triples by batters	1	0.53257	0.11434	4.66	<.0001	146.43162
TEAM_BATTING_HR	Homeruns by batters	1	0.81367	0.08341	9.75	<.0001	386.18301
TEAM_BATTING_SO	Strikeouts by batters	1	-0.04293	0.00516	-8.32	<.0001	24.51099
sqrt_team_baserun_sb		1	1.48711	0.11813	12.59	<.0001	2.48534
TEAM_BASERUN_CS	Caught stealing	1	-0.04685	0.01786	-2.62	0.0088	1.61148
log_team_fielding_e		1	-17.26014	1.44972	-11.91	<.0001	11.67805
TEAM_PITCHING_BB	Walks allowed	1	0.02736	0.00346	7.92	<.0001	2.68970
TEAM_PITCHING_HR	Homeruns allowed	1	-0.29741	0.04276	-6.96	<.0001	103.85884
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.04827	0.00428	11.27	<.0001	17.17343
flag_hbp		1	4.52356	1.11997	4.04	<.0001	1.48047
flag_cs		1	4.01609	1.01191	3.97	<.0001	3.47938
flag_battingso		1	19.55064	3.53153	5.54	<.0001	1.58628
pitchingso_fieldingdp		1	-0.00012909	0.00001688	-7.65	<.0001	7.77531
battingh_batting2b		1	0.00013611	0.00003637	3.74	0.0002	177.66470
battingh_batting3b		1	-0.00027643	0.00007425	-3.72	0.0002	179.99706
battingh_battinghr		1	-0.00031769	0.00004518	-7.03	<.0001	252.46650
fieldinge_pitchingbb		1	-0.00004306	0.00000442	-9.75	<.0001	5.25338
fieldinge_pitchinghr		1	0.00025959	0.00003962	6.55	<.0001	5.05421

Conclusion

Trial and error data exploration/transformation and model building are undertaken to generate an optimal model for predicting the total number of wins for a given baseball team across a set of performance data.

The data exploration phase includes learning about the Moneyball data set in terms of its size, number and types of variables present, distribution/shape, missing values, and correlations between variables. Data exploration facilitates a better understanding of how to transform the data in preparation for model building.

Data preparation includes:

- Fixing missing values using mean, median, and other values
- Deleting/capping outlier values
- Creating flags for missing values and/or “0” values
- Mathematical transformations using (natural) log and square root
- Creating interaction terms between pairs of predictor variables

The model building phase follows data preparation. Different variable selection criteria, such as backward, stepwise, and adjrsq are trialed. Also, different numbers of predictor variables are selected for inclusion in certain model building scenarios.

The “best” model as a result of this exercise is one derived from the backward variable selection method, with a 0.05 level of statistical significance for variables that remain in the model. This model has 20 parameters, including a several flag variables and interaction terms. This model had a relatively high adjusted R^2 and relatively low AIC, indicating a combination of predictive power and parsimony.

Stand Alone Scoring Program

The following SAS code can be used to score new data and predict the number of wins. All variable transformations appear at the top of the code and the regression formula that can be used to score the new data appears at the bottom of the code.

```
libname mydata '/home/joycepang2007/my_courses/donald.wedding/c_8888/PRED411/UNIT01/HW/'
access=readonly;

data SCOREFILE;
set mydata.moneyball_test;
label index='Joyce Pang';
if missing(team_batting_hbp) then flag_hbp = 1; else flag_hbp = 0; drop team_batting_hbp;
if missing(team_baserun_cs) then flag_cs = 1; else flag_cs = 0;
if (TEAM_BATTING_SO = 0) then flag_battingso = 1; else flag_battingso = 0;
if (TEAM_PITCHING_HR = 0) then flag_pitchinghr = 1; else flag_pitchinghr = 0;
log_team_pitching_h = log(team_pitching_h);
log_team_fielding_e = log(team_fielding_e);
if missing(TEAM_BATTING_SO) then TEAM_BATTING_SO = 506.4429825;
if missing(TEAM_BASERUN_SB) then TEAM_BASERUN_SB = 190.5159817;
if missing(TEAM_PITCHING_SO) then TEAM_PITCHING_SO = 661.8479532;
if missing(TEAM_FIELDING_DP) then TEAM_FIELDING_DP = 124.7023810;
if missing(TEAM_BASERUN_CS) then TEAM_BASERUN_CS = 41.2500000;
```

```
sqrt_team_baserun_sb = sqrt(team_baserun_sb);
pitchingso_fieldingdp = team_pitching_so*team_fielding_dp;
baserunsb_baseruncs = team_baserun_sb*team_baserun_cs;
battingh_batting2b = team_batting_h*team_batting_2b;
battingh_batting3b = team_batting_h*team_batting_3b;
battingh_battinghr = team_batting_h*team_batting_hr;
fieldinge_pitchingbb = team_fielding_e*team_pitching_bb;
fieldinge_pitchingh = team_fielding_e*team_pitching_h;
fieldinge_pitchinghr = team_fielding_e*team_pitching_hr;
P_TARGET_WINS =
         66.56674
+        0.05017        *       TEAM_BATTING_H
+        -0.22035       *       TEAM_BATTING_2B
+        0.53257        *       TEAM_BATTING_3B
+        0.81367        *       TEAM_BATTING_HR
+        -0.04293       *       TEAM_BATTING_SO
+        1.48711        *       sqrt_team_baserun_sb
+        -0.04685       *       TEAM_BASERUN_CS
+        -17.26014      *       log_team_fielding_e
+        0.02736        *       TEAM_PITCHING_BB
+        -0.29741       *       TEAM_PITCHING_HR
+        0.04827        *       TEAM_PITCHING_SO
+        4.52356        *       flag_hbp
+        4.01609        *       flag_cs
+        19.55064       *       flag_battingso
+        -0.00012909    *       pitchingso_fieldingdp
+        0.00013611     *       battingh_batting2b
+        -0.00027643    *       battingh_batting3b
+        -0.00031769    *       battingh_battinghr
+        -0.00004306    *       fieldinge_pitchingbb
+        0.00025959     *       fieldinge_pitchinghr;
if P_TARGET_WINS < 38 then P_TARGET_WINS = 38;
if P_TARGET_WINS > 114 then P_TARGET_WINS = 114;
keep INDEX;
keep P_TARGET_WINS;
run;
```

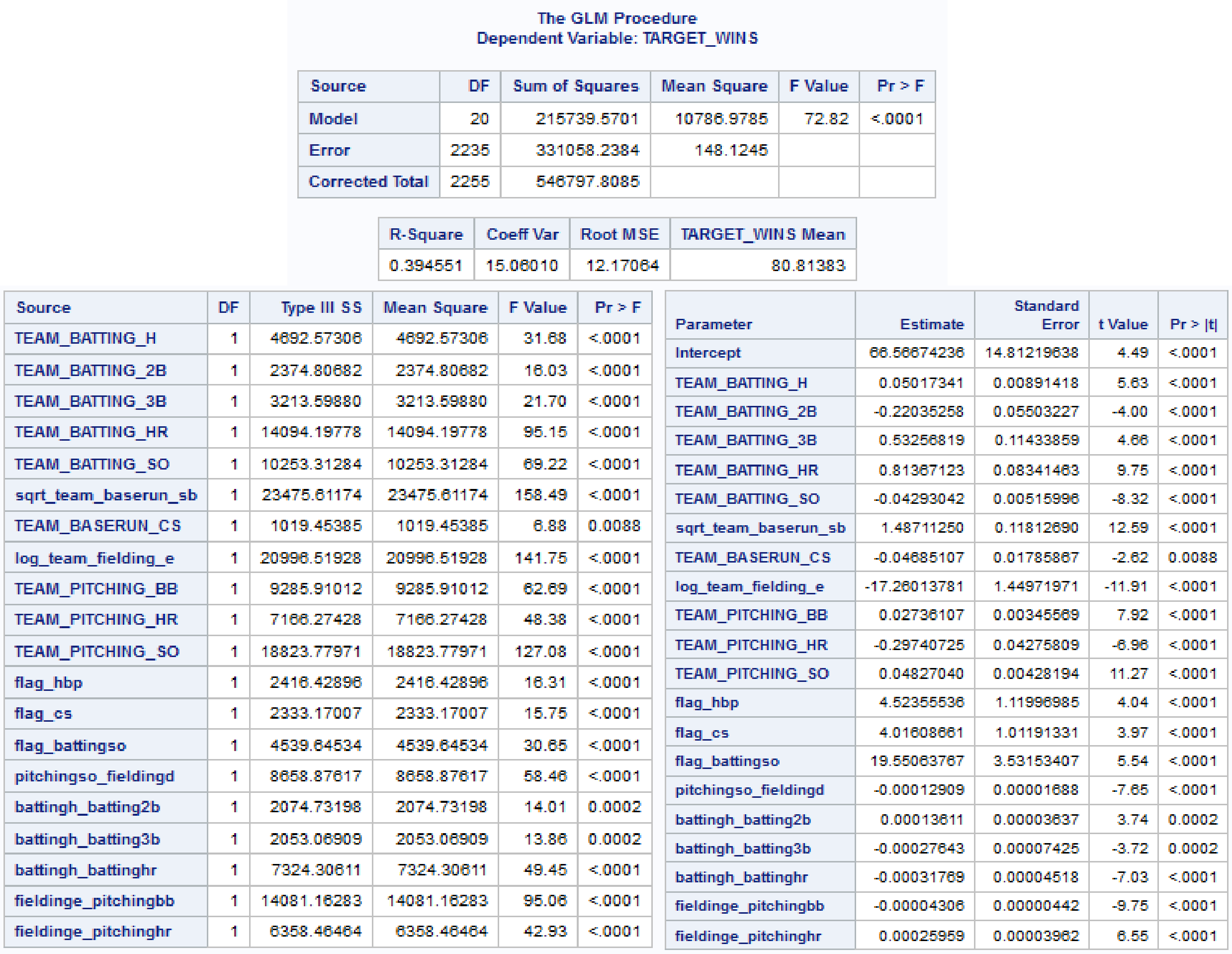
Scored Data File

The scored SAS data file including values for the INDEX and P_TARGET_WINS variables is attached separately.

Scored SAS Data Set +10 Bingo Bonus Points

PROC GLM and PROC GENMOD +20 Bingo Bonus Points

PROC GLM: Using PROC GLM to execute the OLS regression generates the same parameter estimates for my "best" model. The only apparent differences between PROC GLM and PROC REG are in the type of data displayed in the SAS results window. There are some metrics that appear in the results from executing one procedure, but not the other. For example, PROC GLM does not display the adjusted R^2 value, while PROC REG does. Also, the displayed results from PROC GLM include more digits to the right of the decimal points, compared to PROC REG.



The GLM Procedure
Dependent Variable: TARGET_WINS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	215739.5701	10786.9785	72.82	<.0001
Error	2235	331058.2384	148.1245		
Corrected Total	2255	546797.8085			

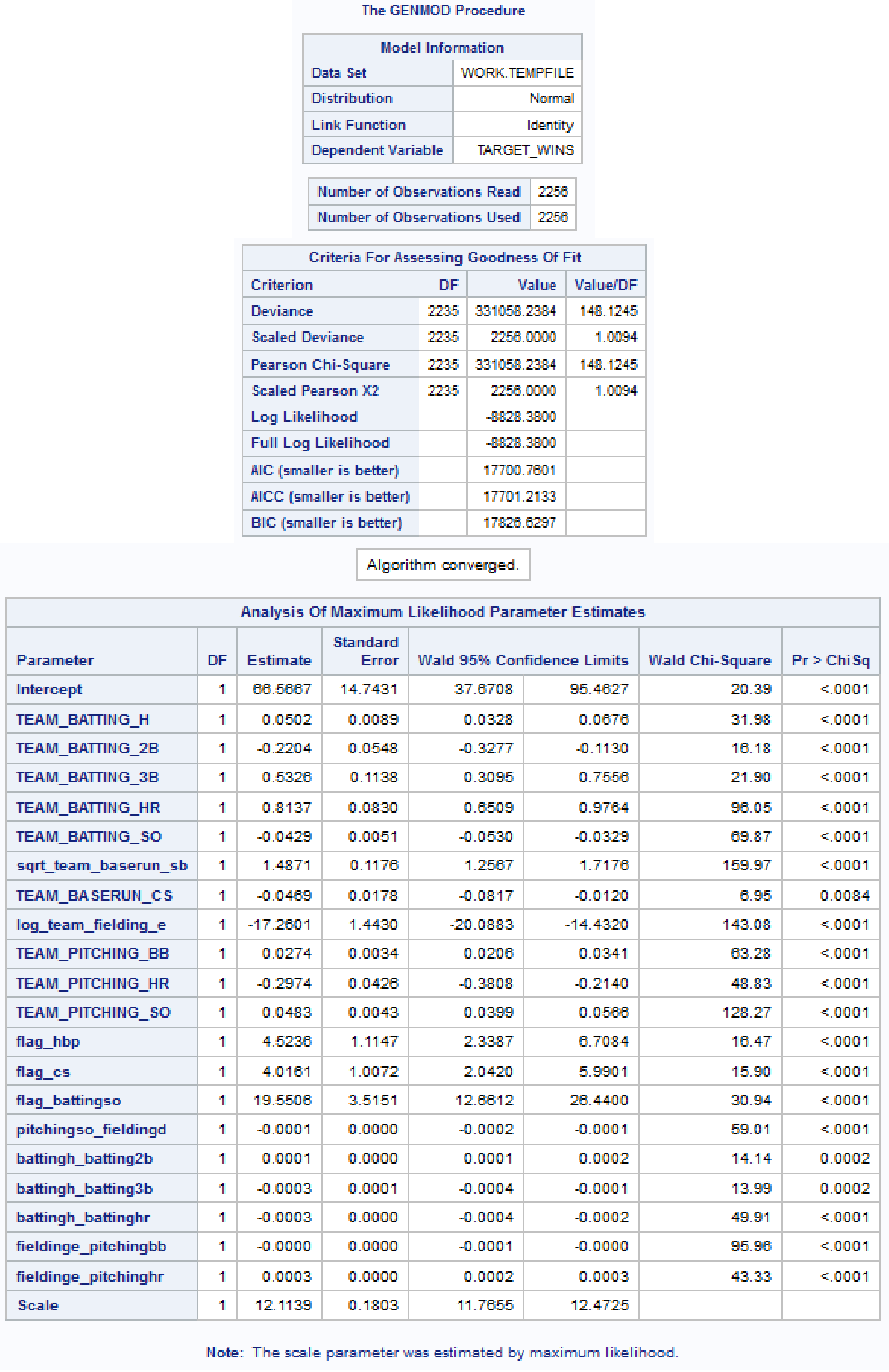
R-Square	Coeff Var	Root MSE	TARGET_WINS Mean
0.394551	15.06010	12.17064	80.81383

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TEAM_BATTING_H	1	4692.57306	4692.57306	31.68	<.0001
TEAM_BATTING_2B	1	2374.80682	2374.80682	16.03	<.0001
TEAM_BATTING_3B	1	3213.59880	3213.59880	21.70	<.0001
TEAM_BATTING_HR	1	14094.19778	14094.19778	95.15	<.0001
TEAM_BATTING_SO	1	10253.31284	10253.31284	69.22	<.0001
sqrt_team_baserun_sb	1	23475.61174	23475.61174	158.49	<.0001
TEAM_BASERUN_CS	1	1019.45385	1019.45385	6.88	0.0088
log_team_fielding_e	1	20996.51928	20996.51928	141.75	<.0001
TEAM_PITCHING_BB	1	9285.91012	9285.91012	62.69	<.0001
TEAM_PITCHING_HR	1	7166.27428	7166.27428	48.38	<.0001
TEAM_PITCHING_SO	1	18823.77971	18823.77971	127.08	<.0001
flag_hbp	1	2416.42896	2416.42896	16.31	<.0001
flag_cs	1	2333.17007	2333.17007	15.75	<.0001
flag_battingso	1	4539.64534	4539.64534	30.65	<.0001
pitchingso_fieldingd	1	8658.87617	8658.87617	58.46	<.0001
battingh_batting2b	1	2074.73198	2074.73198	14.01	0.0002
battingh_batting3b	1	2053.06909	2053.06909	13.86	0.0002
battingh_battinghr	1	7324.30611	7324.30611	49.45	<.0001
fieldinge_pitchingbb	1	14081.16283	14081.16283	95.06	<.0001
fieldinge_pitchinghr	1	6358.46464	6358.46464	42.93	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	66.56674236	14.81219638	4.49	<.0001
TEAM_BATTING_H	0.05017341	0.00891418	5.63	<.0001
TEAM_BATTING_2B	-0.22035258	0.05503227	-4.00	<.0001
TEAM_BATTING_3B	0.53256819	0.11433859	4.66	<.0001
TEAM_BATTING_HR	0.81367123	0.08341463	9.75	<.0001
TEAM_BATTING_SO	-0.04293042	0.00515996	-8.32	<.0001
sqrt_team_baserun_sb	1.48711250	0.11812690	12.59	<.0001
TEAM_BASERUN_CS	-0.04685107	0.01785867	-2.62	0.0088
log_team_fielding_e	-17.26013781	1.44971971	-11.91	<.0001
TEAM_PITCHING_BB	0.02736107	0.00345569	7.92	<.0001
TEAM_PITCHING_HR	-0.29740725	0.04275809	-6.96	<.0001
TEAM_PITCHING_SO	0.04827040	0.00428194	11.27	<.0001
flag_hbp	4.52355536	1.11996985	4.04	<.0001
flag_cs	4.01608661	1.01191331	3.97	<.0001
flag_battingso	19.55063767	3.53153407	5.54	<.0001
pitchingso_fieldingd	-0.00012909	0.00001688	-7.65	<.0001
battingh_batting2b	0.00013611	0.00003637	3.74	0.0002
battingh_batting3b	-0.00027643	0.00007425	-3.72	0.0002
battingh_battinghr	-0.00031769	0.00004518	-7.03	<.0001
fieldinge_pitchingbb	-0.00004306	0.00000442	-9.75	<.0001
fieldinge_pitchinghr	0.00025959	0.00003962	6.55	<.0001

PROC GENMOD: Using PROC GENMOD to execute the OLS regression generates the same parameter estimates for my "best" model, although the estimate values are truncated, compared to those generated from PROC GLM and PROC REG.

The standard errors generated via PROC GENMOD are slightly different than those generated from PROC GLM and PROC REG. Also, model fit appears to be evaluated a different way using PROC GENMOD, via metrics such as Deviance, AIC, AICC, and BIC. PROC GLM and PROC REG both reference R^2 and adjusted R^2 values. Also, the AIC values for this model generated using PROC GENMOD and PROC REG are different. The AIC value from PROC GENMOD is 17700, while the AIC value from PROC REG is 11297. Since lower AIC values indicate better model fit, PROC REG appears to be a better option for this linear regression model than PROC GENMOD.



The GENMOD Procedure

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

Number of Observations Read	2256
Number of Observations Used	2256

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	2235	331058.2384	148.1245
Scaled Deviance	2235	2256.0000	1.0094
Pearson Chi-Square	2235	331058.2384	148.1245
Scaled Pearson X2	2235	2256.0000	1.0094
Log Likelihood		-8828.3800	
Full Log Likelihood		-8828.3800	
AIC (smaller is better)		17700.7601	
AICC (smaller is better)		17701.2133	
BIC (smaller is better)		17826.6297	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	66.5667	14.7431	37.6708	95.4627	20.39	<.0001
TEAM_BATTING_H	1	0.0502	0.0089	0.0328	0.0676	31.98	<.0001
TEAM_BATTING_2B	1	-0.2204	0.0548	-0.3277	-0.1130	16.18	<.0001
TEAM_BATTING_3B	1	0.5326	0.1138	0.3095	0.7556	21.90	<.0001
TEAM_BATTING_HR	1	0.8137	0.0830	0.6509	0.9764	96.05	<.0001
TEAM_BATTING_SO	1	-0.0429	0.0051	-0.0530	-0.0329	69.87	<.0001
sqrt_team_baserun_sb	1	1.4871	0.1176	1.2567	1.7176	159.97	<.0001
TEAM_BASERUN_CS	1	-0.0469	0.0178	-0.0817	-0.0120	6.95	0.0084
log_team_fielding_e	1	-17.2601	1.4430	-20.0883	-14.4320	143.08	<.0001
TEAM_PITCHING_BB	1	0.0274	0.0034	0.0206	0.0341	63.28	<.0001
TEAM_PITCHING_HR	1	-0.2974	0.0426	-0.3808	-0.2140	48.83	<.0001
TEAM_PITCHING_SO	1	0.0483	0.0043	0.0399	0.0566	128.27	<.0001
flag_hbp	1	4.5236	1.1147	2.3387	6.7084	16.47	<.0001
flag_cs	1	4.0161	1.0072	2.0420	5.9901	15.90	<.0001
flag_battingso	1	19.5506	3.5151	12.6612	26.4400	30.94	<.0001
pitchingso_fieldingd	1	-0.0001	0.0000	-0.0002	-0.0001	59.01	<.0001
battingh_batting2b	1	0.0001	0.0000	0.0001	0.0002	14.14	0.0002
battingh_batting3b	1	-0.0003	0.0001	-0.0004	-0.0001	13.99	0.0002
battingh_battinghr	1	-0.0003	0.0000	-0.0004	-0.0002	49.91	<.0001
fieldinge_pitchingbb	1	-0.0000	0.0000	-0.0001	-0.0000	95.96	<.0001
fieldinge_pitchinghr	1	0.0003	0.0000	0.0002	0.0003	43.33	<.0001
Scale	1	12.1139	0.1803	11.7655	12.4725		

Note: The scale parameter was estimated by maximum likelihood.