Data 621 Homework 1: Moneyball

Critical Thinking Group 2 - Armenoush Aslanian-Persico, James Topor, Jeff Nieman, Scott Karr

---

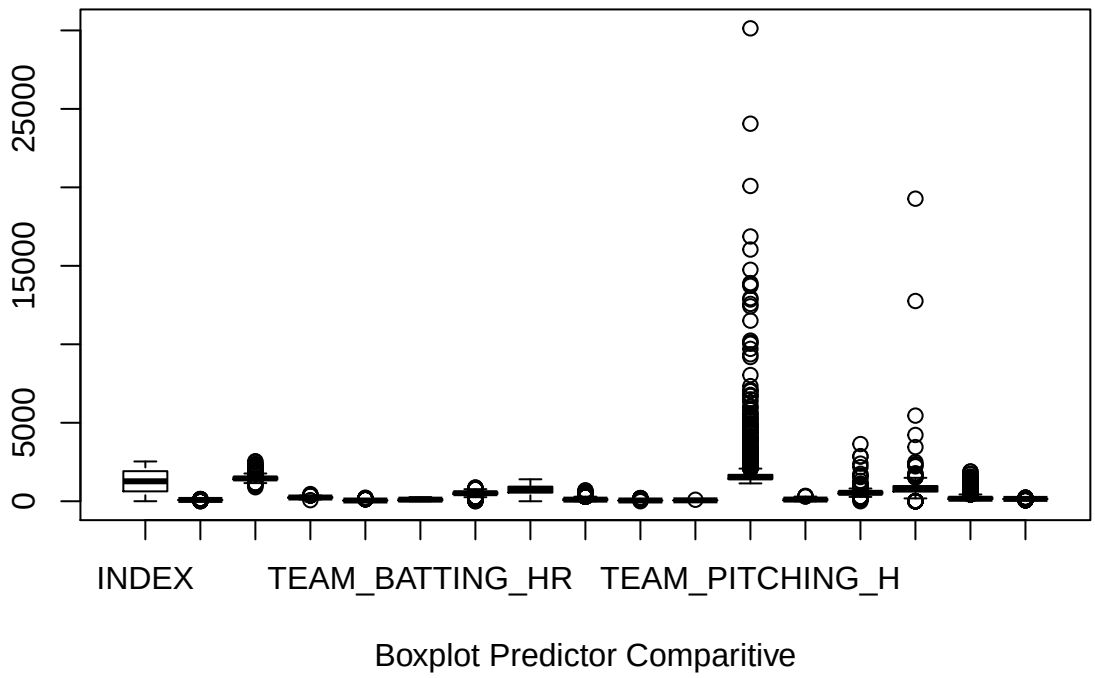
Part 1: Data Exploration

Data Summary

The original Training data set is comprised of 17 elements and 2277 total observations. Of those 17 elements, INDEX is simply an index value used for sorting while TARGET_WINS represents the response variable we are to use within our regression models. The remaining 15 elements are all potential predictor variables for our linear models. A summary table for the data set is provided below.

variables	n	mean	sd	med	min	max	range	skew	kurtosis	se	NAs
INDEX	2276	1269	736	1271	1	2535	2534	0.00	-1.22	15.43	
TARGET_WINS	2276	81	16	82	0	146	146	-0.40	1.03	0.33	
TEAM_BATTING_H	2276	1470	145	1454	891	2554	1663	1.57	7.28	3.03	
TEAM_BATTING_2B	2276	241	47	238	69	458	389	0.22	0.01	0.98	
TEAM_BATTING_3B	2276	55	28	47	0	223	223	1.11	1.50	0.59	
TEAM_BATTING_HR	2276	100	61	102	0	264	264	0.19	-0.96	1.27	
TEAM_BATTING_BB	2276	502	123	512	0	878	878	-1.03	2.18	2.57	
TEAM_BATTING_SO	2174	736	240	750	0	1399	1399	-0.30	-0.32	5.33	102
TEAM_BASERUN_SB	2145	125	88	101	0	697	697	1.97	5.49	1.90	131
TEAM_BASERUN_CS	1504	53	23	49	0	201	201	1.98	7.62	0.59	772
TEAM_BATTING_HBP	191	59	13	58	29	95	66	0.32	-0.11	0.94	2085
TEAM_PITCHING_H	2276	1779	1407	1518	1137	30132	28995	10.33	141.84	29.49	
TEAM_PITCHING_HR	2276	106	61	107	0	343	343	0.29	-0.60	1.28	
TEAM_PITCHING_BB	2276	553	166	536	0	3645	3645	6.74	96.97	3.49	
TEAM_PITCHING_SO	2174	818	553	813	0	19278	19278	22.17	671.19	11.86	102
TEAM_FIELDING_E	2276	246	228	159	65	1898	1833	2.99	10.97	4.77	
TEAM_FIELDING_DP	1990	146	26	149	52	228	176	-0.39	0.18	0.59	286

The first observation of the chart above show that there are missing values in 6 fields. The problem is particularly significant for caught stealing. A survey of box plots show that there are significant outliers in some of the data columns, especially in TEAM_PITCHING_H and TEAM_PITCHING_SO.



Correlation Plot

Using the cor function across the data frame we noticed some strong correlations. TEAM_BATTING_H obviously has some colinearity with TEAM_BATTING_2B, TEAM_BATTING_3B and TEAM_BATTING_HR as these values are a subset of hits. TEAM_BATTING_BB and TEAM_PITCHING_BB have strong correlations and so do TEAM_PITCHING_HR and TEAM_BATTING_HR. Since we are most concerned with wins, the following table shows the correlation when the NA's are ommited:

Value	Correlation with Wins
TEAM_BATTING_H	0.46994665
TEAM_BATTING_2B	0.31298400
TEAM_BATTING_3B	-0.12434586
TEAM_BATTING_HR	0.42241683
TEAM_BATTING_BB	0.46868793
TEAM_BATTING_SO	-0.22889273
TEAM_BASERUN_SB	0.01483639
TEAM_BASERUN_CS	-0.17875598
TEAM_BATTING_HBP	0.07350424
TEAM_PITCHING_H	0.47123431
TEAM_PITCHING_HR	0.42246683
TEAM_PITCHING_BB	0.46839882
TEAM_PITCHING_SO	-0.22936481
TEAM_FIELDING_E	-0.38668800
TEAM_FIELDING_DP	0.13168916

As a result of missing data, severe outliers, and collinearity shown above there is a clear need for some data preparation and transformation.

Part 2: Data Preparation

Our data preparation efforts for the training data set consisted of the creation of one new derived variable, removing four predictor variables, imputing values for the remaining variables that had missing values (NA's), and removal of a relatively small number of records that contained clearly egregious outlier values for particular variables. The results of these efforts were subsequently used as the basis for each of the five different linear models we created and evaluated.

New Variable Creation

We began our data preparation efforts by creating a new variable TEAM_BATTING_1B which represents offensive single base hits. The variable was created by subtracting out the TEAM_BATTING doubles, triples and home runs from the TEAM_BATTING_H variable.

The TEAM_BATTING_H variable was then dropped from the data set. We believe that separating out singles from the other unique hit values will minimize collinearity.

Removal of Variables Due to Collinearity and/or Missing Values

The results of our data exploration efforts led us to drop three other variables from the data set:

1. **TEAM_BATTING_HBP**: The TEAM_BATTING_HBP variable has very little correlation with the TARGET_WINS response variable and also has 2085 missing values out of a total of 2277. Since it would be very difficult to accurately impute such a large proportion of any variable's missing values, we chose to exclude the variable from our analysis.
2. **TEAM_BASERUN_CS**: This variable is strongly correlated (65.5%) with the TEAM_BASERUN_SB variable and is the 2nd largest source of NA's in our data set. These combined facts led us to exclude the variable from our analysis.
3. **TEAM_PITCHING_HR**: This variable is 97% correlated with TEAM_BATTING_HR. In fact, 815 cases (more than 35% of our total cases) have *IDENTICAL* values for pitched and batted HR's. This high degree of correlation may be due to the time series nature of the data: as baseball evolved, more home runs were hit, which naturally causes the number of pitched home runs to increase. The statistics are basically opposite sides of the same coin to a large degree (even if there may be some variability between individual teams in any given year). The fact that these two variables are nearly perfectly correlated indicates that one of them can legitimately be removed from the data set, and we chose **TEAM_PITCHING_HR** since we believe the batting HR metric will be more predictive of TARGET_WINS than will the pitching HR statistic.

These same three variables were subsequently removed from the Evaluation data set to ensure compatibility with our linear models.

Imputation of Missing Values (NA's)

Filling Missing values in the Training Data Set

We then worked on imputing the remaining missing data. To do this we made use of a linear regression approach recommended by Faraway (p.201) and Fox (p.611). We decided against the using the mean or median as a replacement value for NA's since regression yields imputed values that are much more consistent with the actual distribution of the data while introducing much less potential bias than use of a mean or median would incur.

We built regression imputation models for the variables listed below. Each model was analyzed to ensure that there were no collinearity issues and all p-values were $< .05$. While we did not not have the time or resources to run a complete cadre of residual diagnostics on these imputation models, we did verify that their results produced imputed distributions of the subject variables that were consistent with those of the original NA-populated data. It is our belief that this consistency indicates that the resulting predicted values for the missing values are better to use than simply filling the NA's with a mean or median.

The variables we built imputation regression models for are described below. The R code for each of the regression models is contained in the Appendix.

1. **TEAM_BATTING_SO**: The adjusted R^2 value for this regression model was 0.7223 and, after populating the NA's using the imputed values, yielded a distribution that matched that of the variable prior to the NA's being filled. For this variable we imputed a total of 131 missing values via regression.
2. **TEAM_PITCHING_SO**: The adjusted R^2 value for this regression model was 0.9952. A total of 102 missing values were imputed via regression for this variable.
3. **TEAM_BASERUN_SB**: The adjusted R^2 value for this regression model was 0.3427. Despite the adjusted R^2 being low relative to the models described above, after the model yielded a distribution that matched that of the variable prior to the NA's being filled. A total of 131 missing values were imputed via regression for this variable.
4. **TEAM_FIELDING_DP**: The adjusted R^2 for this model was 0.3904. A total of 286 missing values were imputed and the resulting distribution matched that of the variable prior to the NA's being filled.

Filling Missing Values in the Evaluation Data Set

These four imputation regression models were then used to populate missing values for the same variables within the evaluation data set. Clearly, any models we might have constructed for purposes of predicting WINS for the evaluation set would have failed to work on any records within that data set that contained NA's. Applying the imputation models to the evaluation data allows us to avoid that problem.

Removal of Extreme Outliers From the Training Data Set

We completed our data preparation efforts by eliminating some clearly egregious outliers identified via research through *baseball-almanac.com*. This approach is suggested by Sheather (p. 57). For example, the record for the most pitching strikeouts in a single season is 1450 by the 2014 Cleveland Indians. Therefore we know that any records having TEAM_PITCHING_SO values above that point is an aberration, and any records containing such values were summarily removed from the data set.

Similarly, the most errors by team in a single season was 639 by Philadelphia in 1883. Prorating to 162 games we calculated that we should discard any records containing TEAM_FIELDING_E values above 1046.

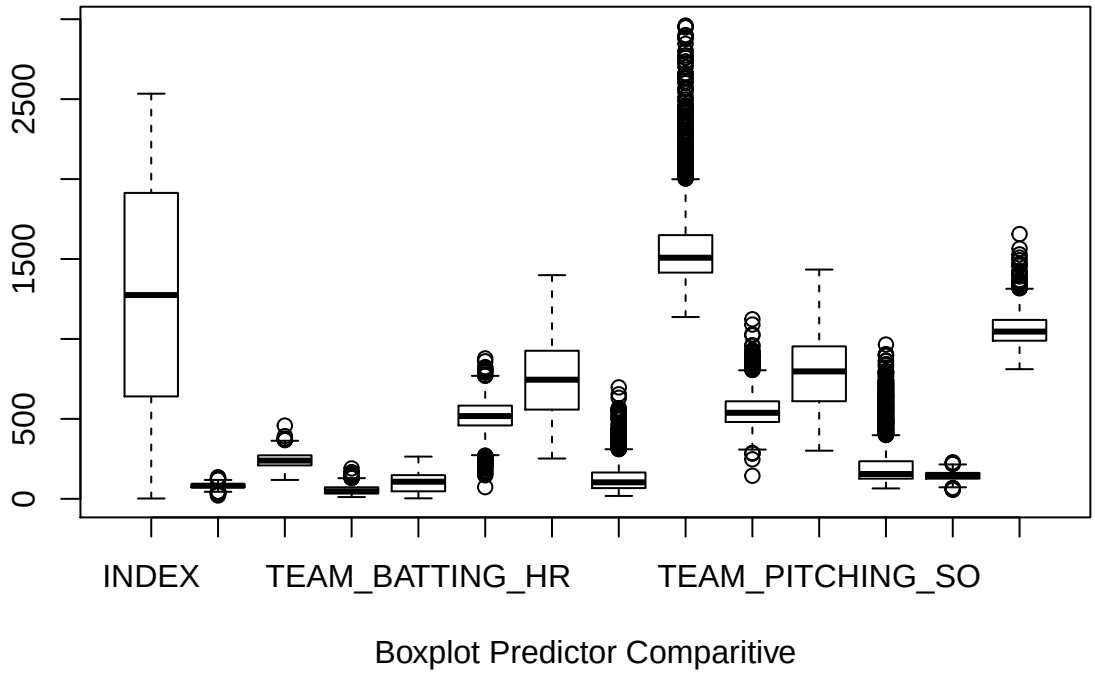
The TEAM_PITCHING_H variable also appeared to have numerous egregious outliers. For example, the most offensive hits by a team in a single season was 1730. As such, it is highly unlikely that any pitching

staff would surrender more than 3000 hits in a single season since such a total would indicate the team was allowing more than 18 hits per game. As such, any records having a TEAM_PITCHING_H value > 3000 was removed from the data set.

Removing the egregious outliers resulting in a sum total of 104 records being removed from the training data set. Removal of these outliers helped to normalize our data and thereby improve the expected performance of our linear models.

Results of Imputation for Missing Values and Outlier Removal Process

The end results of our data transformation process dramatically improved the data as you can see from the following plot. There still are a few outliers but notice the dramatically smaller scale.



Boxplot Predictor Comparitive

```
##      INDEX         TARGET_WINS     TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :   2.0   Min.   : 21.00   Min.   :118.0   Min.   : 11.00
##  1st Qu.: 640.8   1st Qu.: 72.00   1st Qu.:209.0   1st Qu.: 34.00
##  Median :1274.5   Median : 82.00   Median :239.0   Median : 47.00
##  Mean   :1271.3   Mean   : 81.06   Mean   :241.6   Mean   : 54.41
##  3rd Qu.:1913.2   3rd Qu.: 91.00   3rd Qu.:272.0   3rd Qu.: 72.00
##  Max.   :2534.0   Max.   :135.00   Max.   :458.0   Max.   :190.00
##  TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  3     Min.   : 73.0   Min.   : 252.0   Min.   : 18.0
##  1st Qu.: 47     1st Qu.:459.0   1st Qu.: 558.0   1st Qu.: 67.0
##  Median :107     Median :517.5   Median : 745.0   Median :103.0
##  Mean   :103     Mean   :516.2   Mean   : 744.1   Mean   :130.5
##  3rd Qu.:148     3rd Qu.:583.0   3rd Qu.: 926.0   3rd Qu.:164.2
```

```
##  Max.   :264     Max.   :878.0   Max.   :1399.0   Max.   :697.0
##  TEAM_PITCHING_H TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E
##  Min.   :1137    Min.   : 144.0   Min.   : 301.0   Min.   : 65
##  1st Qu.:1415    1st Qu.: 480.8   1st Qu.: 610.8   1st Qu.:126
##  Median :1508    Median : 538.0   Median : 797.0   Median :155
##  Mean   :1575    Mean   : 550.4   Mean   : 789.5   Mean   :213
##  3rd Qu.:1649    3rd Qu.: 610.0   3rd Qu.: 953.2   3rd Qu.:235
##  Max.   :2960    Max.   :1123.0   Max.   :1434.0   Max.   :965
##  TEAM_FIELDING_DP TEAM_BATTING_1B
##  Min.   : 56.0    Min.   : 811
##  1st Qu.:126.0    1st Qu.: 989
##  Median :146.0    Median :1046
##  Mean   :143.2    Mean   :1061
##  3rd Qu.:162.0    3rd Qu.:1119
##  Max.   :228.0    Max.   :1656
```

Our training data set with the NA's filled and the outliers removed can be found here:

https://github.com/spsstudent15/2016-02-621-W1/blob/master/621-HW1-Clean-Data.csv

Our Evaluation data set with the NA's filled can be found here:

https://github.com/spsstudent15/2016-02-621-W1/blob/master/621-HW1-Clean-EvalData-.csv

The summary of the clean data is as follows:

variables	n	mean	sd	med	min	max	range	skew	kurtosis	se	NAs
TARGET_WINS	2172	81	15	82	21	135	114	-0.22	0.08	0.31	
TEAM_BATTING_2B	2172	242	46	239	118	458	340	0.23	-0.13	0.98	
TEAM_BATTING_3B	2172	54	27	47	11	190	179	0.99	0.65	0.58	
TEAM_BATTING_HR	2172	103	59	107	3	264	261	0.16	-0.93	1.27	
TEAM_BATTING_BB	2172	516	100	518	73	878	805	-0.32	0.96	2.15	
TEAM_BATTING_SO	2172	744	226	745	252	1399	1147	0.06	-0.98	4.85	
TEAM_BASERUN_SB	2172	131	93	103	18	697	679	1.75	3.91	1.99	
TEAM_PITCHING_H	2172	1575	256	1508	1137	2960	1823	2.10	5.88	5.48	
TEAM_PITCHING_BB	2172	551	107	538	144	1123	979	0.70	1.40	2.29	
TEAM_PITCHING_SO	2172	789	223	797	301	1434	1133	0.15	-0.63	4.79	
TEAM_FIELDING_E	2172	213	148	155	65	1965	900	2.18	4.55	3.18	
TEAM_FIELDING_DP	2172	143	28	146	56	228	172	-0.30	-0.14	0.59	
TEAM_BATTING_1B	2172	1061	102	1046	811	1656	845	0.89	1.49	2.20	

In comparison with the same chart there are much smaller skews kurtosis and se, and the wild ranges in values such as pitching hits and pitching strikeouts have been reduced.

Other Data Prepation Transformations: Refer to Model Descriptions

Other *model-specific* data transformations were also used, including use of Box-Cox power transforms and linear combinations of variables. These model-specific transformations are discussed within the individual model writeups provided in **Part 3**.

Part 3: Build Models

Model 1: General Model Using Backward Selection

This model applies simple Backward Selection methods through the use of p-values and variance inflation factors (VIF) against all 12 remaining predictor variables:

Simply removing the *TEAM_BATTING_1B* variable yielded a model with all p-values less than .05. However, VIF analysis showed evidence of multiple collinear variables within the model. Subsequent removals of *TEAM_PITCHING_SO* and *TEAM_PITCHING_BB* due to collinearity yielded a model that called for the removal of *TEAM_BATTING_2B* on the basis of its p-value.

The final model of that iteration of the linear modeling process showed clear evidence of a number of outliers as evidenced in R's summary diagnostic plots. Those outliers were removed via a series of additional iterations yielding the following final model, which varies from the inital iteration in that it includes *TEAM_BATTING_2B* due to the fact that removing the outliers improved the statistical significance of the variable :

Coefficient	Variable
66.261	Intercept
- 0.017	TEAM_BATTING_2B
+ 0.150	TEAM_BATTING_3B
+ 0.109	TEAM_BATTING_HR
+ 0.022	TEAM_BATTING_BB
- 0.019	TEAM_BATTING_SO
+ 0.065	TEAM_BASERUN_SB
+ 0.016	TEAM_PITCHING_H
- 0.075	TEAM_FIELDING_E
- 0.109	TEAM_FIELDING_DP

RSE	R^2	Adj. R^2	F Stat.	MSE
11.49	0.3598	0.3572	134.4	132

However, the diagnostic plots of that model showed a lack of linearity between the response variable TARGET_WINS and the predictor variable TEAM_FIELDING_E as evidenced in the Added Variable plots shown in the Appendix. Furthermore, the plots of standardized residuals against each of the predictor variables showed evidence of non-constant variability for variables such as TEAM_BATTING_HR, TEAM_BATTING_SO, TEAM_BASERUN_SB, and TEAM_FIELDING_E.

The TEAM_FIELDING_E variable was subsequently transformed using a Box-Cox recommended power transform of (-1), or (1/y) and the model was re-run. The resulting Added Variable plots showed that all predictors are linearly related to the response, and we see an improvement in the variability of the residuals relative to TEAM_FIELDING_E. Furthermore, the plot of Y against the fitted values showed an improvement in the linearity of the model.

Therefore, this model appears to be an improvement over the first model when the residual plots are considered. The characteristic equation indicated by the model is as follows:

Coefficient	Variable
52.88	Intercept
+ 0.168	TEAM_BATTING_3B
+ 0.096	TEAM_BATTING_HR

Coefficient	Variable
+ 0.027	TEAM_BATTING_BB
- 0.027	TEAM_BATTING_SO
+ 0.034	TEAM_BASERUN_SB
+ 0.004	TEAM_PITCHING_H
+ 3252.31	1/TEAM_FIELDING_E
- 0.102	TEAM_FIELDING_DP

RSE	R^2	Adj. R^2	F Stat.	MSE
11.86	0.3168	0.3143	124.8	141

The coefficients for TEAM_BATTING_3B, TEAM_BATTING_HR, TEAM_BATTING_BB, TEAM_BATTING_SO, and TEAM_BASERUN_SB all make sense intuitively. The TEAM_FIELDING_DP coefficient is less intuitive since one would expect more defensive double plays to improve a team's chances of winning games. However, the variable itself is *negatively* correlated with TARGET_WINS as shown in the **Data Exploration** section above. As such, we shouldn't be surprised to see a negative coefficient for it here. Similarly, the coefficient for TEAM_PITCHING_H is also counterintuitive, but the variable is actually positively correlated with TARGET_WINS as shown in the **Data Exploration** section. Finally, TEAM_FIELDING_E has changed from negative in the earlier model to positive here. However, this is due to the fact that the coefficient now applies to the *transformed* version of the variable rather than the nominal values of the variable.

While this model is an improvement over the initial model, we still have component variables that appear to lack constant variability relative to the residuals for variables such as TEAM_BASERUN_SB. The lack of constant variability in the residuals is likely related to the skewed nature of the distributions of those individual variables.

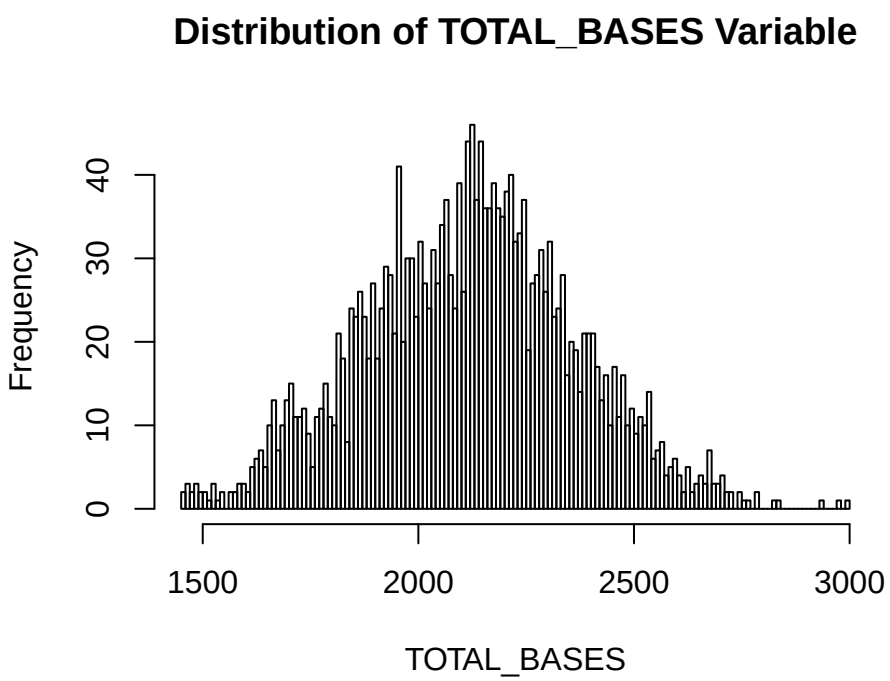
In the other models discussed herein we attempt to address some of the skew issues via various methods, including Box-Cox recommended power transforms and linear combinations of various variables.

Model 2: Total Bases

This model attempts to address some of the lack of constant variability found in the "General Model" discussed above by employing a linear combination of four of the predictor variables to calculate the baseball statistc known as "Total Bases". Total Bases is calculated using what our data set refers to as "TEAM_BATTING" variables as follows:

- Singles + (2 * Doubles) + (3 * Triples) = (4 * Home Runs)

Inclusion of this new variable allows us to eliminate the four component variables from the model. In fact, the TOTAL_BASES variable appears to be nearly normally distributed, thereby negating the skew issues that were evident with its component variables.



This model applies simple Backward Selection methods through the use of p-values and variance inflation factors (VIF) against a derived value for total bases and the remaining 8 predictors:

Three iterations of p-value / VIF backward selection removed TEAM_PITCHING_SO and TEAM_PITCHING_BB from the model. All other variables remained statistically significant with no signficant collinearity. However, evidence of multiple outliers was found via R's summary diagnostic plots. Those outliers were removed via a series of additional iterations yielding the following final model:

Coefficient	Variable
48.486	Intercept
+ 0.022	TEAM_BATTING_BB
- 0.015	TEAM_BATTING_SO
+ 0.063	TEAM_BASERUN_SB
+ 0.010	TEAM_PITCHING_H
- 0.064	TEAM_FIELDING_E
- 0.117	TEAM_FIELDING_DP
+ 0.018	TOTAL_BASES

RSE	R^2	Adj. R^2	F Stat.	MSE
11.7	0.3365	0.3343	156	137

The diagnostic plots of that model show a lack of linearity between the response variable TARGET_WINS and the predictor variable TEAM_FIELDING_E as evidenced in the Added Variable plots shown in the Appendix. Furthermore, the plots of standardized residuals against each of the predictor variables showed evidence of non-constant variability for variables such as TEAM_BATTING_SO, TEAM_BASERUN_SB, and TEAM_FIELDING_E.

The TEAM_FIELDING_E variable was subsequently transformed using a Box-Cox recommended power transform of (-1), or (1/y) and the model was re-run. The resulting Added Variable plots show that all predictors are linearly related to the response, and we see an improvement in the variability of the residuals relative to TEAM_FIELDING_E. Furthermore, the plot of Y against the fitted values shows an improvement in the linearity of the model.

Therefore, this model appears to be an improvement over the first TOTAL_BASES model when the residual plots are considered. The characteristic equation indicated by the model is as follows:

Coefficient	Variable
39.164	Intercept
+ 0.025	TEAM_BATTING_BB
- 0.025	TEAM_BATTING_SO
+ 0.038	TEAM_BASERUN_SB
+ 2714.54	1/TEAM_FIELDING_E
- 0.115	TEAM_FIELDING_DP
+ 0.0197	TOTAL_BASES

RSE	R^2	Adj. R^2	F Stat.	MSE
11.97	0.3048	0.3029	157.5	143

The coefficients for TEAM_BATTING_BB, TEAM_BATTING_SO, TEAM_BASERUN_SB, and TOTAL_BASES all make sense intuitively. The TEAM_FIELDING_DP coefficient is less intuitive since one would expect more defensive double plays to improve a team's chances of winning games. However, the variable itself is *negatively* correlated with TARGET_WINS as shown in the **Data Exploration** section above. As such, we shouldn't be surprised to see a negative coefficient for it here. TEAM_FIELDING_E has changed from negative in the earlier model to positive here. However, this is due to the fact that the coefficient now applies to the *transformed* version of the variable rather than the nominal values of the variable.

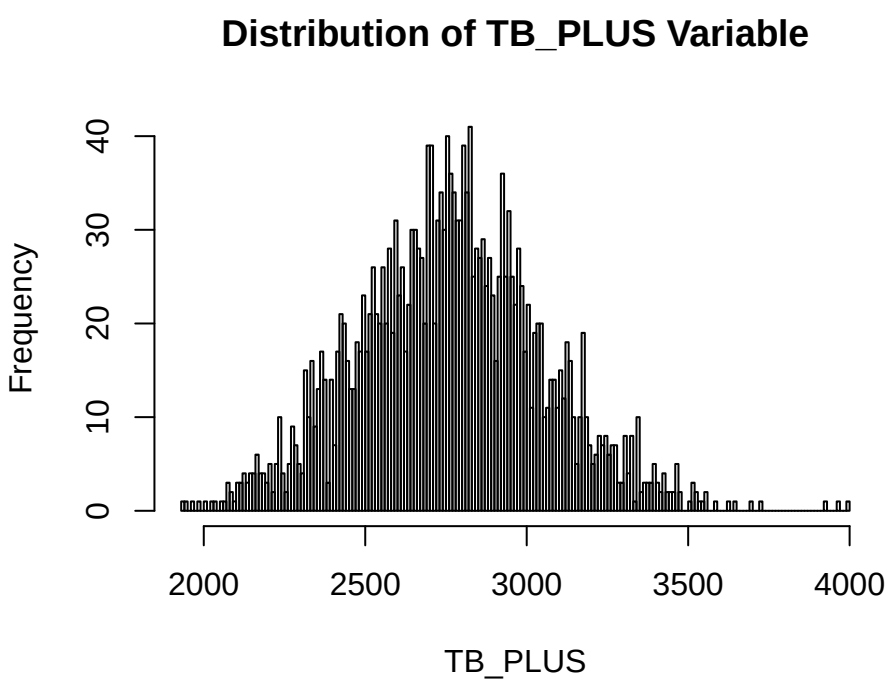
Model 3: Total Bases PLUS

This model attempts to improve upon the results of the "Total Bases" model by extending the TOTAL_BASES variable to include the TEAM_BATTING_BB and TEAM_BASERUN_SB variables. The logic behind adding these two variables to the TOTAL_BASES variable comes from the fact that both, like the component variables of TOTAL_BASES, represent basepath advancements by a team's offense.

"Total Bases Plus"" (referred to as TB_PLUS hereon) is calculated using what our data set refers to as "TEAM_BATTING" and "TEAM_BASERUN" variables as follows:

- Singles + (2 * Doubles) + (3 * Triples) = (4 * Home Runs) + BB + SB

Inclusion of this new variable allows us to eliminate the two additional component variables from the model. In fact, the TB_PLUS variable, like the TOTAL_BASES variable used earlier appears to be nearly normally distributed, thereby negating the skew issues that were evident with its component variables. A histogram of the distribution of the derived TB_PLUS variable is shown below:



This model applies simple Backward Selection methods through the use of p-values and variance inflation factors (VIF) against the derived value for TB_PLUS and the remianing 6 predictor variables:

Four iterations of p-value / VIF backward selection removed TEAM_PITCHING_H, TEAM_PITCHING_SO and TEAM_PITCHING_BB from the model. All other variables remained statistically significant with no signficant collinearity. However, evidence of multiple outliers was found via R's summary diagnostic plots. Those outliers were removed via a series of additional iterations yielding the following final model:

Coefficient	Variable
52.330	Intercept
- 0.016	TEAM_BATTING_SO
- 0.034	TEAM_FIELDING_E
- 0.154	TEAM_FIELDING_DP
+ 0.025	TB_PLUS

RSE	R^2	Adj. R^2	F Stat.	MSE
12.12	0.2944	0.2931	225.5	145

The coefficients for TEAM_BATTING_SO and TEAM_FIELDING_E make sense intuitively: the more strikeouts a team's offense has, the less likely it is to put the ball in play, and the more fielding errors a team commits, the more likely they are to lose games. The TEAM_FIELDING_DP coefficient is less intuitive since one would expect more defensive double plays to improve a team's chances of winning games. However, the variable itself is *negatively* correlated with TARGET_WINS as shown in the **Data Exploration** section above. As such, we shouldn't be surprised to see a negative coefficient for it here. Finally, the coefficient for TB_PLUS is postively correlated with the response variable, which shouldn't surprise us since it encapsulates all of a team's offense hits, stolen bases, and bases on balls.

As with the "General Model" and the "Total Bases" model, the diagnostic plots for this model showed a lack of linearity between the response variable TARGET_WINS and the predictor variable TEAM_FIELDING_E as evidenced in the Added Variable plots shown in the Appendix. Furthermore, the plots of standardized residuals against each of the predictor variables showed evidence of non-constant variability for the variables TEAM_BATTING_SO and TEAM_FIELDING_E.

The TEAM_FIELDING_E variable was subsequently transformed using a Box-Cox recommended power transform of (-1), or (1/y) and the model was re-run. The resulting Added Variable plots showed that all predictors are linearly related to the response, and we found an improvement in the variability of the residuals relative to TEAM_FIELDING_E. Furthermore, the plot of Y against the fitted values shows a non-skewed linear relationship.

Therefore, this model appears to be an improvement over the first TB_PLUS model when the residual plots are considered. Furthermore, the number of predictor variables used here is two fewer than that of the "Total Bases" model discussed earlier. The characteristic equation indicated by the model is as follows:

Coefficient	Variable
42.160	Intercept
- 0.023	TEAM_BATTING_SO
+ 2366.82	1/TEAM_FIELDING_E
- 0.140	TEAM_FIELDING_DP
+ 0.022	TB_PLUS

RSE	R^2	Adj. R^2	F Stat.	MSE
12.13	0.2932	0.2919	223.3	147

As we can see, the coefficient for TEAM_FIELDING_E has changed from negative to positive. However, this is due to the fact that the coefficient now applies to the *transformed* version of the variable rather than the nominal values of the variable. The other coefficients remain the same.

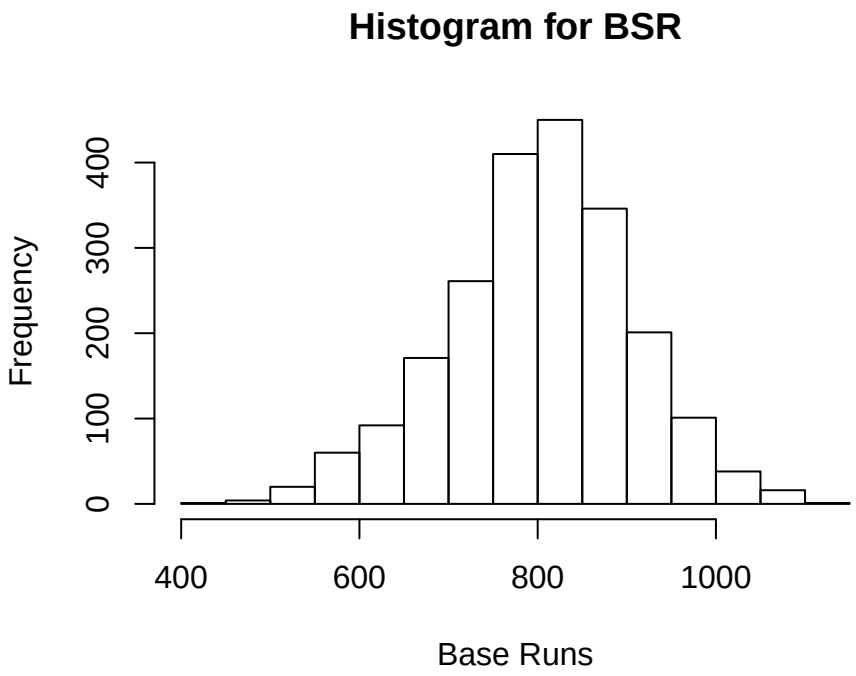
Model 4: Sabermetrics Model

Sabermetrics has become the rage in baseball, actually popularized by Billy Beane and the data set we are exploring. As a result of this, we built a model that centers around one of these advance analytics known as BsR or base runs. This statistic was designed by David Smyth in the 1990's and estimates the amount of runs a team SHOULD score, which made a unique approach as the data set provided did not include runs (see http://tangotiger.net/wiki_archive/Base_Runs.html for more information). The formula is as follows:

BSR = A*B/(B+C) +D where:

- A = TEAM_BATTING_1B + TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_BB
- B = 1.02*(1.4* TEAM_TOTAL_BASES -0.6*TEAM_BATTING_H + 0.1* TEAM_BATTING_BB)
- C = AT BATS - TEAM_BATTING_H (which we approximated with 3*TEAM_BATTING_H as the average batting average is around 0.250)
- D = TEAM_BATTING_HR

Since we eliminated the value of TEAM_BATTING_H we simply summed up singles, doubles, triples and home runs in the actual code, and the approach for TEAM_TOTAL_BASES is described in model 2. The data for BSR exhibit a fairly normal distribution.



Since BSR is a combination of all of the batting variables, we simply eliminated them and created a very strong model on the first iteration. All p-values were very low, and the variation are all below 5 showing no probems with collinearity. The characteristic equation indicated by the model is as follows:

Coefficient	Variable
40.687320	Intercept
+ 0.062189	BSR
- 0.116615	TEAM_FIELDING_DP
- 0.058885	TEAM_FIELDING_E
+ 0.060347	TEAM_BASERUN_SB
- 0.011457	TEAM_PITCHING_SO
+ 0.019419	TEAM_PITCHING_H
- 0.017603	TEAM_PITCHING_BB

RSE	R^2	Adj. R^2	F Stat.	MSE
11.99	0.3229	0.3207	147.4	144

The coefficients for this model overall do make sense. Errors and pitching walks contribute to fewer wins, and stolen bases and the BSR metric have strong influence on increasing wins. Double plays do have a slightly negative value, although this could be explained by a team allowing a large number of baserunners, and in anay case this matches the relationship of the data when you compare against target wins. The positive impact of allowing pitching hits was puzzling, but once again it agrees with the trends we see in the data

provided.

Model 5: Box-Cox First

This model first applied transformations to a simple regression of each predictor against wins and showed improvements in the normality of the distribution and more uniformly distributed residuals. In each case a Box-Cox transformation was applied to improve the skewness of predictor variables prior to becoming candidates for the multi-regression model.

The model then applied a simple Forward Selection strategy, adding variables two-at-a-time until none can be found that improve the model as measured by adjusted R^2.

Pre-model transformation of individual predicotor variables were as follows:

- TEAM_PITCHING_BB (boxcox-transform) λ => y^(1/6)
- TEAM_BATTING_1B (derived variable) => TEAM_BATTING_H - TEAM_BATTING_2B - TEAM_BATTING_3B - TEAM_BATTING_HR, (boxcox-transform) λ => 1/(y^2)
- SLUGGING (derived variable) => 2 * TEAM_BATTING_3B + TEAM_BATTING_HR y^(3/5)
- TEAM_BATTING_SB (boxcox-transform) λ => y^(-1/25)

- FIELDING YIELD (derived variable) => TEAM_FIELDING_E + TEAM_FIELDING_DP y^(-(9/10))
- TEAM_PITCHING_SO (boxcox-transform) λ (y^2/3)

SLUGGING & FIELDING YIELD predictors where derived because the underlying variables lacked enough of a normal distribution and residual uniformity to be added to the multi-linear model on their own. These problems were greatly reduced after applying the derivation and Box-Cox transformation.

The Forward Selection method was applied through the use of p-values and variance inflation factors (VIF) against the following predictor variables:

Steps were as follows:

- Start with p.Hits & p.Walks
- Added b.Singles & b.Slugging
- Added Stolen Bases and Fielding
- Removed p.Hits
- Added b.Walks and p.Strikeouts
- Remove b.Walks

Six iterations of p-value / VIF bacward selection removed TEAM_PITCHING_H, TEAM_PITCHING_SO and TEAM_PITCHING_BB from the model. All other variables remained statistically significant with no signficant collinearity. However, evidence of multiple outliers was found via R's summary diagnostic plots. Those outliers were removed via a series of additional iterations.

The diagnostic plots of that model show relatively good linearity between the response variable TARGET_WINS and the 6 predictor variable as evidenced in the Added Variable plots shown in the Appendix. Furthermore, the plots of standardized residuals against each of the predictor variables showed evidence of relatively uniform variability for each variable except the derived predictor FIELDING which has two clusters.

Note that by combining TEAM_FIELDING_DP and TEAM_FIELDING_E into FIELDING using a Box-Cox transform λ => 1/(y^2) on the derived field prior to adding FIELDING to the model, variablity was somewhat improved and the clustering effect was somewhat mitigated.

Pre-model transformation of individual predicotor variables were then adjusted as follows:

- TEAM_PITCHING_BB
- TEAM_BATTING_1B (derived variable) => TEAM_BATTING_H - TEAM_BATTING_2B - TEAM_BATTING_3B - TEAM_BATTING_HR
- SLUGGING (derived variable)
- TEAM_BATTING_SB
- FIELDING YIELD (derived variable) => TEAM_FIELDING_E + TEAM_FIELDING_DP
- TEAM_PITCHING_SO

Coefficient	Variable
78.730	Intercept
+ 28.760	1/(TEAM_PITCHING_BB^6)
- 21.360	1/(TEAM_BATTING_1B^2)
+ 1.450	(SLUGGING^3) / (SLUGGING^5)
- 125.800	1/(TEAM_BATTING_SB^25)
+ 3.747e6	(FIELDING^10) / (FIELDING^9)
- .1037	(TEAM_PITCHING_SO^2) / (TEAM_PITCHING_SO^3)

RSE	R^2	Adj. R^2	F Stat.	MSE
11.97	0.3142	0.3123	164.2	143

The coefficients for this model are counterintuitive largely due to the transformations applied to each predictor variable. For example, one would think TEAM_BATTING_1B would nominally correlate postively with winning, yet its coefficent is strongly negative. The same is true of TEAM_BATTING_SB and FIELDING.

The sign and magnitude of each coefficient is the result of the transformations as well as the linear model, so it appears that reading the coeffiecients becomes less intuitively tied to the linear relationship between predictors and variables as the extent of transformations increase. This does not invalidate the model, it simply means the coefficents become less useful as a check of the model's fidelity.

Part 4. Select Models

The chart below summarizes the model statistics for all five of our models. The models are listed from left to right in accordance with the order in which they were described in **Part 3** herein (Models 1-5). Please refer to the Appendix for detailed descriptions, diagnostic plots, and procedural details on the derivation of each of these models.

Metric	General Model	Total Bases	TB PLUS	Sabermetrics	Box-Cox First
RSE	11.86	11.97	12.12	11.99	11.97
R^2	0.3168	0.3048	0.2994	0.3229	0.3142
Adj. R^2	0.3143	0.3029	0.2931	0.3207	0.3123
F Stat.	124.8	157.5	224.3	147.4	164.2
MSE	141	143	147	144	143

Each of our five models converged on similar R^2 values, RSE's, and MSE's, and all yielded residuals that were distributed normally without signficant evidence of highly leveraged outliers. No signficant collinearity exists within any of the five models for any of their component predictor variables.

Of the five, the General Model had the least favorable residual characteristics, with multiple predictors showing non-constant variability relative to the residuals, while having a relatively low F-statistic when compared to the other models.

Of the remaining four models, the Total Bases model was a great improvement over the General Model. However, it too displayed some lack of constant variability of residuals relative to a couple of the predictor variables.

The Box-Cox First model makes use of a recommended Box-Cox transform for each individual predictor variable before the linear model is constructed via forward selection. While this model yielded results similar to the others, it was deemed to be overly complex both for the number of variables that comprised the model and the difficulty we would have explaining the predictor coefficients.

Of the remaining two (TB PLUS and Sabermetrics), while the Sabermetrics model yields a slightly large R^2 value, the TB Plus model is simpler in that it makes use of only 4 predictor variables while possessing a much larger F-statistic. Such a large difference in F-statistics indicates that the TB Plus model is explaining more of the variability of the training data than is the Sabermetrics model.

Therefore, **we selected the TB PLUS model** as the basis for our prediction of TARGET_WINS for the Evaluation data set. To ensure the model's efficacy when applied to the Evaluation data set, we first applied a set of transformations to that data set that were identical to those applied to the Training data set during the development of the TB PLUS model. The TB PLUS model was then applied to yield a set of INDEX / TARGET_WINS pairs.

Since displaying the full set of predicted values would consume a large number of pages herein, a sample of the first 10 rows of that data set is displayed below:

INDEX	TARGET_WINS
9	61
10	66
14	72
47	86
60	66
63	73
74	82
83	71
98	69
120	74

The full set of predicted TARGET_WINS can be found at the following web link:

https://github.com/spsstudent15/2016-02-621-W1/blob/master/HW1-PRED-EVAL-WINS-ONLY.csv

---

Part 5. References

Bibliography

Diez, D.M., Barr, C.D., & Cetinkaya-Rundel, M. (2015). OpenIntro Statistics, Third Edition. Open Source. Print

Faraway, J. J. (2015). Extending linear models with R, Second Edition. Boca Raton, Fla: Chapman & Hall/CRC. Print

Faraway, J. J. (2015). Linear models with R, Second Edition. Boca Raton, Fla: Chapman & Hall/CRC. Print

Fox, John (2016). Applied Regression Analysis and Generalized Linear Models, Third Edition. Los Angeles: Sage. Print.

Sheather, Simon J. (2009). A Modern Approach to Regression with R. New York, NY: Springer. Print

Resource Links

http://www.baseball-almanac.com/

https://raw.githubusercontent.com/spsstudent15/2016-02-621-W1

http://tangotiger.net/wiki_archive/Base_Runs.html