Data 621 Homework 1: Moneyball

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

---

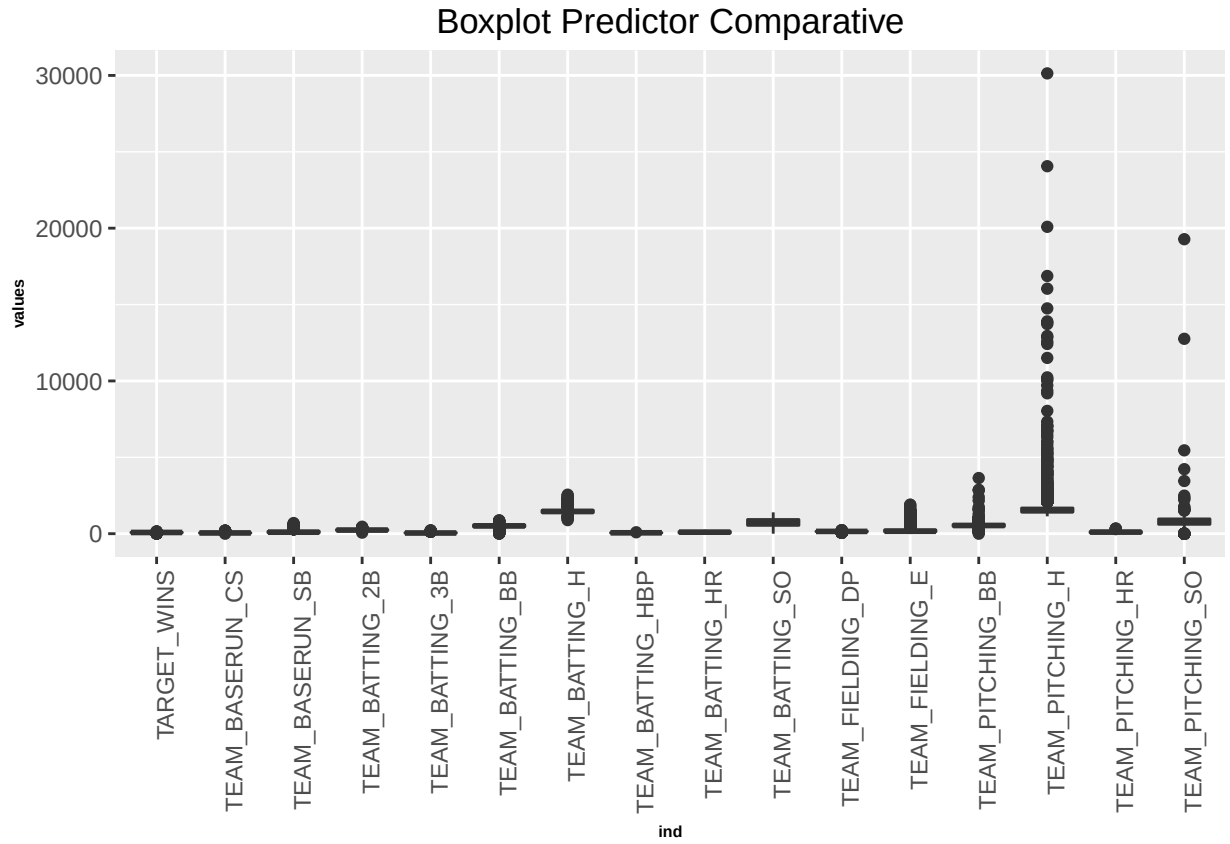
Part 1: Data Exploration

Data Summary

The original Training data set is comprised of 17 elements and 2276 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
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

At first glance this chart shows that there are missing values in 6 variables (particularly TEAM_BATTING_HBP and TEAM_BASERUN_CS). In addition, several values such as TEAM_PITCHING_H, TEAM_PITCHING_BB and TEAM_PITCHING_SO struggle with skew and kurtosis. The box plot summary shown below indicates the presence of some significant outliers in several data columns, particularly in TEAM_PITCHING_H and TEAM_PITCHING_SO.



Correlation Between Predictors and TARGET_WINS

Using the cor function across the data set we noticed some strong correlations between predictor variables. For example, TEAM_BATTING_H showed colinearity with TEAM_BATTING_2B, TEAM_BATTING_3B and TEAM_BATTING_HR since these individual variables are each a subset of hits. TEAM_BATTING_BB and TEAM_PITCHING_BB have strong correlation, as do TEAM_PITCHING_HR and TEAM_BATTING_HR.

The following table shows the correlation of each potential predictor with the TARGET_WINS response variable when the NA's are omitted:

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

Value	Correlation with Wins
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

Conclusion of Data Exploration

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

Part 2: Data Preparation

Our data preparation efforts for the training data set include the creation of one new derived variable, removal of 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.

Step 1: New Variable Creation

We began our data preparation efforts by creating a new variable TEAM_BATTING_1B which represents offensive single base hits (created by subtracting out the TEAM_BATTING doubles, triples and home runs from the TEAM_BATTING_H variable). We believe that separating out singles from the other unique hit values will help to minimize collinearity. The TEAM_BATTING_H variable is then removed from the data set since it is simply a linear combination of its component variables.

Step 2: Removal of Variables Due to Collinearity and/or Missing Values

The results of our data exploration efforts lead 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 contains 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 choose 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.

Step 3: Imputation of Missing Values (NA's)

Filling Missing Values in the Training Data Set

After removing these values, our next step is to impute the remaining missing data using a linear regression approach recommended by Faraway (p.201) and Fox (p.611). We are not 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.

While creating our regression models we make use Variance Inflation Factors (VIF) to verify that there are no collinearity issues and use backward selection to ensure that all p-values are $< .05$. Each model produces imputed distributions of the subject variables that are 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 data are an improvement over simply filling the NA's with a mean or median. Furthermore, the replacement of the NA's with numerical values allow us to run our final models on all records, not just those without NA's. For consistency we make use of the same approach with the evaluation data.

Imputation regression models were created for the following variables:

1. **TEAM_BATTING_SO**: The adjusted R^2 value for this regression model is 0.7223 and yields a distribution matching the variables prior to the NA's replacement. For this predictor we impute a total of 102 missing values via regression.
2. **TEAM_PITCHING_SO**: The adjusted R^2 value for this regression model is 0.9952. We impute 102 missing values for pitching strikeouts.
3. **TEAM_BASERUN_SB**: The adjusted R^2 value for this regression model is 0.3427. Despite the adjusted R^2 being low relative to the models described above, the model yields a distribution matching that of the variable beforehand. Our model replaces 131 missing stolen base values.
4. **TEAM_FIELDING_DP**: The adjusted R^2 for this model is 0.3904. We impute 286 missing values with a similar distribution to the previous data.

Step 4: Removal of Extreme Outliers From the Training Data Set

Our final data preparation step is to eliminate some clearly egregious outliers identified via research through *baseball-almanac.com*, as 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 are aberrations.

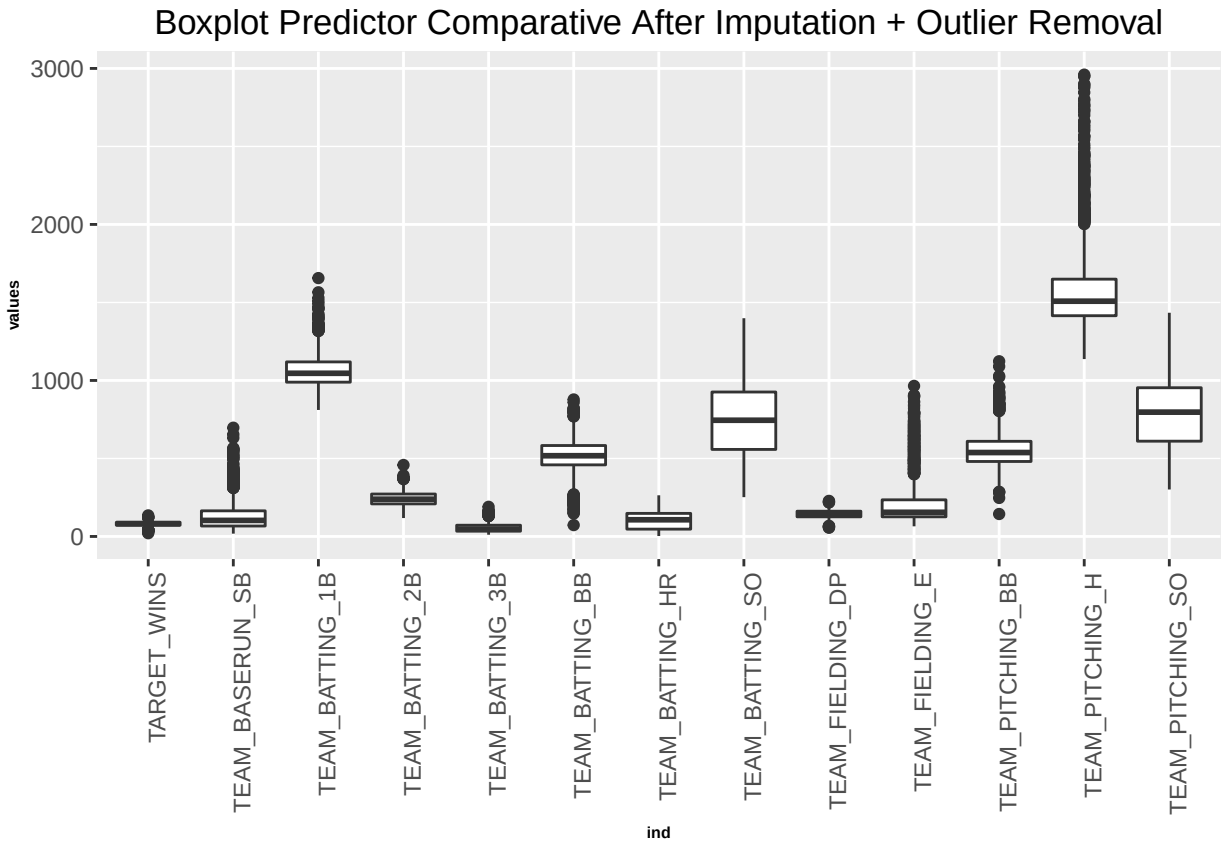
Similarly, the most errors by team in a single season are 639 by Philadelphia in 1883. Prorating to 162 games we calculate that we should discard any records containing TEAM_FIELDING_E values above 1046.

The TEAM_PITCHING_H variable also appears to have numerous egregious outliers. For example, the most offensive hits by a team in a single season are 1730. As such, it is highly unlikely that any pitching staff would surrender more than 3000 hits in a single season. Such a total would indicate that the team allows more than 18 hits per game. As such, any records having a TEAM_PITCHING_H value > 3000 are removed from the data set.

As result of this research, we feel confident in removing 104 records with eggregious outliers that are impossible from a historical perspective. Using this SME knowledge will help to normalize our data and improve the expected performance of our linear models.

Results of Imputation for Missing Values and Outlier Removal Process

The charts below show that our data transformation process is dramatically improving the data, There are still a few outliers, but on a dramatically smaller scale, with a particularly significant change for the TEAM_PITCHING_H and TEAM_PITCHING_SO variables.



In addition, the chart below shows how we've improved the skew and kurtosis relative to the original data set.

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	

variables	n	mean	sd	med	min	max	range	skew	kurtosis	se	NAs
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	

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

Step 5: Other Data Preparation Transformations: Refer to Model Descriptions

We did use other *model-specific* data transformations, including 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

Approach:

Our first 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 yields a model with all p-values less than .05. However, VIF analysis shows evidence of multiple collinear variables within the model. Subsequent removals of *TEAM_PITCHING_SO* and *TEAM_PITCHING_BB* due to collinearity yields a model calling for the removal of *TEAM_BATTING_2B* on the basis of its p-value.

The final model of these iterations shows clear evidence of a number of outliers as evidenced in R's summary diagnostic plots. Removal of these outliers via a series of additional modeling iterations yields the following final model (which once again includes *TEAM_BATTING_2B* since removal of 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

Coefficient	Variable
- 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

Additional Iterations:

However, the diagnostic plots of that model show a lack of linearity between the response variable TARGET_WINS and the predictor variable TEAM_FIELDING_E. Furthermore, the plots of standardized residuals against each of the predictor variables show evidence of non-constant variability for variables such as TEAM_BATTING_HR, TEAM_BATTING_SO, TEAM_BASERUN_SB, and TEAM_FIELDING_E. Therefore, we transform The TEAM_FIELDING_E variable using a Box-Cox recommended power transform of (-1), or (1/y) and rebuild the model. 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 show an improvement in the linearity of the model.

The characteristic equation for this improved model is as follows:

Coefficient	Variable
52.88	Intercept
+ 0.168	TEAM_BATTING_3B
+ 0.096	TEAM_BATTING_HR
+ 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

Conclusions:

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 surprises since baseball fans believe that more defensive double plays will improve a team's chances of winning games. However, the variable itself is *negatively* correlated with TARGET_WINS (see the **Data Exploration** section), which validates the negative coefficient. Similarly, the coefficient for TEAM_PITCHING_H is also counterintuitive, but the variable is actually positively correlated with TARGET_WINS. Finally, TEAM_FIELDING_E changes from negative in the earlier model to positive here. However, 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 earlier iterations, we still see 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 our next models we attempt to address some of the skew issues by creating linear combinations of various variables.

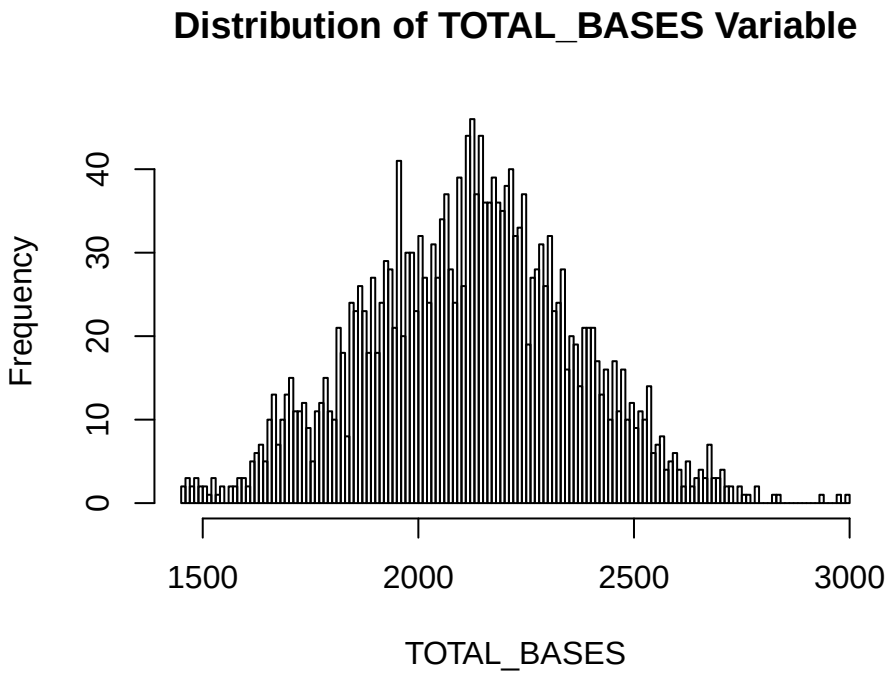
Model 2: Total Bases

Approach:

This model employs 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 remove TEAM_PITCHING_SO and TEAM_PITCHING_BB from the model. All other variables remain statistically significant with no signficant collinearity. However, evidence of multiple outliers are found through R's summary diagnostic plots. Removing those outliers via several additional iterations resulting in the following model:

Coefficient	Variable
48.486	Intercept

Coefficient	Variable
+ 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

Transformation Iterations

Once again, the diagnostic plots of that model show a lack of linearity between the response variable TARGET_WINS and one of the predictor variables (TEAM_FIELDING_E). Furthermore, the plots of standardized residuals against each of the predictor variables show evidence of non-constant variability for variables such as TEAM_BATTING_SO, TEAM_BASERUN_SB, and TEAM_FIELDING_E.

A Box-Cox recommended power transform of (-1), or (1/y) was then applied to TEAM_FIELDING_E and the model was rebuilt. The resulting Added Variable plots show that all predictors are linearly related to the response, and the variability of the residuals improve. 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 and the 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

Conclusions:

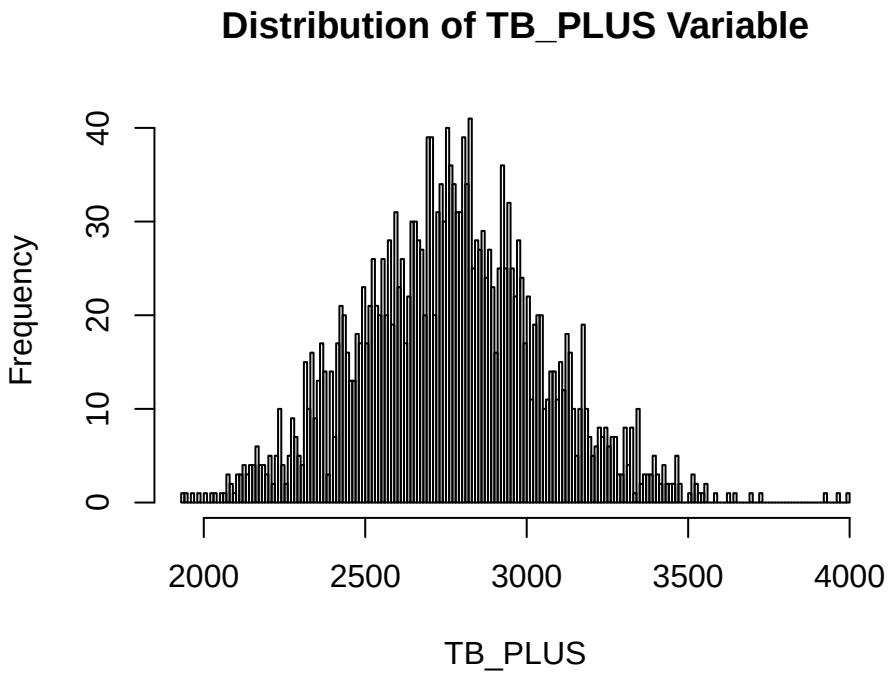
Like the first model,the coefficients for TEAM_BATTING_BB, TEAM_BATTING_SO, TEAM_BASERUN_SB, and TOTAL_BASES all make sense intuitively. The TEAM_FIELDING_DP coefficient is negative, matching its correlation with TARGET_WINS. However, the coefficent for TEAM_FIELDING_E changes from negative in the earlier model to positive here, as it now applies to the *transformed reciprocal of the variable.

Model 3: Total Bases PLUS

Our third model improves upon 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 TOTAL_BASES comes from the fact that both (as with the component variables of TOTAL_BASES) represent basepath advancements by a team's offense. "Total Bases Plus" (TB_PLUS) is calculated using TEAM_BATTING and TEAM_BASERUN variables as follows:

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

Including 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 from model 2 appears to be nearly normally distributed, thereby negating any skew issues evident in its component variables. A histogram of the distribution of the derived TB_PLUS variable is shown below:



This model also 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 remove TEAM_PITCHING_H, TEAM_PITCHING_SO and TEAM_PITCHING_BB from the model. All other variables remain statistically significant with no signficant collinearity. However, once again we found evidence of multiple outliers via R's summary diagnostic plots, and removal of those outliers via a series of additional iterations yields 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

Transformation Iterations

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. We see the same negative trend with TEAM_FIELDING_DP as in models 1 and 2. Most importantly, the coefficient for TB_PLUS positively correlates with the response variable as expected. As with the first two models, the diagnostic plots for this approach unfortunately show a lack of linearity between the response variable TARGET_WINS and the predictor variable TEAM_FIELDING_E. Furthermore, the plots of standardized residuals against each of the predictor variables demonstrate evidence of non-constant variability for the variables TEAM_BATTING_SO and TEAM_FIELDING_E.

As in model 2, we transform the TEAM_FIELDING_E predictor using its reciprocal. 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. 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

Conclusion:

This model is an improvement over the first TB_PLUS model when the residual plots are considered, and the number of predictor variables used is two fewer than that of the "Total Bases" model. As in model 2, the coefficient for TEAM_FIELDING_E changes from negative to positive, and once again 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 4: Sabermetrics Model

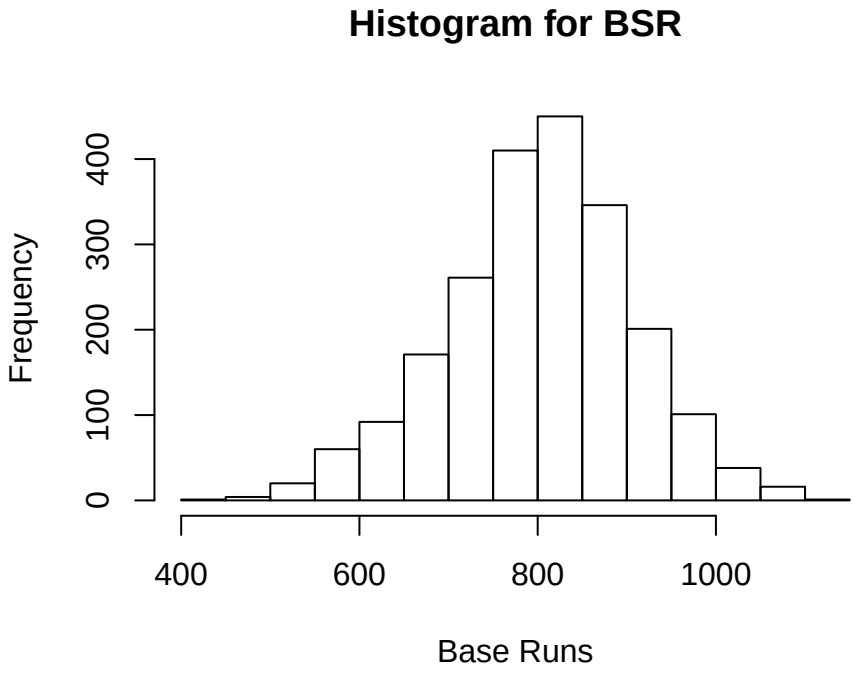
Approach

Sabermetrics has become the rage in baseball, popularized by Billy Beane and the data set we are exploring. As a result, we built a model that centers around one of these advanced analytics known as BSR or base runs. This statistic (designed by David Smyth in the 1990's) estimates the amount of runs a team *SHOULD* score, adding an intriguing element to a data set which does not include runs (see http://tangotiger.net/wiki_archive/Base_Runs.html for more information). The formula For BSR 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 eliminate the value of TEAM_BATTING_H, we sum up singles, doubles, triples and home runs as in the approach used for the TOTAL_BASES model described above. The data for BSR exhibit a fairly normal distribution.



Since BSR is a combination of all of the batting variables, we eliminate them from the regression resulting in a very strong model on the first iteration. All p-values are very low, and the variation values 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

RSE	R^2	Adj. R^2	F Stat.	MSE

Conclusion:

The coefficients for this model largely 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 as mentioned above it matches the variable's correlation with TARGET_WINS). The positive impact of allowing pitching hits is puzzling, but once again it agrees with the trends we see in the original Training data set.

Model 5: Box-Cox First

Approach:

This model first applies transformations to a simple regression of each predictor against wins, resulting in improvements in both the normality of the distributions of the predictors and the distribution of residuals relative to each predictor variable. In each case, we apply a Box-Cox transformation to improve the skew of predictor variables BEFORE creating the multi-regression model. The model then applies 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. We also derive SLUGGING and FIELDING_YIELD variables by combining some of the variables.

Pre-model transformation of individual predicotor variables are 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)

Multiple iterations backwards and forwards based on p-value and vif values result in a model whose diagnostic plots show relatively good linearity between the response variable TARGET_WINS and the 6 predictor variable. Furthermore, the plots of standardized residuals against each of the predictor variables show evidence of relatively uniform variability for each variable except the derived predictor FIELDING which has two clusters. The model equation is as follows:

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

Conclusion:

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 correlate positively 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 are the result of the transformations as well as the linear model. 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

Step 1: Compare Key statistics

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**.

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

Step 2: Pick the top two

Of the five, we eliminate the General Model as it has the least favorable residual characteristics, with multiple predictors showing non-constant variability relative to the residuals. It also creates 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 displays some lack of constant variability of residuals relative to a couple of the predictor variables, so we can eliminate that one as well.

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 yields results similar to the others, we believe it is overly complex due to both the number of variables that comprise the model and the difficulty we have in explaining the predictor coefficients.

Step 3: Pick the "best" model

Of the remaining two models, the Sabermetrics model yields a slightly larger R^2 value showing a slightly better possible fit. However, the TB Plus model is simpler as it make as it uses 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 select the TB PLUS model** as the basis for our prediction of TARGET_WINS for the Evaluation data set.

Step 4: Apply to the evaluation data

To ensure the model's efficacy when applied to the Evaluation data set, we apply the same set of transformations used on the Training data set prior to building our individual models. The results of those transformations (which include filling in the missing values) can be found here:

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

Then, the additional transformations required for the TB PLUS model are applied to ensure conformity between the Training and Evaluation data sets relative to the structure of the model.

The TB PLUS model is 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, a sample of the first 10 rows of that data set is displayed as an example:

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 and is also presented at the beginning of the Appendix:

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

Summary statistics for TARGET_WINS for the evaluation data, the "prepped" training data set, and the original training data set are shown below (NOTE: "n" represents the number of records within a data set).

Data Set	n	Mean	sd	Med.	Min	Max	Range	Skew	Kurtosis	SE
Evaluation	259	81	9	81	50	104	54	-0.05	0.35	0.55
Training	2276	81	16	82	0	146	146	-0.40	1.03	0.33
"Prepped"	2172	81	15	82	21	135	114	-0.22	0.08	0.31

Using the Model for Inference

The standard error metrics derived via the model can be used to compute confidence intervals (CIs), prediction intervals (PIs), and perform hypothesis tests on the coefficients. However, such computations are beyond the scope of this assignment.

Conclusion:

The predicted wins we derive for the Evaluation data appear to make sense. They have a mean and median of 81 and show little skew. The predictions range between 50 and 104, reasonable but not as varied as the ones in both the original training data and the "Prepped" data we used for building our models.

Our selected model appears to performs well relative to the training data set we were provided. Though the training data set presented a variety of challenges, including signifcant outliers and missing data, we were able to address those challenges through the creation of a new derived variable (TEAM_BATTING_1B, or singles), the removal of three data fields (TEAM_HBP, TEAM_BASERUN_CS, and TEAM_PITCHING_HR), the removal of 104 records due to extreme outliers, and the imputation of the remaining missing values via linear regression models.

After developing 5 distinct models using backward iterations, forward iterations, linear combinations of variables, and Box-Cox transformations we selected what we believe to be our best model. However, the models we constructed would likely be greatly improved if other commonly used baseball statistics were available, such as runs scored, at bats, etc. Given the lack of such data within the training data set, we believe we've put forth a credible predictive model that can be used when such common baseball statistics aren't available.

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, FL: Chapman & Hall/CRC. Print

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

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

Appendix

The Appendix to this document containing all of the R code used for this assignment (as well as other relevant output) can be found in a separate PDF file accessible via this web link:

https://github.com/spsstudent15/2016-02-621-W1/blob/master/HW1_Appendix.pdf