Unit 01 Homework – Moneyball

KAGGLE name: *NikhilAgarwal*

Nikhil Agarwal
Northwestern University
PREDICT 411, Section 55

I am requesting a total of **30** bingo bonus points for the unit 1 homework. Please see my justification below.

Points Requested	**Category**	**Justification**
5	Decision Tree	I used R to construct a decision tree. The results for one variable were implemented into the submitted SAS code
5	Macro	I used a small amount of macro coding in my SAS code
20	R code	I constructed quite a bit of this homework in R. Please see separate R code (NikhilAgarwal_Unit1HW_Rcode.txt)

Table of Contents

INTRODUCTION

The intent of this assignment is to ultimately develop an ordinary least squares (OLS) function that can be used to predict wins for a baseball team given a certain set of variables. Information from over 2000 teams, spanning from1871 through 2006, was used to help construct a model. Prior to developing a single model, multiple regression models were explored (single and multivariate) using primarily the stepwise function. Various selection parameters (e.g., SBC, AIC, Adjusted R-Squared, etc.) were used to determine the best model.

RESULTS

Data Exploration

The original dataset contains 15 distinct variables (outlined in Table 1). These variables can be considered potential predictors. Note that the response variable will be TARGET_WINS. Of particular note is the “Theoretical Effect” column. This column succinctly describes whether the potential predictor variable has a positive or negative impact on wins (i.e., the response variable TARGET_WINS).

Table 1: Brief description of default variables

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

Continuing on the data exploration journey, it is also wise to identify if any of the variables are missing. Figure 1 illustrates some basic statistics on the dataset.

Variable	N	N Miss	Minimum	Maximum	Median	Mean	Std Dev
INDEX	2276	0	1.0000000	2535.00	1270.50	1268.46	736.3490405
TARGET_WINS	2276	0	0	146.0000000	82.0000000	80.7908612	15.7521525
TEAM_BATTING_H	2276	0	891.0000000	2554.00	1454.00	1469.27	144.5911954
TEAM_BATTING_2B	2276	0	69.0000000	458.0000000	238.0000000	241.2469244	46.8014146
TEAM_BATTING_3B	2276	0	0	223.0000000	47.0000000	55.2500000	27.9385570
TEAM_BATTING_HR	2276	0	0	264.0000000	102.0000000	99.6120387	60.5468720
TEAM_BATTING_BB	2276	0	0	878.0000000	512.0000000	501.5588752	122.6708615
TEAM_BATTING_SO	2174	102	0	1399.00	750.0000000	735.6053358	248.5264177
TEAM_BASERUN_SB	2145	131	0	697.0000000	101.0000000	124.7617716	87.7911660
TEAM_BASERUN_CS	1504	772	0	201.0000000	49.0000000	52.8038564	22.9563376
TEAM_BATTING_HBP	191	2085	29.0000000	95.0000000	58.0000000	59.3560209	12.9671225
TEAM_PITCHING_H	2276	0	1137.00	30132.00	1518.00	1779.21	1406.84
TEAM_PITCHING_HR	2276	0	0	343.0000000	107.0000000	105.6985940	61.2987469
TEAM_PITCHING_BB	2276	0	0	3645.00	536.5000000	553.0079086	166.3573617
TEAM_PITCHING_SO	2174	102	0	19278.00	813.5000000	817.7304508	553.0850315
TEAM_FIELDING_E	2276	0	65.0000000	1898.00	159.0000000	246.4806678	227.7709724
TEAM_FIELDING_DP	1990	286	52.0000000	228.0000000	149.0000000	146.3879397	26.2263853

Figure 1: Descriptive Statistics on default variables

Note how six of the potential predictor variables have missing values. On a side note, the variable INDEX is simply a unique identifier that will not be used for any modeling purpose. The variable TEAM_BATTING_HBP is missing over 2000 values. Recall that this dataset only contains 2276 observations. Therefore, the variable TEAM_BATTING_HBP will be dropped from the analysis as there are too many missing values and imputing would most likely lead to false assumptions and conclusions.

Table 2 highlights the correlation of each predictor variable to the response variable of TARGET_WINS. The closer the value is to +1 or -1, the stronger the linear relationship. Unsurprisingly, no single predictor variable is highly correlated with TARGET_WINS. Recall from Table 1 that certain predictor variables have a negative impact on the overall wins. However, in Table 2, it is evident that the variable TEAM_BASERUN_CS has a very slight positive relationship with TARGET_WINS. A similar surprise can be found with the variables TEAM_PITCHING_BB and TEAM_PITCHING_HR – which both have a positive correlation, yet they have a negative impact on wins. In other words, the higher the values for these variables, the 'higher' the wins.

Table 2 : Correlation table of default variables to TARGET_WINS

Variable	Correlation
TEAM_BATTING_H	0.38877
TEAM_BATTING_2B	0.2891
TEAM_BATTING_3B	0.14261
TEAM_BATTING_HR	0.17615
TEAM_BATTING_BB	0.23256
TEAM_BATTING_HBP	0.0735
TEAM_BATTING_SO	-0.03175
TEAM_BASERUN_SB	0.13514
TEAM_BASERUN_CS	0.0224
TEAM_FIELDING_E	-0.17648
TEAM_FIELDING_DP	-0.03485
TEAM_PITCHING_BB	0.12417
TEAM_PITCHING_H	-0.10994
TEAM_PITCHING_HR	0.18901
TEAM_PITCHING_SO	-0.07844

Another key check is to determine if there are outliers. All 15 predictor variables (along with the response variable, TARGET_WINS) were checked for outliers. Figure 2 is an example of a histogram and boxplot for the variable TEAM_BATTING_H. Note the many circles that are outside the whiskers in the boxplot. This is a strong indicator that outliers may be present.

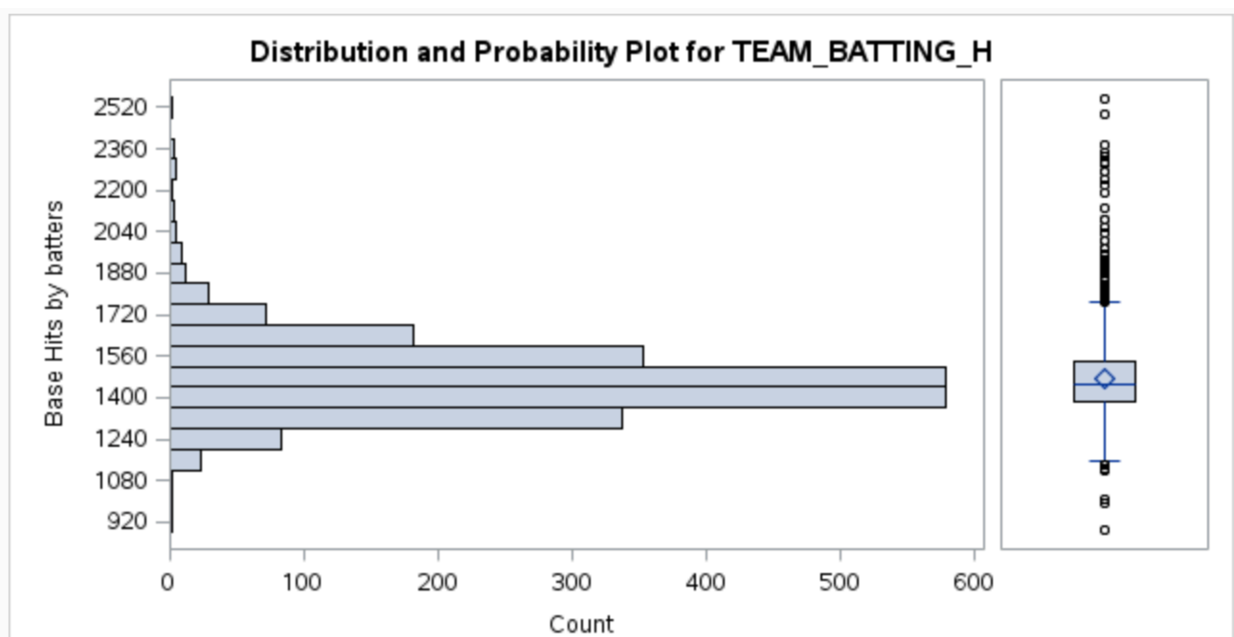

Figure 2: Example histogram & boxplot for variable TEAM_BATTING_H

It was found that almost all of the variables had outliers. Therefore, value caps will be deployed on each variable to ensure that outliers do not skew the model unnecessarily. The caps are explained in greater detail in the section, "Data Preparation".

Data Preparation

Imputation

Figure 1 shows that there are six predictor variables with missing values. Recall that the variable TEAM_BATTING_HBP will be dropped as it is missing a significant amount of information. Therefore, five of variables require imputation. Table 3 summarizes the imputation method used for each of the four variables. All five of these variables were removed from the modelling process and the imputed values were stored in different variables starting with the prefix "imp_". If there was a non-null value for any of the variables in Table 3, then that value was used in lieu of an imputed value. Furthermore, a flag variable was created (with the prefix "m_") to indicate if an imputed value was entered (indicated with a 1 meaning true) or if the original value was used (indicated with a 0 meaning false).

Table 3: Imputation Methods for Variables with Missing Data

Variable	Method of Imputation	Imputed Variable Name
TEAM_PITCHING_SO	Decision Tree	imp_TEAM_PITCHING_SO
TEAM_FIELDING_DP	Mean	imp_TEAM_FIELDING_DP
TEAM_BASERUN_CS	Median	imp_TEAM_BASERUN_CS
TEAM_BASERUN_SB	Median	imp_TEAM_BASERUN_SB
TEAM_BATTING_SO	Median	imp_TEAM_BATTING_SO

Using R, a simple decision tree was constructed for the variable TEAM_PITCHING_SO (see Figure 3).

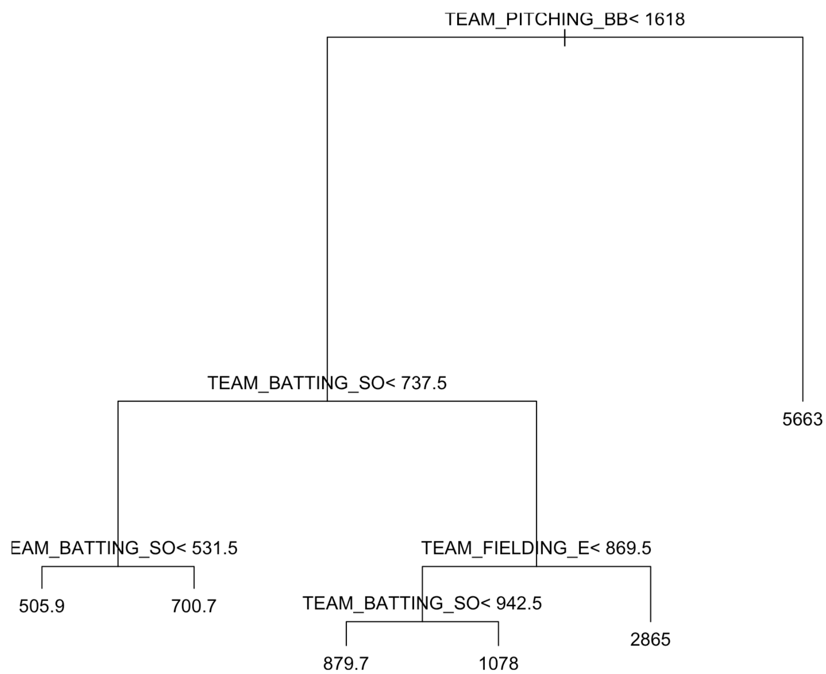

Figure 3: Decision tree for variable TEAM_PITCHING_SO

The following logic briefly describes (converting logic syntax to English) the decision tree for the variable TEAM_PITCHING_SO:

```
If TEAM_PITCHING_BB is greater than 1618 then TEAM_PITCHING_SO will be 5663. If that is not true, then if TEAM_BATTING_SO is less than 737.5 then the imputed value of TEAM_PITCHING_SO will be 609. If that is also not true, then if TEAM_FIELDING_E is less than 869 then the imputed TEAM_PITCHING_SO value is 970. Otherwise, the imputed TEAM_PITCHING_SO value will be 2865.
```

The mean value found amongst non-null data for TEAM_FIELDING_DP was used as the imputed value if there was a missing value. For this variable, it was discovered that the maximum value was not significantly higher than the 99th percentile value (228 vs. 204, respectively). Therefore, the average was deemed appropriate. However, the same cannot necessarily be said for the other three variables (TEAM_BASERUN_CS, TEAM_BASERUN_SB, and TEAM_BATTING_SO). Since the imputation process is taking place prior to the reduction of outliers, it is appropriate to use the 50th percentile data point (median) as the default value for any missing value for these variables.

Outliers

In order to reduce the effects of outliers, all of the 14 predictor variables' 1 and 99th percentiles were identified as the value thresholds (i.e., lower and upper limits). The intent is not to necessarily eliminate all outliers, but to reduce their effect on the overall model. These percentiles still allow for some 'peak' and 'valley' points that illustrate the teams' abilities to be below or above expectations (in terms of wins). Table 4 highlights the lower (1 percentile) and the upper (99th percentile) thresholds used for each variable.

Table 4: Lower and Upper Thresholds for variables

Variable	Lower Threshold	Upper Threshold
TEAM_BATTING_H	1188	1950
TEAM_BATTING_2B	141	352
TEAM_BATTING_3B	17	134
TEAM_BATTING_HR	4	235
TEAM_BATTING_BB	79	755
imp_TEAM_BATTING_SO	67	1200
imp_TEAM_BASERUN_SB	23	439
imp_TEAM_BASERUN_CS	16	143
TEAM_PITCHING_H	1244	7093
TEAM_PITCHING_HR	8	244
TEAM_PITCHING_BB	237	924
imp_TEAM_PITCHING_SO	205	1474
TEAM_FIELDING_E	86	1237
imp_TEAM_FIELDING_DP	79	204

For the actual simulated results (part of the scoring process), the wins were also capped using the 1 percentile and 99th percentile values. For simplicity's sake, the minimum number of wins in the scoring process was limited to 40 wins and the maximum wins was limited to 114.

Data transformations

The original dataset values have been adjusted to match the performance of 162 games. In the early years of baseball, 162 games were not necessarily played by each team in each season. Therefore, each of the variables outlined in table YY were divided by 162 to obtain a per-game assessment. These values were stored in new variable names using a similar nomenclature found in table YY followed by "_pg".

In an effort to minimize potential variation and skewness in the data, both log and square root transformations were applied and new variables were created. All new variables that used log transformations are accompanied with the prefix "log_" and new variables that used square root transformations are accompanied with the prefix "sqrt_".

Another new variable constructed is known as SB_PCT. This variable simply describes the stolen base percentage as a ratio of stolen bases over the total of stolen bases and caught stealing. It was calculated using the following equation:

Equation 1: SB_PCT Calculation

$$SB\_PCT = \frac{imp\_TEAM\_BASE\_RUN\_SB}{(imp\_TEAM\_BASE\_RUN\_SB\ + imp\_TEAM\_BASERUN\_CS)}$$

Model Development

Eight different regression models were constructed, but for the sake of succinctness, only three of them will be discussed in this report. Furthermore, all of the models were constructed using the stepwise selection criteria. This approach ensures that all of the predictor variables are statistically significant and has the ability to remove variables once they enter the model and fail to be statistically significant after newer variables are introduced.

As a primer, all of the models constructed have the following equation format:

$$TARGET\_WINS = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

where β_0 is the intercept, β_n is the parameter estimate of the variable number *n*, X_n is the observed value in the dataset for the variable number *n*, and ε is the standard error term.

Model 1

This first model was designed to explore 19 predictor variables (including the imputation flag variables) with the response variable being TARGET_WINS. Table 5 illustrates a summary of the variables chosen, their coefficients, and the corresponding VIF values. Note how this model only uses 12 of the 19 predictor variables.

Table 5: Model 1 Parameter Estimates & VIF

Variable	Variable Code	Coefficient Notation	Coefficient Value	VIF
Intercept		β0	21.05224	0
TEAM_BATTING_H_pg	X1	β1	6.10805	2.56182
TEAM_BATTING_3B_pg	X2	β2	18.34527	3.10942
TEAM_BATTING_HR_pg	X3	β3	12.14337	4.95189
TEAM_BATTING_BB_pg	X4	β4	6.78694	18.9289
TEAM_PITCHING_BB_pg	X5	β5	-4.18658	11.94574
TEAM_FIELDING_E_pg	X6	β6	-4.15511	5.40092
TEAM_PITCHING_H_pg	X7	β7	0.48377	6.53825
m_TEAM_PITCHING_SO	X8	β8	8.17901	1.9066
imp_TEAM_BATTING_SO_pg	X9	β9	-3.73472	23.05112
imp_TEAM_BASERUN_SB_pg	X10	β10	5.0427	1.88297
imp_TEAM_PITCHING_SO_pg	X11	β11	2.66629	14.97049
imp_TEAM_FIELDING_DP_pg	X12	β12	-16.86646	1.51045

This model appears to have several variables with high VIF values. Generally, a VIF value of 5 or higher tends to suggest multicollinearity and VIF values at ten or higher strongly suggest that multicollinearity is present. On another note, it is curious that the coefficient value for the variable imp_TEAM_FIELDING_DP_pg (recall that this variable refers to the double plays made by the fielding team) is negative. This negative coefficient value suggests that this variable has a negative impact on wins. Intuitively, we would expect that as more double plays are made, then the fielding team may end up winning more games. However, this could also be interpreted that just because a fielding team does well (conducting more double plays), this does not necessarily mean that its batting performance is good.

Table 6 highlights some of the goodness of fit diagnostics of this model.

Table 6: Model 1 Goodness of Fit Diagnostics

Characteristic	Value
R-Square	0.29888
Adjusted R-Square	0.29516
RMSE	13.2247
AIC	11766.60
BIC	11768.77
SBC	11841.10

It is important to note that although the adjusted R-square value is quite low, this does not necessarily mean the model is incorrect or insufficient. Furthermore, for this analysis, the only key metric that will be used to determine the most appropriate model will be SBC. This is explained further in section: **Model Selection**.

Finally, this model was adjusted slightly through the removal of the variables: TEAM_PITCHING_BB_pg and imp_TEAM_BATTING_SO_pg. When the model was reran, the goodness of fit diagnostics were similar to what was found in Table 6.

Model 2

For this model, the log transformed predictor variables were used (along with the imputation flags). As seen in Model 1, this model also consists of 19 variables with 1 response variable (TARGET_WINS). Table 7 illustrates a summary of the variables chosen, their coefficients, and the corresponding VIF values. Note how this model also only uses 12 of the 19 predictor variables.

Table 7: Model 2 Parameter Estimates & VIF

Variable	Variable Code	Coefficient Notation	Coefficient Value	VIF
Intercept		β0	-71.17207	0
m_TEAM_PITCHING_SO	X1	β1	3.79362	1.45456
log_TEAM_BATTING_H_pg	X2	β2	71.79477	3.5304
log_TEAM_BATTING_2B_pg	X3	β3	-4.10007	2.58181
log_TEAM_BATTING_3B_pg	X4	β4	6.10749	2.67269
log_TEAM_BATTING_HR_pg	X5	β5	-8.94744	51.72002
log_TEAM_BATTING_BB_pg	X6	β6	5.31283	2.80443
log_TEAM_PITCHING_HR_pg	X7	β7	12.14965	41.76897
log_TEAM_FIELDING_E_pg	X8	β8	-11.03641	5.75686
log_imp_TEAM_BASERUN_SB_pg	X9	β9	4.68466	1.95589
log_imp_TEAM_BASERUN_CS_pg	X10	β10	-2.91443	1.39726
log_imp_TEAM_PITCHING_SO_pg	X11	β11	-1.96853	2.39799
log_imp_TEAM_FIELDING_DP_pg	X12	β12	-15.71114	1.64452

Similar to Model 1, this model appears to have two variables with very high VIF values. Intuitively, it is somewhat clear to see the relationship between log_TEAM_BATTING_HR and log_TEAM_PITCHING_HR. Of particular interest is the coefficient of log_TEAM_BATTING_2B. Since this value is negative, it could be interpreted as saying "for a unit increase in X3, it could be expected that TARGET_WINS would decrease by approximately 4.1%. This is somewhat odd as the more double hits a batting team has, one would think that the team would make more runs. However, from a different perspective, this also could be construed as increasing the probability of the fielding team of tagging a player out.

Table 8 highlights some of the goodness of fit diagnostics of this model. Note the R-Square value of 0.29682. This can be interpreted as roughly 29.6% of the variation found in wins can be explained by the chosen predictor variables.

Table 8: Model 2 Goodness of Fit Diagnostics

Characteristic	Value
R-Square	0.29682
Adjusted R-Square	0.29310
RMSE	13.2440
AIC	11773.27
BIC	11775.41
SBC	11847.76

This model was slightly modified with the removal of the variable log_TEAM_BATTING_HR_pg. This resulted in slightly different goodness of fit diagnostics (see Table 9).

Table 9: Model 2 (revised) Goodness of Fit Diagnostics

Characteristic	Value
R-Square	0.29594
Adjusted R-Square	0.29252
RMSE	13.2495
AIC	11774.14
BIC	11776.27
SBC	11842.90

Note how the SBC value is lower, but both the AIC and BIC values have slightly worsened (as compared to the values seen in Table 8). Furthermore, notice how the R-square value did not change significantly.

Model 3

The objective of any of the models is to project the overall wins. In that spirit, 48 variables (including the imputation flags) were entered into the model development. This consisted of all the "_pg", "log_", and "sqrt_" variables. After running the regression model with the stepwise selection, 26 variables were retained. Table 10 illustrates a summary of the variables chosen, their coefficients, and the corresponding VIF values.

Table 10: Model 3 Parameter Estimates & VIF

Variable	Variable Code	Coefficient Notation	Coefficient Value	VIF
Intercept		β0	4116.5805	0
TEAM_BATTING_H_pg	X1	β1	5.38229	8.05094
TEAM_BATTING_2B_pg	X2	β2	1184.45396	42542
TEAM_BATTING_3B_pg	X3	β3	23.86704	3.47063
TEAM_BATTING_BB_pg	X4	β4	177.71041	2714.76142
TEAM_PITCHING_BB_pg	X5	β5	-150.1981	18019
TEAM_FIELDING_E_pg	X6	β6	-32.50384	1208.32683
TEAM_PITCHING_H_pg	X7	β7	-18.28425	4719.54227
m_TEAM_PITCHING_SO	X8	β8	3.45968	2.76542
imp_TEAM_BATTING_SO_pg	X9	β9	-45.36556	1119.30827
imp_TEAM_BASERUN_SB_pg	X10	β10	7.99394	2.67856
imp_TEAM_PITCHING_SO_pg	X11	β11	-1.74724	32.14662
log_TEAM_BATTING_2B_pg	X12	β12	1783.33331	41027
log_TEAM_BATTING_HR_pg	X13	β13	-5.86404	43.92345
log_TEAM_BATTING_BB_pg	X14	β14	299.27984	2189.74418
log_TEAM_PITCHING_BB_pg	X15	β15	-382.89843	17887
log_TEAM_FIELDING_E_pg	X16	β16	-69.01397	1012.53768
log_TEAM_PITCHING_H_pg	X17	β17	-292.20308	4859.0128
log_imp_TEAM_BATTING_SO_pg	X18	β18	-112.73409	719.89872
log_imp_TEAM_FIELDING_DP_pg	X19	β19	-12.60626	1.78729
sqrt_TEAM_BATTING_2B_pg	X20	β20	-5838.73625	166133
sqrt_TEAM_BATTING_BB_pg	X21	β21	-957.78231	9401.39843
sqrt_TEAM_PITCHING_HR_pg	X22	β22	30.34083	36.99459
sqrt_TEAM_PITCHING_BB_pg	X23	β23	971.79692	71382
sqrt_TEAM_FIELDING_E_pg	X24	β24	170.69434	4162.83487
sqrt_TEAM_PITCHING_H_pg	X25	β25	306.85225	18557
sqrt_imp_TEAM_BATTING_SO_pg	X26	β26	301.42524	3256.29868

Prior to discussing this model in greater detail, Table 11 highlights some of the goodness of fit diagnostics of this model.

Table 11: Model 3 Goodness of Fit Diagnostics

Characteristic	Value
R-Square	0.41201
Adjusted R-Square	0.40521
RMSE	12.1485
AIC	11394.11
BIC	11396.54
SBC	11548.82

This model is of absolute amazement! There are enormous amounts of multicollinearity and the coefficient values for some of the variables are absolutely large. What is also very surprising is the fact this model has a much higher adjusted R-square value and a significantly lower SBC value. It is important to note that the R-squared value simply describes the amount of variation in the response variable explained by the model. Also interesting about this model is the fact that many of the coefficient values for the parameters make sense in terms of the sign (positive or negative). However, two of the variables chosen (sqrt_TEAM_BATTING_2B_pg and sqrt_TEAM_BATTING_BB_pg) have negative values which would seem contrary. Intuitively, both of these variables should help yield more wins, but in this model they lead to a reduction in wins.

Model Selection

In this analysis, three of the eight different models constructed were discussed. Recall that all of the models constructed used a stepwise selection criterion. The key metric that was used for model selection is the SBC (Schwarz Bayesian Criteria). In a paper by Dennis Neal, the metrics AIC, BIC, and SBC were compared for ten independent variables used in a given model with varying sample sizes. The conclusion from the paper: "SBC performed best by correctly selecting the true model the most consistently after enumerating all possible true models from the simulated data. SBC consistently performed best for both sample sizes n = 1000 and n = 100" (Beal, 2007). Although the models discussed in this paper use more than ten variables, the consistency of SBC found by Dennis Neal is used as a guiding light to create a simple 'go/no-go gauge' for model usage.

Table 12 summarizes the SBC values found for each of the three models discussed.

Table 12: SBC Summary for all 3 Models

Model	SBC
Model 1	11841.10
Model 2	11847.76
Model 3	11548.82

Based on the criterion of SBC alone, Model 3 is the 'best' model to deploy. Although eight models were constructed, Model 3 had the lowest SBC value of any of the models constructed. There are a couple caveats that should be noted. First, Model 3 has a significant amount of multicollinearity. This may result in parameter coefficient values that are not necessarily meaningful. However, given Model 3's performance (in terms of SBC), it is worthwhile overlooking this shortcoming for the **initial** deployment. Second, a couple of the variables in this model have a sign issue that is not easy to explain. Since the intent of the model is to project overall wins, for this first iteration, it is acceptable to use this model due to its lower SBC value (compared to all the other models constructed). At this point, it may not be as essential to be concerned with individual parameters as continuing evaluation of the model is necessitated anyways.

Prior to deployment, it would be prudent to consult a subject matter expert to review each parameter's sign. For instance, if a parameter is negative, does that make sense? Furthermore, it would also be wise to see if additional data sources or data points could be retrieved and potentially implemented into the model. Nevertheless, as this is the first model, implementing into production will enable the team to understand it's performance. It is suggested that this model be monitored for a period of three months (at a minimum) to better understand it's performance as new data are consumed.

Conclusion

Eight different models were developed (of which only three have been discussed in this analysis) to predict the number of wins spanning from 1871 through 2006. The original dataset only contained 15 parameters and like any sport, baseball has dozens (if not hundreds) more parameters that could be used. Of note is the fact that many of the data points have been corrected to match the performance of today's 162 game season. This may have inevitably led to many values being extreme outliers or having outlandish results. Although a model has been chosen, future work is required to better understand some of the non-intuitive sign issues as well as the extreme VIF values. This will require further investigation which is outside the scope of this analysis. Finally, it would be prudent to monitor this chosen model's performance for the next three months as new data are fed into it. Next steps may include (but not limited to) refining the model and focusing the reduction of multicollinearity issues.

Bibliography

Beal, D. J. (2007). *Information Criteria Methods in SAS for Multiple Linear Regression Models.* Retrieved October 02, 2016, from Institute for Advanced Analytics - North Carolina State University: http://analytics.ncsu.edu/sesug/2007/SA05.pdf