Assignment #1: Moneyball

PREDICT 411 Section 55

Scott M. Morgan

KAGGLE: ScottMorgan

Scott Morgan
KAGGLE: ScottMorgan

DATE: July 8, 2017

TO: Mr. George Steinbrenner, Principal Owner and Managing Partner

FROM: Mr. Scott Morgan, Senior Analytics Associate

SUBJECT: Scouting Metrics Based on Predictive Modeling (i.e. Sabermetrics)

EXECUTIVE SUMMARY

It has come to the attention of senior leadership the need to devise a more robust scouting system grounded in statistical analysis with the goal of increasing the win percentage of the New York Yankees baseball organization. The factors which lead a professional sports team to win more games is a source of great interest to owners, scouts, coaches and players alike. The knowledge of a team's performance-based metrics that increase the propensity of winning more games, and the development of predictive models based on those metrics, is particularly important to stakeholders charged with personnel decisions. Front office employees can use this information to target players that might be considered "undervalued" by traditional metrics but fit a certain statistical profile.

The following report is a detailed account of the modeling process. The basic managerial recommendation is to focus on recruiting players with the following characteristics:

- *Positive Traits (Seek)*
  - Batting: Base hits, triples and walks
  - Baserunning: Stolen bases

- *Negative Traits (Avoid)*
  - Batting: Strikeouts
  - Baserunning: Caught stealing
  - Fielding: Errors
  - Pitching: Walks allowed

Additionally, it could be useful to pay attention to incomplete or questionable reporting of player statistics as certain missing data points have shown to be conducive to generating more wins. In particular, recruitment personnel should watch for missing data pertaining to batter strikeouts, baserunner stolen bases and fielding double plays; all of which have been shown to be impactful to wins and could constitute "under-valued" player additions. From an organizational standpoint, we may also wish to practice brevity when reporting these metrics as they may represent a competitive advantage to our organization if opponents are unable to analyze them; limiting their ability develop strategy.

Scott Morgan
KAGGLE: ScottMorgan

INTRODUCTION

The purpose of this report is to build regression models to predict the number of wins a baseball team earns in a given season. To accomplish this objective, I use historical baseball data provided by the course instructor and SAS studio to perform the analysis. In a world where data is becoming increasingly abundant, analysis is often conducted and conclusions are drawn naively without the appropriate formation, statistical knowledge and validation. The accessibility of free open source tools puts analytical power in the hands of anyone who can download software. This has precarious implications for business decision makers regardless of industry who rely on predictive analytics to make informed, forward-looking decisions. In this exercise, we use functionality within SAS to produce an end-to-end predictive modeling process by using an assortment of observed baseball statistics to predict the number of wins in a season with an emphasis on simplicity and interpretability.

RESULTS

In the subsequent sections, we generate and evaluate a series of predictive models. We first use the functionality within SAS Studio to perform a brief exploratory data analysis (EDA) to build an understanding of potential predictor variables and their relationship to the response. Following this, we examine the variables for deficiencies such as missing data and outliers as a precursor to preparing the data set for modeling through imputation, elimination and transformation. Finally, we construct three predictive models. The first will be a rudimentary model using the untransformed, raw data to provide a baseline comparison based on Adjusted R-Squared. The second and third models will be more sophisticated iterations which utilize a clean data set and an assortment of construction techniques. We then recommend the most logical, effective solution for use by management.

Exploratory Data Analysis (EDA). While we will ultimately be using variable selection algorithms as an aid in constructing the predictive models, this portion of the analysis is intended to be a building block to ascertain variables of interest as well as a precursor to cleaning the data set. I begin the analysis by examining the variables in the baseball data set using SAS content procedure. The data set contains 2276 observations and 16 variables; one variable is a unique identifier and thus is excluded from the analysis. Each record represents a professional baseball team from the years 1871 to 2006, inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. All variables in the data set are non-negative discrete variables. For purposes of this project, TARGET_WINS will be the variable we are attempting to predict.

While the naming convention of the variables is relatively effective, the data dictionary is a helpful resource. The data set itself has a wide variety of statistics that appear analytically interesting and appropriate to build a model with to predict the number of wins. Table 1 below provides an alphabetic list of the possible predictor variables, descriptions, theoretical effects on the response variable and their respective correlations to TARGET_WINS (red represents negatively correlated while green is positively correlated). The theoretical effect is particularly important as we will reference this logic during the examination of coefficients in the model building process. To summarize the effects and our overarching hypothesis, metrics that are traditionally considered beneficial to winning games will have a positive impact on wins vis-à-vis the coefficients and metrics that are detrimental will have a negative impact.

Scott Morgan
KAGGLE: ScottMorgan

Table 1: Alphabetic List of Variables, Theoretical Effect and Correlations

Variable	Type	Label	Theoretical Effect	Correlations
TEAM_BASERUN_CS	Num	Caught stealing	Negative Impact on Wins	0.02
TEAM_BASERUN_SB	Num	Stolen bases	Positive Impact on Wins	0.14
TEAM_BATTING_2B	Num	Doubles by	Positive Impact on Wins	0.29
TEAM_BATTING_3B	Num	Triples by batters	Positive Impact on Wins	0.14
TEAM_BATTING_BB	Num	Walks by batters	Positive Impact on Wins	0.23
TEAM_BATTING_H	Num	Base Hits by	Positive Impact on Wins	0.39
TEAM_BATTING_HBP	Num	Batters hit by	Positive Impact on Wins	0.07
TEAM_BATTING_HR	Num	Homeruns by	Positive Impact on Wins	0.18
TEAM_BATTING_SO	Num	Strikeouts by	Negative Impact on Wins	-0.03
TEAM_FIELDING_DP	Num	Double Plays	Positive Impact on Wins	-0.03
TEAM_FIELDING_E	Num	Errors	Negative Impact on Wins	-0.18
TEAM_PITCHING_BB	Num	Walks allowed	Negative Impact on Wins	0.12
TEAM_PITCHING_H	Num	Hits allowed	Negative Impact on Wins	-0.11
TEAM_PITCHING_HR	Num	Homeruns	Negative Impact on Wins	0.19
TEAM_PITCHING_SO	Num	Strikeouts by	Positive Impact on Wins	-0.08

Correlations. Using the PROC CORR function we begin our analysis of the variables. One of the key assumptions of ordinary least squares regression is the relationship between the independent and dependent variables is linear. From the table above, we can see that TEAM_BATTING_H, TEAM_BATTING_BB, TEAM_BATTING_2B are moderately positively correlated with TARGET_WINS. Intuitively these make sense and are in the predicted direction. Conversely, only TEAM_FIELDING_E and TEAM_PITCHING_H are mildly negatively correlated with TARGET_WINS. These are also logical and in the predicted direction. The magnitude of these findings is somewhat disappointing as we hope for a correlation of +0.5 (-0.5) or more (less). Exhibits 1 and 2 below provide scatterplots for these variables to the response variable.

Exhibit 1: Scatterplot(s) for Variables with Highest Correlation to the Response Variable

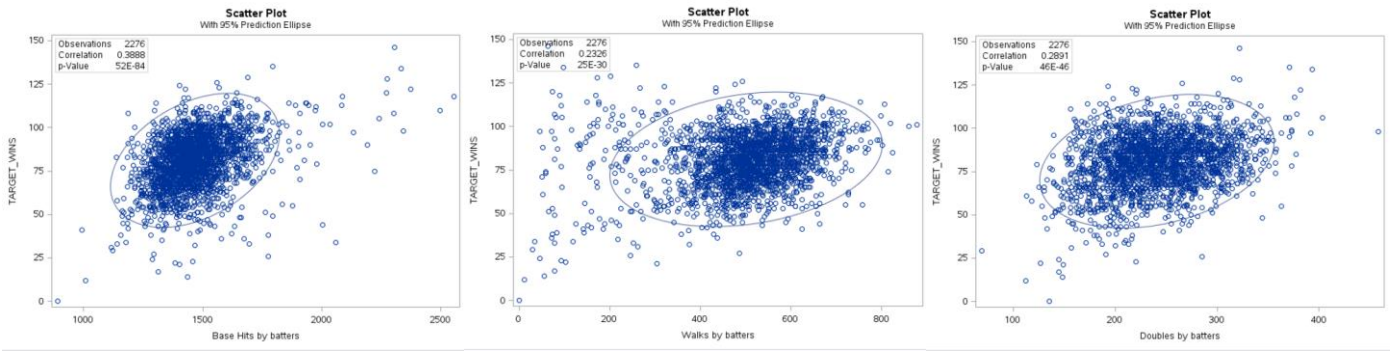
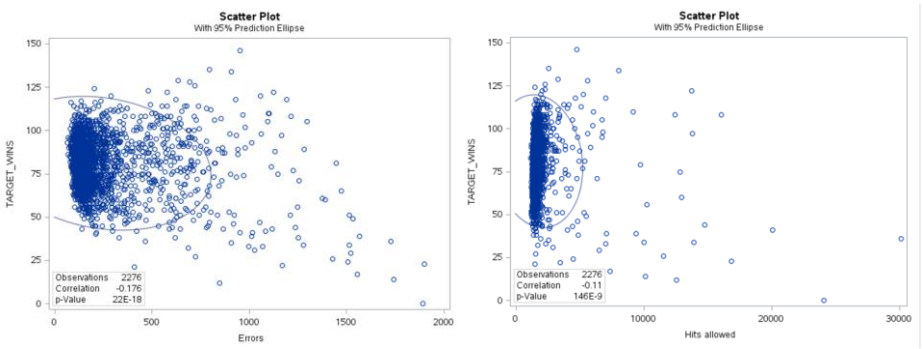


Exhibit 2: Scatterplot(s) for Variables with Lowest Correlation to the Response Variable



The results from Exhibit 1 are encouraging as there appears to be recognizable positive relationships between the chosen independent variables and the response. In Exhibit 2 there appears to be a fairly clear negative relationship between TEAM_FIELDING_E (Errors) and TARGET_WINS while the same cannot be said for TEAM_PITCHING_H (Hits allowed). In addition to providing insight into the linear relationships of the variables, these charts also further assist in identifying possible outliers, which potentially exist in every variable displayed. This represents an area were further analysis and possible action is needed.

Locally Estimated Scatterplot Smoother (LOESS). The above variables are further analyzed by graphing them as scatterplots and superimposed with the Locally Estimated Scatterplot Smoother (LOESS). Exhibit 3 illustrates the scatterplots of the highest correlated variables with and without the LOESS curve. The LOESS scatterplot is of interest because it is a locally ("neighborhood") weighted, non-parametric fitting technique. The biggest advantage of LOESS is that it does not require a function to fit a model to all the data in a sample. As it pertains to the variables in Exhibit 3, we can see how the distribution and presence of outliers is modestly impacting the regression lines. This influence is much more extreme in Exhibit 4.

Exhibit 3: Scatterplot(s) for Variables with Highest Correlation to the Response Variable with and without LOESS

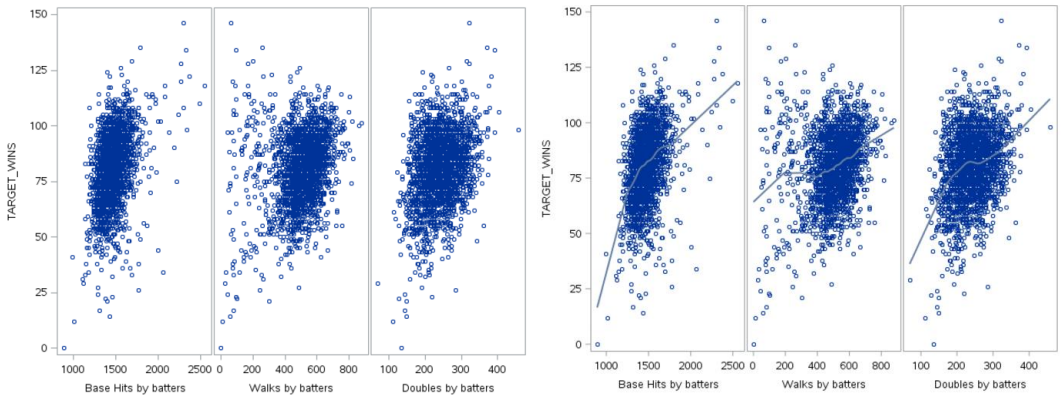
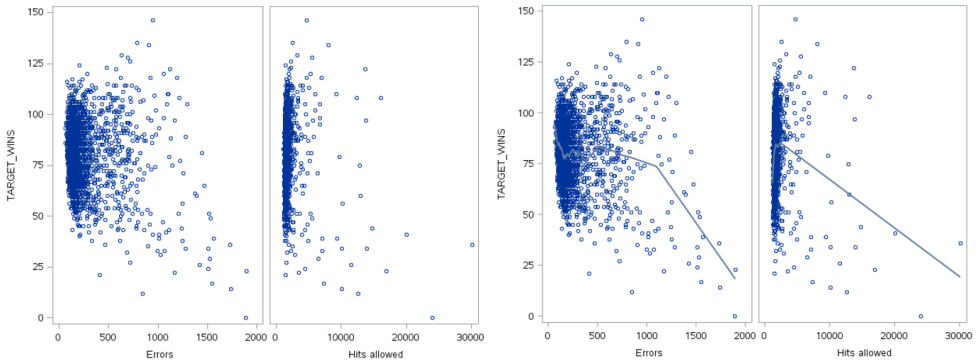


Exhibit 4: Scatterplot(s) for Variables with Lowest Correlation to the Response Variable with and without LOESS



Outliers. As we begin to transition into the data preparation portion of the modeling process, we take a step back to identify potential outliers in the variables. Table 2 below provides a detailed statistical summary of all the variables in the data set. Over half of the variables, including the response, appear relatively normally distributed as observed skewness measures are close to zero though many appear leptokurtic in nature. For reference, a standard normal distribution has a skewness of zero and a kurtosis of three. Variables of concern include: TEAM_BATTING H,TEAM_BASERUN CS,TEAM PITCHING H,TEAM_PITCHING BB, TEAM PITCHING SO and TEAM FIELDING E. There could be a number reasons for the presence outliers. First, there are data integrity issues when dealing with publicly available data sets. The original proprietors of the data altered and scaled the data differently than what was initially measured. Additionally, having not collected these data ourselves we are forced to rely on the original practitioners in terms of collection techniques and accuracy. We discuss how to handle the presence of outliers in the subsequent section on data preparation.

Missing Data. Similar to identifying outliers, Table 2 assists in finding variables where data is missing. Missing data is a key issue because nearly all standard statistical methods presume complete information for all the variables included in an analysis (Soley-Bori, 2013). Incomplete data sets can reduce sample sizes, skew distributions and ultimately weaken the predictive power of regression models. In the baseball data set, there are several variables with missing data: TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS TEAM_BATTING_HBP, TEAM_PITCHING_SO and TEAM_FIELDING_DP. We again reserve discussing the resolution of this issue to the following discussion on data preparation.

Table 2: The MEANS Procedure (Raw Data Set)

Variable	Mean	N	N Miss	Min	5th Pctl	Med	95th Pctl	Max	Kurtosis	Skewness
TARGET_WINS	80.79	2276	0	0	54	82	104	146	1.04	-0.40
TEAM_BATTING_H	1469.27	2276	0	891	1280	1454	1696	2554	7.31	1.57
TEAM_BATTING_2B	241.25	2276	0	69	167	238	320	458	0.01	0.22
TEAM_BATTING_3B	55.25	2276	0	0	23	47	108	223	1.51	1.11
TEAM_BATTING_HR	99.61	2276	0	0	14	102	199	264	-0.96	0.19
TEAM_BATTING_BB	501.56	2276	0	0	246	512	671	878	2.19	-1.03
TEAM_BATTING_SO	735.61	2174	102	0	359	750	1104	1399	-0.32	-0.30
TEAM_BASERUN_SB	124.76	2145	131	0	35	101	302	697	5.51	1.98
TEAM_BASERUN_CS	52.80	1504	772	0	24	49	91	201	7.66	1.98
TEAM_BATTING_HBP	59.36	191	2085	29	40	58	83	95	-0.05	0.32
TEAM_PITCHING_H	1779.21	2276	0	1137	1316	1518	2563	30132	142.28	10.34
TEAM_PITCHING_HR	105.70	2276	0	0	18	107	210	343	-0.60	0.29
TEAM_PITCHING_BB	553.01	2276	0	0	377	536.5	757	3645	97.27	6.75
TEAM_PITCHING_SO	817.73	2174	102	0	420	813.5	1173	19278	673.36	22.21
TEAM_FIELDING_E	246.48	2276	0	65	100	159	716	1898	11.01	2.99
TEAM_FIELDING_DP	146.39	1990	286	52	98	149	186	228	0.19	-0.39

The initial EDA uncovered several interesting idiosyncrasies in what seemed like an otherwise straightforward data set. The correlations of the possible predictor variables with the response were not as robust as expected and several procedures must be performed to clean the data and possibly improve fit in preparation for generating regression models.

Data Preparation. In this section we prepare the data by changing the original values as well as create several new variables. Logarithm transformations were performed on the five variables identified in the EDA for use strictly in the final model. We reserve discussing this logic until presentation of that model. Before moving on to the model portion of the discussion, we present a summary of the new data set excluding the transformations.

Outlier Resolution. In attempt to create a better fit for our models, we first modify the data set to eliminate possible outliers as they can significantly influence regression results. To accomplish this, we impute outlier values, replacing them with the 5th and 95th percentile breakpoints for each respective variable. TEAM_BATTING_H, TEAM_PITCHING_H, and TEAM_PITCHING_SO have extreme low and high values, so we impute at both breakpoints. TEAM_BASERUN_CS, TEAM_PITCHING_BB and TEAM_FIELDING_E appear to only have extreme high values, therefore we impute using only the 95th percentile breakpoints.

Missing Data Resolution. Similar to the outlier remedy, we impute missing values in the data set by replacing them with the median values of the distribution. While there are more sophisticated replacement techniques, median imputation is used because it is a number that is already present in the data set and is less susceptible to outlier errors as compared to mean imputation; keeping in mind this could also lead to biased estimates of the residuals. In addition to median imputation, a missing flag for each variable is generated to determine whether there is a difference in outcomes associated with missing versus complete data. We apply this methodology to TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS TEAM_PITCHING_SO and TEAM_FIELDING_DP. Lastly, the variable

TEAM_BATTING_HBP has 2085 missing values (92%). Given this significant discrepancy, the varible has been removed from the new data set.

The resultant structure of the new data set (not shown) has no missing values and five additional variables in the form of missing data flags. The aforementioned imputations appear to have been effective as the skewness and kurtosis of the data has been reduced. Outlier deletion and imputation are intended to improve the robustness of models and can have powerful effects on fit. While the effects are sometimes not in the desired direction, as we will soon learn, these analytical activities can be beneficial if they improve the linear relationship between two or more variables.

Regression Model Construction. In this section, we generate three linear regression models and assess model adequacy.

Model 1: Baseline. For our initial model, we use the raw, untransformed data. It will skip all records with missing observations, but this will give a high Adjusted R-squared target to strive for. Using the SAS simple linear regression procedure, we generate the following parameter estimates (Table 3):

Table 3: Parameter Estimates for Model 1

Variable	Label	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	60.2883	19.6784	3.0600	0.0025	0.00
TEAM_BATTING_H	Base Hits by batters	1.9135	2.7614	0.6900	0.4893	117182.00
TEAM_BATTING_2B	Doubles by batters	0.0264	0.0303	0.8700	0.3848	1.69
TEAM_BATTING_3B	Triples by batters	-0.1012	0.0775	-1.3100	0.1935	1.30
TEAM_BATTING_HR	Homeruns by batters	-4.8437	10.5085	-0.4600	0.6454	307480.00
TEAM_BATTING_BB	Walks by batters	-4.4597	3.6362	-1.2300	0.2217	196285.00
TEAM_BATTING_SO	Strikeouts by batters	0.3420	2.5988	0.1300	0.8955	194175.00
TEAM_BASERUN_SB	Stolen bases	0.0330	0.0287	1.1500	0.2507	1.95
TEAM_BASERUN_CS	Caught stealing	-0.0110	0.0714	-0.1500	0.8773	1.91
TEAM_BATTING_HBP	Batters hit by pitch	0.0825	0.0496	1.6600	0.0982	1.10
TEAM_FIELDING_E	Errors	-0.1720	0.0414	-4.1600	<.0001	1.26
TEAM_FIELDING_DP	Double Plays	-0.1082	0.0365	-2.9600	0.0035	1.10
TEAM_PITCHING_BB	Walks allowed	4.5109	3.6337	1.2400	0.2161	196404.00
TEAM_PITCHING_H	Hits allowed	-1.8910	2.7610	-0.6800	0.4943	116042.00
TEAM_PITCHING_HR	Homeruns allowed	4.9304	10.5066	0.4700	0.6395	306962.00
TEAM_PITCHING_SO	Strikeouts by pitchers	-0.3736	2.5971	-0.1400	0.8858	194632.00

This results in the following model in equation form:

TARGET_WINS = 60.2883

+	1.9135	x	Base Hits by batters
+	0.0264	x	Doubles by batters
-	0.1012	x	Triples by batters
-	4.8437	x	Homeruns by batters
-	4.4597	x	Walks by batters
+	0.3420	x	Strikeouts by batters
+	0.0330	x	Stolen bases
-	0.0110	x	Caught stealing
+	0.0825	x	Batters hit by pitch
-	0.1720	x	Errors
-	0.1082	x	Double Plays
+	4.5109	x	Walks allowed
-	1.8910	x	Hits allowed
+	4.9304	x	Homeruns allowed
-	0.3736	x	Strikeouts by pitchers

In terms of Model 1, seven of the 15 independent variable coefficients make sense while the remaining variables are not in the predicted direction. The interpretation of the Model 1 is simple given the absence of transformations. Within the context of this model, if all of the independent variables were equal to zero, then we expect a team to win 60 games in a season, which is reasonable given historic standards. Additionally, for every increase of 1 unit across all of the independent variables, we expect the average win total to remain static (~-0.12). This result would be of little use to decision makers as it does not generate any actionable insight. The structural characteristics of the original data set is the likely culprit and its impact on Model 1 needs to be explored further.

Metrics that anecdotally positively impact wins which matched the theoretical effect include Base Hits by batters, Doubles by batters, Stolen bases and Batters hit by pitch. Metrics that should have negative impacts on wins which matched the theoretical effect include Caught stealing, Errors and Hits allowed. However, only Errors was statistically significant as determined by the p-value.

Several of coefficients did have the intended effect. Triples, walks, homeruns by batters should intuitively be positive as they help teams score points while strikeouts by team batters should negatively impact the propensity to win games. In terms of fielding, double plays should positively impact wins as these help earn outs against the opposing team, limiting their ability to score additional points by retiring batters. Pitching statistics like allowing walks and homeruns should also negatively impact wins as these allow the opponent to score runs. Lastly, strikeouts by pitchers should have a positive impact on wins because, similar to fielding, they limit the opposing team's ability to score runs. Of these counterintuitive results, only Double Plays was found to be statistically significant.

To check for multicollinearity, we use SAS to calculate the variance inflation factors (VIF) to measure how much variance of the coefficients is "inflated" by the existence of correlation among the predictor variables in the model. The guidelines we used for interpreting the VIF are from Minitab.com. VIFs larger than 10 imply serious problems with multicollinearity (Montegomery, Peck & Vinning, 2012). As Table 3 above illustrates, multicollinearity is present and substantial in eight of the variables.

We next shift our attention to the analysis of variance (ANOVA) presented in Table 4 below. Of the 2276 observations in the sample, the number of observations with missing values is 191 (not shown) , which is not ideal. The F Value is low but statistically significant. R-Squared and Adjusted R-Squared are 55% and 51%, respectively, which is lower than we would like but acceptable. We will continue to monitor the Adjusted R-Squared as we build out the models with different predictor variables.

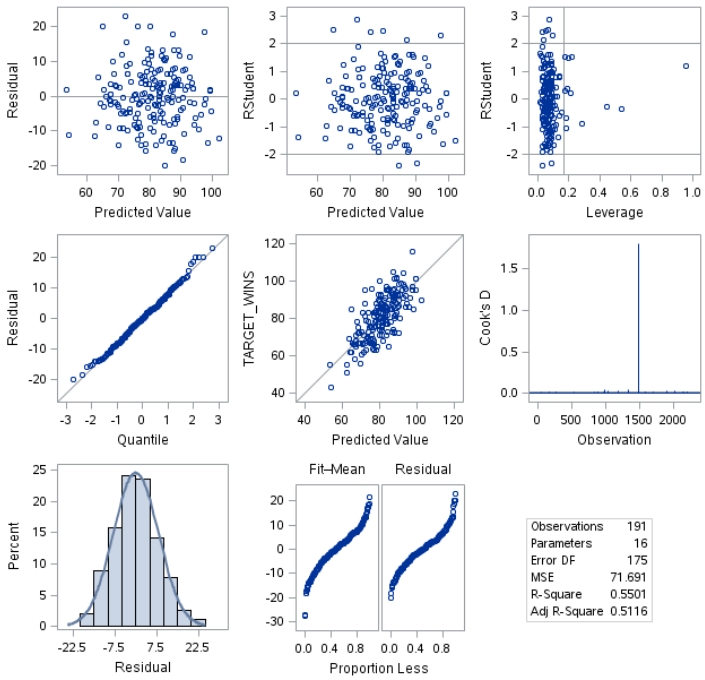
Table 4: Model 1 ANOVA Table

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	15	15341	1022.73892	14.27	<.0001
Error	175	12546	71.6908		
Corrected Total	190	27887			

Root MSE	8.46704	**R-Square**	0.5501
Dependent Mean	80.9267	**Adj R-Sq**	0.5116
Coeff Var	10.46261		

Studying the automatically generated ODS output (Exhibit 5) from SAS we assess the goodness-of-fit for Model 1. The scatterplot for the residual values appears symmetrically distributed, tending to cluster towards the middle of the plot. The y-axis range of the residuals is a bit wider than desired. The Q-Q plot of the residuals suggests a relatively normal distribution. The histogram of the residuals confirms this. Cook's Distance (Cook's D) denotes the presence of a singular large influential point and several very small ones. The left pane of the Fit-Mean Residual plot is slightly lower than the right, indicating that there is still unexplained variation in the response variable. Overall the diagnostics suggests that model is adequate though the high degree of multicollinearity is troublesome and worth monitoring.

Exhibit 5: Fit Diagnostics For Model 1



Model 2: Cleaned Data Set with Hybrid Selection. For the second model, I utilize the cleaned data set as well as a combination of automated and manual variable selection techniques (i.e. "Hybrid"). Below is a description of this process in sequential order followed by an analysis of the final Model 2.

1. Stepwise selection is used as a starting point to narrow down the initial number of variables. Stepwise selection is chosen because of its ability to manage large amounts of potential predictor variables. The stepwise regression selects 14 of independent variables, all with statistically significant p-values. The Adjusted R-Squared is 37%, which is lower than expected given the remedial alterations to the data set. Of the 11 original variables (non-flagged), eight are in the predicted direction. All three flagged variables (M_TEAM_BATTING_SO, M_TEAM_BASERUN_SB, M_TEAM_FIELDING_DP) were statistically significant. The effects of multicollinearity also appear to be dampened considerably.

2. Now that a set of variables has been identified, we reference the theoretical effects from Table 1 and remove those variables with coefficients in the wrong direction as our goal is to build a model that makes intuitive sense. Allowing more hits and homeruns while having fewer double plays does not make sense to winning games and if implemented could bring up numerous ethics and sportsmanship concerns. As these are not practical, Hits allowed, Homeruns allowed and Double Plays are removed from the candidate list of variables and the regression is re-run with the remaining 11 independent variables.

3. The resulting model produces an Adjusted R-Squared of 33%, unfortunately lower than the initial Model 2 and Model 1. All of the coefficients are in the predicted direction, which is very promising. Additionally, the majority of variables are statistically significant and multicollinearity is not a concern. We stop our selection process here and move on to presenting the model and analyzing the fit.

Using the SAS simple linear regression procedure, we generate the following parameter estimates (Table 5):

Table 5: Parameter Estimates for Baseline Model 2

Variable	Label	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	6.48175	5.1634	1.26	0.2095	0
TEAM_BATTING_H	Base Hits by batters	0.049	0.00284	17.26	<.0001	1.42445
TEAM_BATTING_3B	Triples by batters	0.09911	0.0156	6.35	<.0001	2.61398
TEAM_BATTING_BB	Walks by batters	0.02528	0.00587	4.31	<.0001	7.13002
TEAM_BATTING_SO	Strikeouts by batters	-0.00543	0.00175	-3.1	0.002	2.49282
TEAM_BASERUN_SB	Stolen bases	0.07083	0.00508	13.95	<.0001	2.58923
TEAM_BASERUN_CS	Caught stealing	-0.1042	0.02035	-5.12	<.0001	1.24289
TEAM_PITCHING_BB	Walks allowed	-0.00025126	0.00518	-0.05	0.9613	4.39294
TEAM_FIELDING_E	Errors	-0.07748	0.00508	-15.26	<.0001	10.39252
M_TEAM_BATTING_SO	Missing Strikeouts Batters	8.1919	1.48473	5.52	<.0001	1.29971
M_TEAM_BASERUN_SB	Missing Stolen Bases	38.84626	2.08789	18.61	<.0001	3.25689
M_TEAM_FIELDING_DP	Missing Double plays	3.32836	1.52932	2.18	0.0296	3.53919

This results in the following model in equation form:

TARGET_WINS = 6.48175

+	0.0490	x	Base Hits by batters
+	0.0991	x	Triples by batters
+	0.0253	x	Walks by batters
-	0.0054	x	Strikeouts by batters
+	0.0708	x	Stolen bases
-	0.1042	x	Caught stealing
-	0.0003	x	Walks allowed
-	0.0775	x	Errors
+	8.1919	x	Missing Flag Strikeouts Batters
+	38.8463	x	Missing Stolen Bases
+	3.3284	x	Missing Flag Double plays

In terms of Model 2, all the independent variable coefficients are in the predicted directions. The large coefficients associated with the missing flag variables is interesting and noteworthy as there appears to be predictive power in data that is not reported. All predictors are statistically significant with exception of Walks allowed and the Missing Flag for Double plays. Interestingly, three of the five variables analyzed in the EDA made it in the final version of the model. Similar to Model 1, the interpretation of the model is relatively simple given the absence of transformations. Within the context of Model 2, if all of the independent variables were equal to zero, then we expect a team to win 6 games in a season. This is ostensibly low given the worst win total in the modern baseball area (1900 – present) is 20 games, however the probability that any given team not producing any statistics is basically impossible. Additionally, for every increase of 1 unit in all of the independent variables, we expect the average win total to increase by about 50 games, which is more in line with expectations and historical standards.

I do not to iterate the model further for fear of overfitting. Multicollinearity is in an acceptable range. We next shift our attention to the ANOVA in Table 6 below. Of the 2276 observations in the sample, there are no missing variables (not shown). The F Value is higher than Model 1 and statistically significant. R-Squared and Adjusted R-Squared are 34% and 33%, respectively, which is lower than we would like.

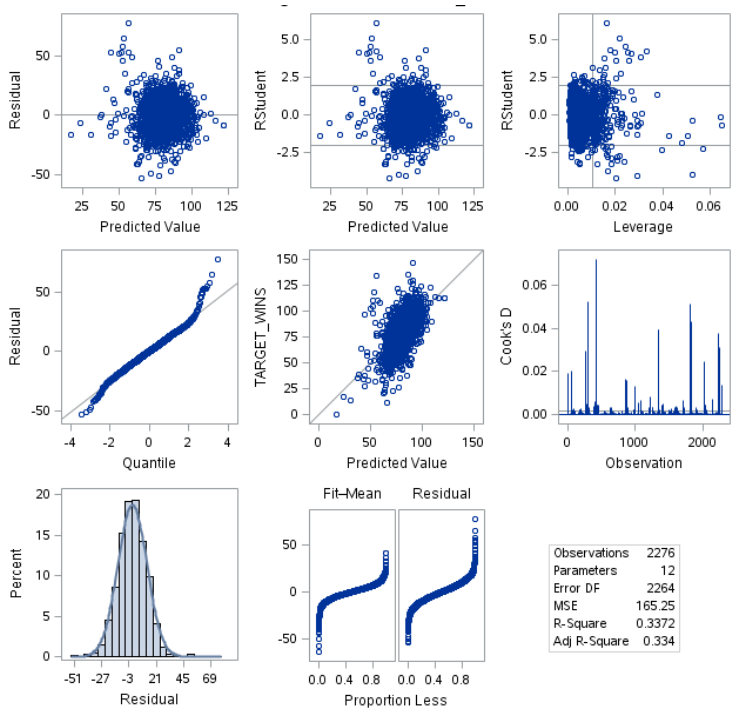
Table 6: Model 2 ANOVA Table

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	11	190373	17307	104.73	<.0001
Error	2264	374123	165.24882		
Corrected Total	2275	564496			

Root MSE	12.85491	**R-Square**	0.3372
Dependent Mean	80.79086	**Adj R-Sq**	0.334
Coeff Var	15.91135		

Studying the automatically generated ODS output (Exhibit 6) from SAS we assess the goodness-of-fit for Model 2. The scatterplot for the residual values appears to be concentrated and pool towards the center of the plot which is not perfect but acceptable. The y-axis range of the residuals is again a bit wider than desired. The Q-Q plot of the residuals shows there is a slight systematic pattern of progressive departure from normality, particularly in the upper right corner suggesting positive skewness. The histogram is relatively normally distributed with slight positive skewness, corroborating our observations from the Q-Q plot. Cook's D denotes the presence of a multiple possible influential points. The left pane of the Fit-Mean Residual plot is lower than the right, indicating that there is still unexplained variation in the response variable.

Exhibit 6: Fit Diagnostics For Model 2



To conclude our evaluation of Model 2, while it might not be as appealing as Model 1 in terms of summary statistics (Adjusted R-Squared, R-Squared, etc.) and fit, it is superior by way of predicted coefficient directionality and thus makes more sense intuitively.

Model 3: Cleaned Data Set with Log Transformations. The last model we present uses the natural logarithms of the five variables identified during the EDA that are "highly" correlated (+/-) with the response: Doubles by batters, Walks by batters, Base Hits by batters, Errors and Hits allowed. The reasoning for the transformation is measurements that cannot be negative often benefit from log re-expression. Using the SAS simple linear regression procedure, we generate the following parameter estimates (Table 7):

Table 7: Parameter Estimates for Baseline Model 3

Parameter Estimates					
Variable	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	-4.18354	0.45681	-9.16	<.0001	0
log_TEAM_BATTING_H	1.08522	0.07677	14.14	<.0001	2.07217
log_TEAM_BATTING_BB	0.15068	0.01618	9.31	<.0001	2.07785
log_TEAM_BATTING_2B	-0.01463	0.03084	-0.47	0.6353	2.2521
log_TEAM_PITCHING_H	0.0039	0.02136	0.18	0.8552	2.10778
log_TEAM_FIELDING_E	-0.04413	0.01051	-4.2	<.0001	2.1548

This results in the following model in equation form:

Log_TARGET_WINS = -4.18354

+	1.08522	x	Log of Base Hits by batters
+	0.15068	x	Log of Walks by batters
-	0.01463	x	Log of Doubles by batters
+	0.00390	x	Log of Hits allowed
-	0.04413	x	Log of Errors

In terms of Model 3, Hits allowed and Doubles by batters are in the wrong direction and are not statistically significant. The interpretation of the model is slightly different from the previous models given the log transformations. Within the context of Model 3, if we change all of the independent variables by 1%, then we expect a team to win -3% fewer games in a season. This is not in-line with expectations and difficult to interpret given the results are in percentages instead of games won.

Multicollinearity is in an acceptable range. We next shift our attention to the ANOVA in Table 8 below. Of the 2276 observations in the sample, there is one missing variables (not shown). The F Value is the highest of the three models and statistically significant. R-Squared and Adjusted R-Squared, however, are the lowest, both at 22%.

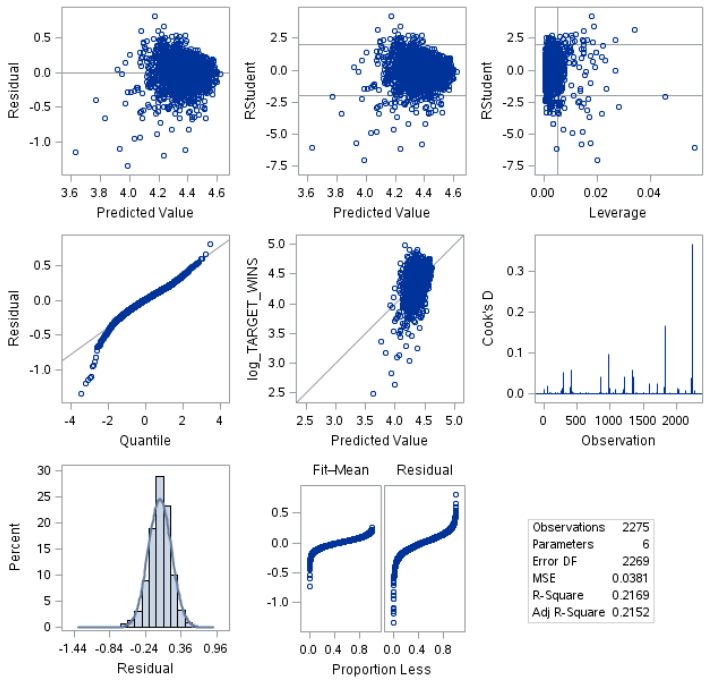
Table 8: Model 3 ANOVA Table

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	23.95156	4.79031	125.71	<.0001
Error	2269	86.46425	0.03811		
Corrected Total	2274	110.41581			

Root MSE	0.19521	**R-Square**	0.2169
Dependent Mean	4.37051	**Adj R-Sq**	0.2152
Coeff Var	4.46652		

Studying the automatically generated ODS output (Exhibit 7) from SAS we assess the goodness-of-fit for Model 3. The scatterplot for the residual values shows pooling to the right of the chart which is not ideal. The Q-Q plot of the residuals also suggests there is a systematic pattern of progressive departure from normality, particularly in the lower left corner signifying negative skewness; the histogram confirms this. Cook's D denotes the presence of multiple possible influential points. The left pane of the Fit-Mean Residual plot is much lower than the right, indicating that there is still unexplained variation in the response variable.

Exhibit 7: Fit Diagnostics For Model 3



Several concerns arise when using transformed variables as we did in Model 3. First, the log transformation appears to have decreased the predictive power of the model and created negative skewness in the residuals. While the number of outliers seems to have been reduced in comparison to Model 2, Cook's D suggests there remain several influential points. The interpretation of the log of TARGET_WINS is difficult as well. In the case of a Y variable transformation, we interpret the coefficient's impact in terms of a percentage change of the predicted wins.

Model Selection. We have presented three different predictive models and now must chose the "Best Model." For purposes of this exercise we use a combination of quantitative and qualitive measures to assess the best model. Table 11 below provides this comparison for each of the three models. The main quantitative criteria we will be evaluating on are the F-value, Root MSE, Adjusted R-Squared, Mallow's Cp, AIC and BIC. Our guiding rules are to maximize F-value and Adjusted R-Squared and minimize Root MSE, Mallow's Cp, AIC and BIC.

Table 11: Model Comparison Statistics

Model	Model 1	Model 2	Model 3
Independent Variables	15	11	5
Variables in Predict Direction	7	11	3
Missing Values	191	0	1
F-Value	14.27	104.73	125.71
Root MSE	8.46	12.855	0.19521
Adj R-Sq	0.5116	0.3340	0.2152
Cp	16.00	12.00	6.00
AIC	831.311	11636.53	-7427.26
BIC	836.220	11638.66	-7425.23

Based on the quantitative criteria, Model 1 and Model 3 appear generally superior to Model 2, particularly in regards to Root MSE, AIC and BIC. Please note the AIC and BIC measures for Model 3 are negative. Hoffam (2004) recommends when these metrics are less than zero to compare which is smallest among the models. In this case, Model 3 would be superior to Model 1 and Model 2. The Mallow's C(p) of Model 3 is also superior to Model 1 and Model 2. Model 1 trumps both other models in terms of Adjusted R-Squared. However, both Model 1 and Model 3 have missing data and, more importantly, contain coefficients that are not in the predicted direction. For these reasons, we recommend the use of Model 2 to management.

CONCLUSION

Building sound regression models is a cornerstone of predictive analytics and to creating a culture of proactive data-driven decision making. For this exercise, a large historical baseball data set was used to construct regression models and select a final model using a variety of criteria. My basic managerial recommendation is to stress simplicity. While there is intellectual appeal in understanding as many factors as possible in what might be predictive of wins, it is best to isolate those few variables that are easily interpretable and implementable but perhaps less statistically robust. For this reason, the model selected does not have any sign issues with the variables, all coefficients are in the predictive direction.

Scott Morgan
KAGGLE: ScottMorgan

BINGO BONUS 1: You'll notice that I used SAS Macros in my file titled Homework_01_Scott_Morgan_Analysis.sas. I even included the periods to look like a pro.

BINGO BONUS 2: I ran each of the three models through the PROC GLM and PROC GENMOD functions; this code is located at the end of the file titled *Homework_01_Scott_Morgan_Analysis_Code*.

PROC GLM appears to produce the same results with slightly different output. PROC GENMOD produced much different outputs compared to PRCO REG and PROC GLM, I assume this is because it uses maximum likelihood estimated. It also included a "Scale" term in the parameter estimates which I am unfamiliar with. The table below provides a comparison of the three models from the assignment. From referencing Hoffman (2003), the Deviance should decrease as we improve the models. This is the case with Model 3. Also, the Log Likelihood should be larger as we improve fit – this again is the case with Model 3. The AIC and BIC are also superior for Model 3 so that is perhaps a good alternative model.

PROC GENMOD

	Model 1	Model 2	Model 3
Deviance	12545.89	374123.3215	86.4643
Scaled Deviance	191	2276	2275
Pearson Chi-Square	12545.89	374123.3215	86.4643
Scaled Pearson X2	191	2276	2275
Log Likelihood	-670.6728	-9035.769	491.5447
Full Log Likelihood	-670.6728	-9035.769	491.5447
AIC (smaller is better)	1375.3456	18097.5379	-969.0894
AICC (smaller is better)	1378.8832	18097.6989	-969.04
BIC (smaller is better)	1430.6343	18172.0302	-928.9813

Another interesting exercise was changing the distribution type for the GENMOD function. As OLS regression should not be used with discrete variables, I tried setting the DIS setting to Poisson as baseball statistics are essentially counts which are non-negative integers. The model coefficients were much different. Below are the same comparative statistics. Following the above logic Model 3 appears to be a good model. Perhaps more work with the coefficients could be done to improve directionality but that is beyond the scope of this excerise.

PROC GENMOD DIST = Poisson

	Model 1	Model 2	Model 3
Deviance	160.0209	5033.7453	20.6234
Scaled Deviance	160.0209	5033.7453	20.6234
Pearson Chi-Square	160.2445	4971.7833	20.2594
Scaled Pearson X2	160.2445	4971.7833	20.2594
Log Likelihood	52549.6835	624900.418	4724.4435
Full Log Likelihood	-674.1926	-9581.379	-3820.4432
AIC (smaller is better)	1380.3851	19186.7579	7652.8865
AICC (smaller is better)	1383.5116	19186.8958	7652.9235
BIC (smaller is better)	1432.4215	19255.52	7687.2649

REFERENCES

Hoffman, J. P. (2003). Generalized linear models: An applied approach. Pearson.

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2012). *Introduction to Linear Regression Analysis.* (5th Edition). New York, NY: Wiley.

Soley-Bori, Marina. “Dealing with missing data: Key assumptions and methods for applied analysis” Boston University School of Public Health (2013, May 2013).