Bingo Bonus

The following are the Bingo Bonus points that are covered throughout this homework assignment.

- (20 Points) PROC GLM and PROC GENMOD were used to do the OLS Regression. The results are discussed in the model building section of this document.
- (10 Points) Use of SAS Macros or good programming technique. This can be seen in the submitted code.
- (10 Points) SCORED FILE was submitted as a SAS DATA file. This can be seen with the submitted files.

Therefore a total of 40 Bingo Bonus points are requested.

Introduction

This assignment discusses the analysis of historical baseball team data from 1871 to 2006 inclusive. This analysis is done in order to create a model that will predict the number of wins for each team based on their historical data using Ordinary Least Squares regression. A number of models are built using different variable selection techniques and the data being analyzed is imputed to address missing values and outliers as much as possible. The selected model is further discussed with regards to its value and strength to predict the number of wins. Once the final model is chosen, a test set of data is used to produce a predicted value for the number of wins in the test data. These predictions are submitted to determine the ranking of the derived model as it compares to the other students in the class.

Data Exploration

Initial exploration of the data set provides details on each of the 17 variables that are present within the data set. All the variables are numeric continuous variables. The following table lists each of the variables along with a short description.

Variable	Description
INDEX	Index of record
TARGET_WINS	Number of wins for season
TEAM_BATTING_H	Base Hits by batters
TEAM_BATTING_2B	Doubles by batters
TEAM_BATTING_3B	Triples by batters
TEAM_BATTING_HR	Homeruns by batters
TEAM_BATTING_BB	Walks by batters
TEAM_BATTING_SO	Strikeouts by batters
TEAM_BASERUN_SB	Stolen bases

TEAM_BASERUN_CS	Caught stealing
TEAM_BATTING_HBP	Batters hit by pitch
TEAM_PITCHING_H	Hits allowed
TEAM_PITCHING_HR	Homeruns allowed
TEAM_PITCHING_BB	Walks allowed
TEAM_PITCHING_SO	Strikeouts by pitchers
TEAM_FIELDING_E	Errors
TEAM_FIELDING_DP	Double Plays

An explanation was given that the data was normalized in order for all records to match up with a 162 game season. This normalization has most likely introduced some outlier values within the data. This will be discussed further later in this document.

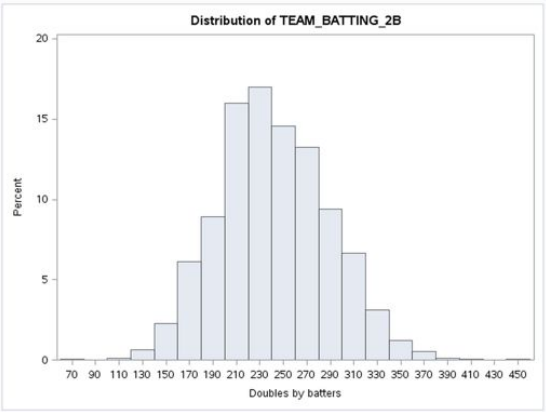
The first step of exploring the data was to collect some basic metrics. This includes number of records, number of records with missing data for each variable, mean, median, standard deviation, minimum value, and maximum value. These metrics are shown in the table below for the 17 variables.

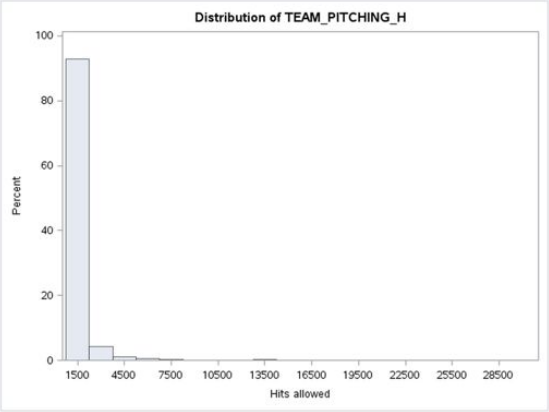
The MEANS Procedure

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

From the table above it can be seen that there are a number of variables that contain missing data (N Miss). The method used to impute these values is discussed in the next section. Looking at the means and median you can gain a sense of how skewed the data may, or may not, be. For instance, the TEAM_PITCHING_H variable shows a maximum of 30,132 and the mean is higher than the median value. This is usually an indication that there are outliers within that variable. Also the value of 30,132 for the number of hits allowed in a 162 game season is not realistic. This may have occurred during the normalization process of the data.

To confirm some of the distributions the following histograms were created.

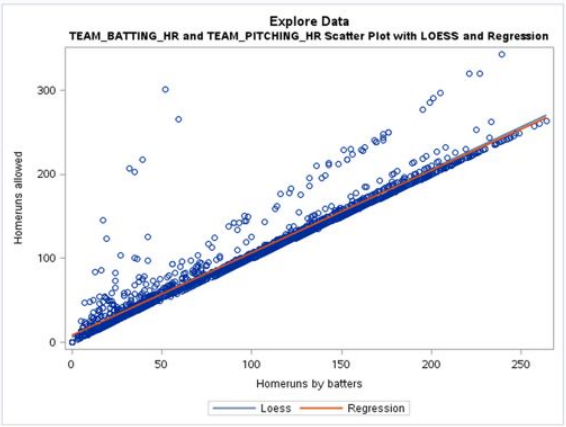




As shown above, the TEAM_BATTING_2B variable appears to have a normal distribution while the TEAM_PITCHING_H variable shows a heavily skewed distribution. Therefore as stated above, the TEAM_PITCHING_H variable most likely has some large outliers that need to be dealt with.

In addition to looking at the distributions across the variables, the correlations between variables were analyzed to identify any multicollinearity that may exist amongst the variables. The Pearson Correlation Coefficients were calculated for each of the variables (this matrix was very large and is the reason for not including within this document). Based on the Correlation Coefficients calculated there are no extremely high correlations between the number of wins and the other variables. Therefore a number of variables will need to be leveraged in order to produce a model that minimizes the errors. The highest correlation with TARGET_WINS is with TEAM_BATTING_H with a Correlation Coefficient of 38.9%. This seems logical since the number of base hits by the batter would contribute to the number of wins.

Analyzing correlations between the independent variables does show some strong correlations. One is a 96.9% correlation between Homeruns allowed by a team and home runs by a batter. The following is a scatterplot showing the correlation between the two.



As shown above, the strong correlation between the two variables is very obvious. When building a model, the variables chosen need to be examined to ensure this strong correlation is not biasing the parameters. As shown in the models built, the stronger models do not include both of these variables since this would cause the VIF to be elevated.

Data Preparation

As discussed in the previous section there are a number of variables that have missing values. These variables, and the number of missing values, are listed in the following table.

Variable	Number of Missing Values
TEAM_BATTING_SO	102
TEAM_BASERUN_SB	131
TEAM_BASERUN_CS	772
TEAM_PITCHING_SO	102
TEAM_FIELDING_DP	286
TEAM_BATTING_HBP	2085

As shown above, the TEAM_BATTING_HBP (batters hit by pitch) variable is missing values for the majority of the 2276 records in the training data set (missing values for 2085). For this reason this variable was dropped from the data set so it is not considered when building models. The reason for this is that it would be very difficult to impute the data with any confidence of being close to what the actual value should have been.

For the other variables above, the missing values were imputed using the median value for each of the related variables. Therefore the following values were used to impute the missing data. These values were obtained from the data exploration step.

Variable	Median
TEAM_BATTING_SO	750
TEAM_BASERUN_SB	101
TEAM_BASERUN_CS	49
TEAM_PITCHING_SO	814
TEAM_FIELDING_DP	149

In addition to imputing the missing data, a new variable was created for each of the five variables to represent a flag that indicates if missing data was imputed. These additional variables will be used within the models. After imputing the missing data and creating the necessary flags, the following table shows the data metrics. As shown, there is no missing data within the training data set and there are five new variables listed in the dataset that are used for the missing data flags.

Proc Means - Missing Data Imputed.

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	1st Pctl	99th Pctl	Maximum
INDEX		2276	0	1268.46	1270.50	736.3490405	1.0000000	26.0000000	2510.00	2535.00
TARGET_WINS		2276	0	80.7908612	82.0000000	15.7521525	0	38.0000000	114.0000000	146.0000000
TEAM_BATTING_H	Base Hits by batters	2276	0	1469.27	1454.00	144.5911954	891.0000000	1188.00	1950.00	2554.00
TEAM_BATTING_2B	Doubles by batters	2276	0	241.2469244	238.0000000	46.8014146	69.0000000	141.0000000	352.0000000	458.0000000
TEAM_BATTING_3B	Triples by batters	2276	0	55.2500000	47.0000000	27.9385570	0	17.0000000	134.0000000	223.0000000
TEAM_BATTING_HR	Homeruns by batters	2276	0	99.6120387	102.0000000	60.5468720	0	4.0000000	235.0000000	264.0000000
TEAM_BATTING_BB	Walks by batters	2276	0	501.5588752	512.0000000	122.6708615	0	79.0000000	755.0000000	878.0000000
TEAM_BATTING_SO	Strikeouts by batters	2276	0	736.2504394	750.0000000	242.9094339	0	72.0000000	1192.00	1399.00
TEAM_BASERUN_SB	Stolen bases	2276	0	123.3941125	101.0000000	85.4056525	0	24.0000000	438.0000000	697.0000000
TEAM_BASERUN_CS	Caught stealing	2276	0	51.5136204	49.0000000	18.7458723	0	18.0000000	125.0000000	201.0000000
TEAM_PITCHING_H	Hits allowed	2276	0	1779.21	1518.00	1406.84	1137.00	1244.00	7093.00	30132.00
TEAM_PITCHING_HR	Homeruns allowed	2276	0	105.6985940	107.0000000	61.2987469	0	8.0000000	244.0000000	343.0000000
TEAM_PITCHING_BB	Walks allowed	2276	0	553.0079086	536.5000000	166.3573617	0	237.0000000	924.0000000	3645.00
TEAM_PITCHING_SO	Strikeouts by pitchers	2276	0	817.5632689	814.0000000	540.5445719	0	208.0000000	1464.00	19278.00
TEAM_FIELDING_E	Errors	2276	0	246.4806678	159.0000000	227.7709724	65.0000000	86.0000000	1237.00	1898.00
TEAM_FIELDING_DP	Double Plays	2276	0	146.7161687	149.0000000	24.5378080	52.0000000	80.0000000	202.0000000	228.0000000
M_TEAM_BATTING_SO		2276	0	0.0448155	0	0.2069441	0	0	1.0000000	1.0000000
M_TEAM_BASERUN_SB		2276	0	0.0575571	0	0.2329552	0	0	1.0000000	1.0000000
M_TEAM_BASERUN_CS		2276	0	0.3391916	0	0.4735390	0	0	1.0000000	1.0000000
M_TEAM_PITCHING_SO		2276	0	0.0448155	0	0.2069441	0	0	1.0000000	1.0000000
M_TEAM_FIELDING_DP		2276	0	0.1256591	0	0.3315376	0	0	1.0000000	1.0000000

As discussed in the previous section there were a number of variables that appeared to have outliers. After reviewing the histograms for the variables, there were six variables that outliers appeared to be a heavy influence to the distribution. These variables are:

- TEAM_BASERUN_SB – Stolen bases
- TEAM_BASERUN_CS – Caught steeling
- TEAM_PITCHING_H – Hits allowed
- TEAM_PITCHING_BB – Walks allowed
- TEAM_PITCHING_SO – Strikeouts by pitcher
- TEAM_FIELDING_E – Errors
- TEAM_BATTING_H – Base hits by batter

In order to minimize the influence of these outliers the 1st and 99th percentile values was determined in the above table. These values provided a range for the variables where 98% of the data falls within. If any values fall outside of these ranges they are adjusted to the 1st or 99th percentile value, depending on which they are closer to. The following table shows the metrics for the variables after the missing data and outliers have been imputed.

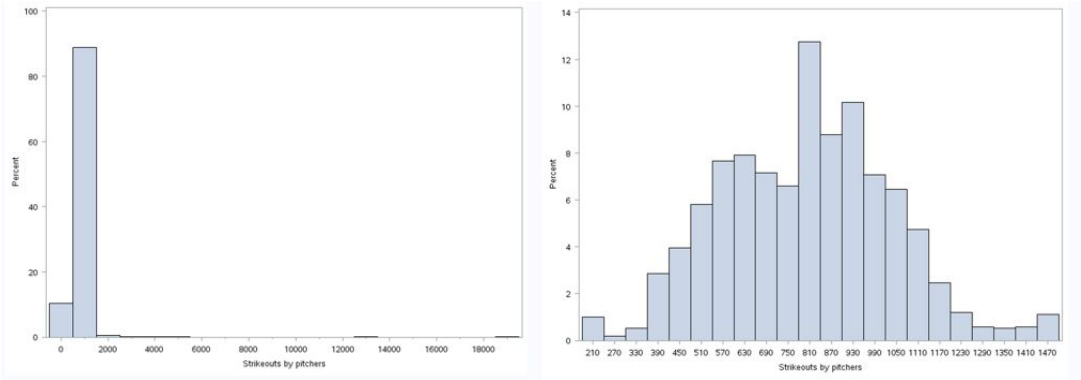
Verify Outliers corrected

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	1st Pctl	99th Pctl	Maximum
INDEX		2276	0	1268.46	1270.50	736.3490405	1.0000000	26.0000000	2510.00	2535.00
TARGET_WINS		2276	0	80.7908612	82.0000000	15.7521525	0	38.0000000	114.0000000	146.0000000
TEAM_BATTING_H	Base Hits by batters	2276	0	1469.27	1454.00	144.5911954	891.0000000	1188.00	1950.00	2554.00
TEAM_BATTING_2B	Doubles by batters	2276	0	241.2469244	238.0000000	46.8014146	69.0000000	141.0000000	352.0000000	458.0000000
TEAM_BATTING_3B	Triples by batters	2276	0	55.2500000	47.0000000	27.9385570	0	17.0000000	134.0000000	223.0000000
TEAM_BATTING_HR	Homeruns by batters	2276	0	99.6120387	102.0000000	60.5468720	0	4.0000000	235.0000000	264.0000000
TEAM_BATTING_BB	Walks by batters	2276	0	501.5588752	512.0000000	122.6708615	0	79.0000000	755.0000000	878.0000000
TEAM_BATTING_SO	Strikeouts by batters	2276	0	736.2504394	750.0000000	242.9094339	0	72.0000000	1192.00	1399.00
TEAM_BASERUN_SB	Stolen bases	2276	0	122.5272406	101.0000000	81.1137942	24.0000000	24.0000000	438.0000000	438.0000000
TEAM_BASERUN_CS	Caught stealing	2276	0	51.2196837	49.0000000	16.8490670	18.0000000	18.0000000	125.0000000	125.0000000
TEAM_PITCHING_H	Hits allowed	2276	0	1716.39	1518.00	789.6833894	1244.00	1244.00	7093.00	7093.00
TEAM_PITCHING_HR	Homeruns allowed	2276	0	105.6985940	107.0000000	61.2987469	0	8.0000000	244.0000000	343.0000000
TEAM_PITCHING_BB	Walks allowed	2276	0	547.7759227	536.5000000	116.3176358	237.0000000	237.0000000	924.0000000	924.0000000
TEAM_PITCHING_SO	Strikeouts by pitchers	2276	0	800.5896309	814.0000000	236.1816182	208.0000000	208.0000000	1464.00	1464.00
TEAM_FIELDING_E	Errors	2276	0	244.1463093	159.0000000	214.8264296	86.0000000	86.0000000	1237.00	1237.00
TEAM_FIELDING_DP	Double Plays	2276	0	146.7161687	149.0000000	24.5378080	52.0000000	80.0000000	202.0000000	228.0000000
M_TEAM_BATTING_SO		2276	0	0.0448155	0	0.2069441	0	0	1.0000000	1.0000000
M_TEAM_BASERUN_SB		2276	0	0.0575571	0	0.2329552	0	0	1.0000000	1.0000000
M_TEAM_BASERUN_CS		2276	0	0.3391916	0	0.4735390	0	0	1.0000000	1.0000000
M_TEAM_PITCHING_SO		2276	0	0.0448155	0	0.2069441	0	0	1.0000000	1.0000000
M_TEAM_FIELDING_DP		2276	0	0.1256591	0	0.3315376	0	0	1.0000000	1.0000000

As shown in the table the maximum and minimum of the 6 variables that were identified with strong outliers now are within a range that is more reasonable. To

provide some additional information on the impact of the above steps, below are two histograms for TEAM_PITCHING_H that show the before and after the outliers are corrected.



The plot on the right is more reasonable and may have a higher predictive value within the models.

The data set now has been adjusted for missing data and outliers. This data is used for building some of the models. In addition to this data set another data set was created that made some transformations on some of the variables. The second training data set has the same missing data steps performed. But when making changes to minimize any impacts from outliers a different technique was used when changing 3 of the 6 variables altered for outliers. Instead of using the 1st and 99th percentiles, a binning technique was used for TEAM_PITCHING_H, TEAM_PITCHING_BB, and TEAM_PITCHING_SO. For these variables, three variables were created (QUANT_TEAM_PITCHING_H, QUANT_TEAM_PITCHING_BB, and QUANT_TEAM_PITCHING_SO) that will have values from 1 to 10. Each number represents an additional 10 percentile.

Binning the three variables discussed above creates a second data set that is used for additional model building. These two data sets are listed below:

- IMP_OUTLIERCOR_MB – Missing data imputed. The 6 outlier variables are constrained to have values between the 1st and 99th percentile.
- IMP_OUTLIERCOR_MB2 – Missing data imputed. Three of the 6 variables are constrained to have values between the 1st and 99th percentile. The other 3 variables binned into corresponding variables based on multiples of 10 percentiles.

Some experimentation was done by transforming some variables by taking the logarithm of the variable, are by combining variables. However, there were no combinations attempted that improved the models that were created. Therefore, these were not done for the final model evaluations.

Build Models

The following sections describe the different models built using the modified data sets discussed above.

Model 1

This model uses the IMP_OUTLIERCOR_MB data set. For this model all of the variables were used without any selection method. This is a baseline model to see where the prediction falls based on all the variables. As shown in the following output you will notice the R-Square is 0.4219. Therefore using these variables explains 42% of the variability in the TARGET_WINS variable.

The REG Procedure
Model: MODEL1
Dependent Variable: TARGET_WINS

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	238189	13233	91.53	<.0001
Error	2257	326308	144.57592		
Corrected Total	2275	564496			

Root MSE	12.02397	R-Square	0.4219
Dependent Mean	80.79086	Adj R-Sq	0.4173
Coeff Var	14.88284		

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

M_TEAM_PITCHING_SO =	M_TEAM_BATTING_SO

Parameter Estimates

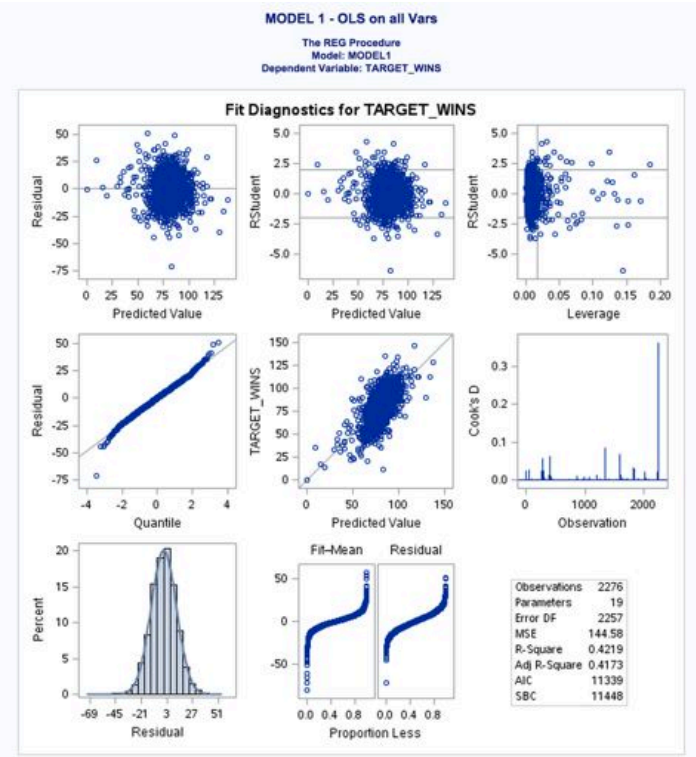
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	20.80130	5.42328	3.84	0.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.04247	0.00364	11.66	<.0001	4.36280
TEAM_BATTING_2B	Doubles by batters	1	-0.03352	0.00855	-3.92	<.0001	2.51853
TEAM_BATTING_3B	Triples by batters	1	0.07276	0.01604	4.54	<.0001	3.15814
TEAM_BATTING_HR	Homeruns by batters	1	-0.02047	0.03017	-0.68	0.4974	52.49226
TEAM_BATTING_BB	Walks by batters	1	0.06167	0.00957	6.44	<.0001	21.70311
TEAM_BATTING_SO	Strikeouts by batters	1	-0.00394	0.00571	-0.69	0.4903	30.24343
TEAM_BASERUN_SB	Stolen bases	1	0.06201	0.00537	11.55	<.0001	2.98453
TEAM_BASERUN_CS	Caught stealing	1	-0.03994	0.01826	-2.19	0.0288	1.48977
TEAM_PITCHING_H	Hits allowed	1	0.00687	0.00092522	7.43	<.0001	8.40008
TEAM_PITCHING_HR	Homeruns allowed	1	0.07492	0.02686	2.79	0.0053	42.65287
TEAM_PITCHING_BB	Walks allowed	1	-0.03171	0.00810	-3.91	<.0001	13.97154
TEAM_PITCHING_SO	Strikeouts by pitchers	1	-0.00585	0.00493	-1.19	0.2352	21.33982
TEAM_FIELDING_E	Errors	1	-0.06906	0.00402	-17.16	<.0001	11.76016
TEAM_FIELDING_DP	Double Plays	1	-0.09698	0.01366	-7.10	<.0001	1.76899
M_TEAM_BATTING_SO		B	7.83072	1.50618	5.20	<.0001	1.52878
M_TEAM_BASERUN_SB		1	39.32623	1.98759	19.79	<.0001	3.37353
M_TEAM_BASERUN_CS		1	0.67687	0.94083	0.72	0.4719	3.12336
M_TEAM_PITCHING_SO		0	0	.	.	.	.
M_TEAM_FIELDING_DP		1	5.53893	1.54747	3.58	0.0004	4.14184

One item to note is the comments regarding the variables M_TEAM_BATTING_SO and M_TEAM_PITCHING_SO. There is a perfect correlation between these two variables. After further investigation it was found that for every record that had a missing value for the TEAM_BATTING_SO variable, it also had a missing value for TEAM_PITCHING_SO. Due to this, the missing value flags were both set to 1 for the same records. Using both of these missing value flags will result in biased parameter coefficients. SAS notes this in the output. Therefore, only one of these missing value flags should be used and found in the final model used.

The other values to examine are the Variance Inflation values (VIF) in the tables above. Any values that are above 10 are a possible indication of strong

multicollinearity. As one can see above, there are a number of values considerably higher than 10. Therefore, this confirms the strong multicollinearity within this model.

The following are the plots created for model 1. There are a few of these plots that should be examined in order to ensure the assumptions required for OLS are not being violated. The residual plot (upper left) should show a random distribution, and there should not be any patterns. Examining the Residual Plot for model 1 shows that the points are randomly distributed. The density of the points falls off at the outer edges, which is most likely caused by some additional outliers, but the majority of the points are random in their distribution.



To confirm the normality assumption holds true we can look at the QQ plot (first plot in the second row) and standardized residuals histogram (plot in bottom left). As you can see above in the QQ plot, the points follow fairly closely to the 45 degree line. There is some slight deviation at the ends, which is most likely due to outlier influences. Looking at the normal distribution histogram you can see that the residuals appear to follow a normal distribution.

The last plot that should be looked at is the Cook's Distance plot (third plot in the second row). This plot helps identify points that have a high influence to the overall fit. Looking at the Cook's Distance plot for this model you can see all the points are below 1, which indicates there are no extremely high influencers. However, there are some points that appear further away from the rest. This is an indication that

there are still outliers, but since they are within acceptable limits they are not heavily influencing the model.

In addition to the plots in the above diagram, there are also some metrics that were produced that assist in evaluating the model. The following are the metrics and their values:

Metric	Value
MSE	144.58
Adj R-Square	0.4173
AIC	11339
SBC	11448

These metrics are used when selecting which model to use, and will be discussed further in the model selection section of this document.

The resulting equation for this model is below:

WINS_TARGET =		20.80130		
	+	0.04247	* TEAM_BATTING_H	Base Hits by batters
	-	0.03352	* TEAM_BATTING_2B	Doubles by batters
	+	0.07276	* TEAM_BATTING_3B	Triples by batters
	-	0.02047	* TEAM_BATTING_HR	Homeruns by batters
	+	0.06167	* TEAM_BATTING_BB	Walks by batters
	-	0.00394	* TEAM_BATTING_SO	Strikeouts by batters
	+	0.06201	* TEAM_BASERUN_SB	Stolen bases
	-	0.03994	* TEAM_BASERUN_CS	Caught stealing
	+	0.00687	* TEAM_PITCHING_H	Hits allowed
	+	0.07492	* TEAM_PITCHING_HR	Homeruns allowed
	-	0.03171	* TEAM_PITCHING_BB	Walks allowed
	-	0.00585	* TEAM_PITCHING_SO	Strikeouts by pitchers
	-	0.06905	* TEAM_FIELDING_E	Errors
	-	0.09698	* TEAM_FIELDING_DP	Double Plays
	+	7.83072	* M_TEAM_BATTING_SO	Missing flag
	+	39.32823	* M_TEAM_BASERUN_SB	Missing flag
	+	0.67687	* M_TEAM_BASERUN_CS	Missing flag
	+	5.53893	* M_TEAM_FIELDING_DP	Missing flag

When examining the signs for each of the parameter estimates it can be seen that some signs do not make sense. For instance, the homeruns allowed by the pitcher is positive. A pitcher allowing homeruns would be a negative impact to the number of games won. This counterintuitive value may be due to the multicollinearity that is discussed above. The discrepancies with the signs is further discussed in the model selection section.

Model 2

This model also uses the IMP_OUTLIERCOR_MB data set. However, instead of using all the variables the stepwise selection approach was used so that SAS could select the variables that should be used. As shown in the following output you will notice

the R-Square is 0.4213. Therefore using these variables explains 42% of the variability in the target variable.

Model 2 - Stepwise

The REG Procedure
Model: MODEL1
Dependent Variable: TARGET_WINS

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	237835	15856	109.70	<.0001
Error	2260	326661	144.54042		
Corrected Total	2275	564496			

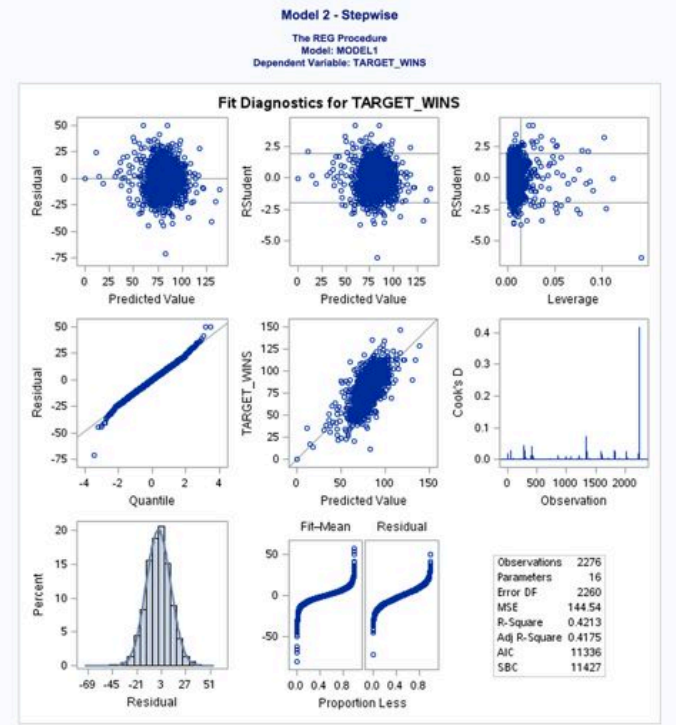
Root MSE	12.02250	R-Square	0.4213
Dependent Mean	80.79086	Adj R-Sq	0.4175
Coeff Var	14.88101		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	20.74403	5.01906	4.13	<.0001	0
TEAM_BATTING_H	Base Hits by batters	1	0.04293	0.00361	11.90	<.0001	4.28142
TEAM_BATTING_2B	Doubles by batters	1	-0.03439	0.00849	-4.05	<.0001	2.48338
TEAM_BATTING_3B	Triples by batters	1	0.07955	0.01534	5.19	<.0001	2.89000
TEAM_BATTING_BB	Walks by batters	1	0.05317	0.00768	6.92	<.0001	13.97156
TEAM_BASERUN_SB	Stolen bases	1	0.06237	0.00521	11.96	<.0001	2.81507
TEAM_BASERUN_CS	Caught stealing	1	-0.04329	0.01691	-2.56	0.0106	1.27815
TEAM_PITCHING_H	Hits allowed	1	0.00665	0.00089498	7.43	<.0001	7.86182
TEAM_PITCHING_HR	Homeruns allowed	1	0.05551	0.00820	6.77	<.0001	3.97897
TEAM_PITCHING_BB	Walks allowed	1	-0.02406	0.00616	-3.91	<.0001	8.07668
TEAM_PITCHING_SO	Strikeouts by pitchers	1	-0.00933	0.00186	-5.02	<.0001	3.03385
TEAM_FIELDING_E	Errors	1	-0.06848	0.00397	-17.25	<.0001	11.45437
TEAM_FIELDING_DP	Double Plays	1	-0.09850	0.01344	-7.33	<.0001	1.71139
M_TEAM_BATTING_SO		1	8.00614	1.46713	5.46	<.0001	1.45089
M_TEAM_BASERUN_SB		1	39.74474	1.86652	21.29	<.0001	2.97580
M_TEAM_FIELDING_DP		1	5.34603	1.52120	3.51	0.0004	4.00341

Examination of the Variance Inflation values (VIF) for the parameters in the table above shows that all but two are below 10. However, the two, TEAM_BATTING_BB and TEAM_FIELDING_E, are not much above 10. Therefore, there may be some slight multicollinearity.

The following are the plots created for model 2. Examining the Residual Plot shows that the points are randomly distributed. The density of the points falls off at the outer edges, which is most likely caused by some additional outliers, but the majority of the points are random in their distribution.



As you can see above in the QQ plot, the points follow fairly closely to the 45 degree line. There is some slight deviation at the ends, which is most likely due to outlier influences. Looking at the normal distribution histogram you can see that the residuals appear to follow a normal distribution.

Looking at the Cook's Distance plot for this model you can see all the points are below 1, which indicates there are no extremely high influencers. However, there are some points that appear further away from the rest. There appears to be fewer points elevated compared to what model 1 had shown. This is an indication that there are still outliers, but since they are within acceptable limits they are not heavily influencing the model.

In addition to the plots in the above diagram, there are also some metrics that were produced that assist in evaluating the model. The following are the metrics and their values:

Metric	Value
MSE	144.54
Adj R-Square	0.4175
AIC	11336
SBC	11427

These metrics will be discussed further in the model selection section of this document.

The resulting equation for this model is below:

WINS_TARGET =		20.74403		
	+	0.04293	* TEAM_BATTING_H	Base Hits by batters
	-	0.03439	* TEAM_BATTING_2B	Doubles by batters
	+	0.07955	* TEAM_BATTING_3B	Triples by batters
	+	0.05317	* TEAM_BATTING_BB	Walks by batters
	+	0.06237	* TEAM_BASERUN_SB	Stolen bases
	-	0.04329	* TEAM_BASERUN_CS	Caught stealing
	+	0.00665	* TEAM_PITCHING_H	Hits allowed
	+	0.05551	* TEAM_PITCHING_HR	Homeruns allowed
	-	0.02406	* TEAM_PITCHING_BB	Walks allowed
	-	0.00933	* TEAM_PITCHING_SO	Strikeouts by pitchers
	-	0.06848	* TEAM_FIELDING_E	Errors
	-	0.09850	* TEAM_FIELDING_DP	Double Plays
	+	8.00614	* M_TEAM_BATTING_SO	Missing flag
	+	39.74474	* M_TEAM_BASERUN_SB	Missing flag
	+	5.34603	* M_TEAM_FIELDING_DP	Missing flag

Since stepwise selection was used, there are fewer variables selected for this model. This model still has some parameter signs that are counterintuitive. For instance, the homeruns allowed by the pitcher is positive. A pitcher allowing homeruns would be a negative impact to the number of games won. This counterintuitive value may be due to the multicollinearity that is discussed above. The discrepancies with the signs are further discussed in the model selection section.

Model 3

This model uses the IMP_OUTLIERCOR_MB2 data set, which uses binning for some of the variables. Also, the backwards selection method was used in this model. As shown in the following output you will notice the R-Square is 0.4245. Therefore using these variables explains 42.5% of the variability in the target variable.

MODEL 3 - backward using binning data

The REG Procedure
Model: MODEL1
Dependent Variable: TARGET_WINS

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	239604	14094	97.96	<.0001
Error	2258	324892	143.88484		
Corrected Total	2275	564496			

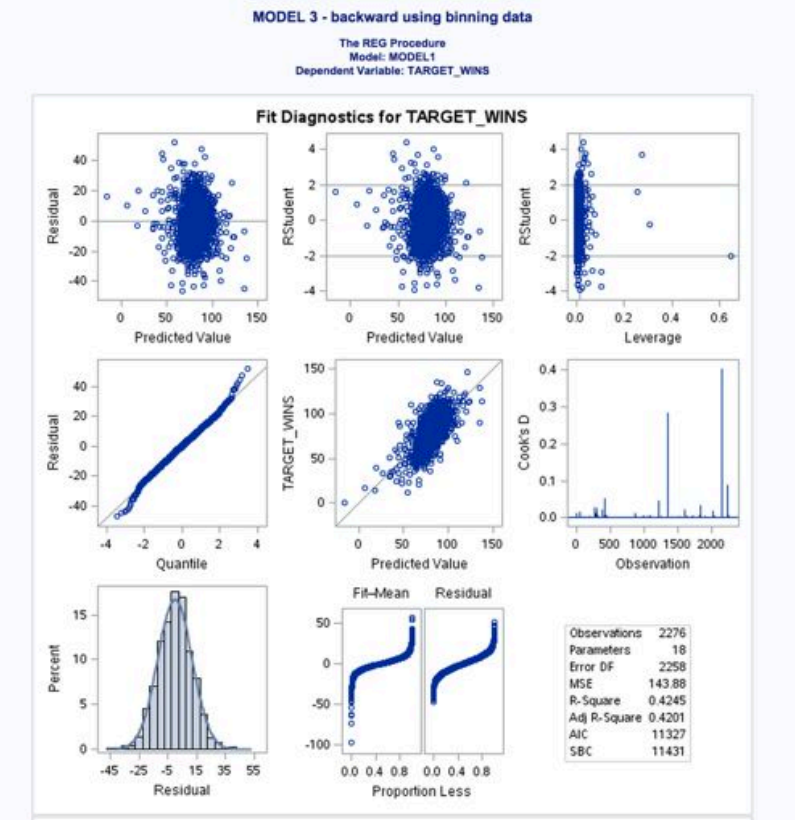
Root MSE	11.99520	R-Square	0.4245
Dependent Mean	80.79086	Adj R-Sq	0.4201
Coeff Var	14.84722		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	14.58021	5.17348	2.82	0.0049	0
TEAM_BATTING_H	Base Hits by batters	1	0.05821	0.00352	16.54	<.0001	4.09693
TEAM_BATTING_2B	Doubles by batters	1	-0.02563	0.00879	-2.92	0.0036	2.67534
TEAM_BATTING_3B	Triples by batters	1	0.06636	0.01526	4.35	<.0001	2.87403
TEAM_BATTING_BB	Walks by batters	1	0.02599	0.00322	8.07	<.0001	2.46988
TEAM_BATTING_SO	Strikeouts by batters	1	-0.02343	0.00292	-8.02	<.0001	7.95535
TEAM_BASERUN_SB	Stolen bases	1	0.05889	0.00523	11.26	<.0001	2.84644
TEAM_BASERUN_CS	Caught stealing	1	-0.04614	0.01693	-2.72	0.0065	1.28708
TEAM_PITCHING_H	Hits allowed	1	0.00056274	0.00030890	1.82	0.0686	2.98596
TEAM_PITCHING_HR	Homeruns allowed	1	0.06129	0.00823	7.45	<.0001	4.02337
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.00139	0.00065918	2.11	0.0348	2.00742
TEAM_FIELDING_E	Errors	1	-0.06177	0.00360	-17.18	<.0001	9.43582
TEAM_FIELDING_DP	Double Plays	1	-0.09444	0.01343	-7.03	<.0001	1.71681
M_TEAM_BATTING_SO		1	7.22886	1.47492	4.90	<.0001	1.47303
M_TEAM_BASERUN_SB		1	37.28053	1.99445	18.69	<.0001	3.41318
M_TEAM_FIELDING_DP		1	4.69601	1.50209	3.13	0.0018	3.92127
QUANT_TEAM_PITCHING_H		1	-0.84945	0.15949	-5.33	<.0001	3.30984
QUANT_TEAM_PITCHING_SO		1	0.89336	0.18886	4.73	<.0001	3.39153

Examination of the Variance Inflation values (VIF) for the parameters in the table above shows that all are below 10. This is a good indication that the effects of multicollinearity are not very high.

The following are the plots created for model 3. Examining the Residual Plot shows that the points are randomly distributed. The density of the points falls off at the outer edges, which is most likely caused by some additional outliers, but the majority of the points are random in their distribution.



As you can see above in the QQ plot, the points follow fairly closely to the 45 degree line. There is some deviation at the ends, which is most likely due to outlier influences. The deviation appears to be slightly larger in this model than the other models. Outliers may be having some additional impacts to the model. Looking at the normal distribution histogram you can see that the residuals appear to follow a normal distribution.

Looking at the Cook's Distance plot for this model you can see all the points are below 1, which indicates there are no extremely high influencers. However, there are some points that appear further away from the rest. This is an indication that there are still outliers, but since they are within acceptable limits they are not heavily influencing the model.

In addition to the plots in the above diagram, there are also some metrics that were produced that assist in evaluating the model. The following are the metrics and their values:

Metric	Value
MSE	143.88
Adj R-Square	0.4201
AIC	11327
SBC	11431

These metrics will be discussed further in the model selection section of this document.

The resulting equation for this model is below:

WINS_TARGET =		14.58021		
	+	0.05821	* TEAM_BATTING_H	Base Hits by batters
	-	0.02563	* TEAM_BATTING_2B	Doubles by batters
	+	0.06636	* TEAM_BATTING_3B	Triples by batters
	+	0.02599	* TEAM_BATTING_BB	Walks by batters
	-	0.02343	* TEAM_BATTING_SO	Strikeouts by batters
	+	0.05889	* TEAM_BASERUN_SB	Stolen bases
	-	0.04614	* TEAM_BASERUN_CS	Caught stealing
	+	0.00056274	* TEAM_PITCHING_H	Hits allowed
	+	0.06129	* TEAM_PITCHING_HR	Homeruns allowed
	+	0.00139	* TEAM_PITCHING_SO	Strikeouts by pitchers
	-	0.06177	* TEAM_FIELDING_E	Errors
	-	0.09444	* TEAM_FIELDING_DP	Double Plays
	+	7.22886	* M_TEAM_BATTING_SO	Missing flag
	+	37.28053	* M_TEAM_BASERUN_SB	Missing flag
	+	4.69601	* M_TEAM_FIELDING_DP	Missing flag
	-	0.84945	* QUANT_TEAM_PITCHING_H	Percentile bin
	+	0.89336	* QUANT_TEAM_PITCHING_SO	Percentile bin

Since backwards selection was used there are fewer variables selected for this model compared to model 1. This model still has some parameter signs that are counterintuitive. For instance, the homeruns allowed by the pitcher is positive. A pitcher allowing homeruns would be a negative impact to the number of games won. This counterintuitive value may be due to the multicollinearity that is discussed above. The discrepancies with the signs are further discussed in the model selection section.

Model 4

This model also uses the IMP_OUTLIERCOR_MB2 data set, which uses binning for some of the variables. Also, the backwards selection method was used in this model. The difference with this model is to reduce the starting list of variables that the selection is run against. There are some variables that are suspected to have some correlation with other variables so they were removed. The removed variables were TEAM_PITCHING_H and M_TEAM_PITCHING_SO. As shown in the following output you will notice the R-Square is 0.4243. Therefore using these variables explains 42% of the variability in the target variable.

MODEL 4 - backward using bining data - Reduced Set 1

The REG Procedure
Model: MODEL1
Dependent Variable: TARGET_WINS

Number of Observations Read	2276
Number of Observations Used	2276

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	239538	14090	97.91	<.0001
Error	2258	324959	143.91447		
Corrected Total	2275	564496			

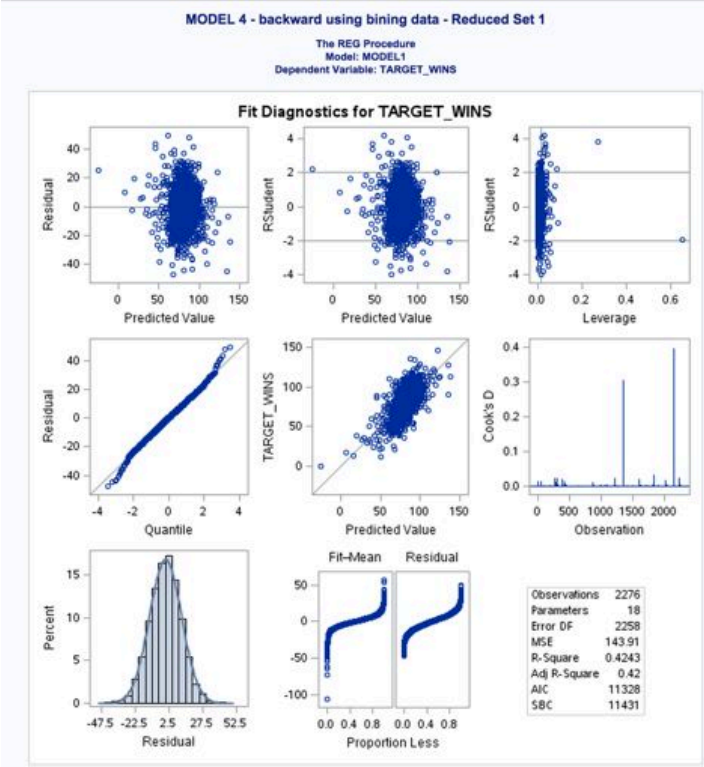
Root MSE	11.99644	R-Square	0.4243
Dependent Mean	80.79086	Adj R-Sq	0.4200
Coeff Var	14.84875		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	17.83319	5.50614	3.24	0.0012	0
TEAM_BATTING_H	Base Hits by batters	1	0.05915	0.00352	16.82	<.0001	4.08587
TEAM_BATTING_2B	Doubles by batters	1	-0.02334	0.00876	-2.66	0.0078	2.65797
TEAM_BATTING_3B	Triples by batters	1	0.06546	0.01522	4.30	<.0001	2.85648
TEAM_BATTING_BB	Walks by batters	1	0.01501	0.00680	2.21	0.0275	11.00160
TEAM_BATTING_SO	Strikeouts by batters	1	-0.02436	0.00281	-8.67	<.0001	7.37207
TEAM_BASERUN_SB	Stolen bases	1	0.05747	0.00518	11.10	<.0001	2.78773
TEAM_BASERUN_CS	Caught stealing	1	-0.04586	0.01693	-2.71	0.0068	1.28631
TEAM_PITCHING_HR	Homeruns allowed	1	0.06158	0.00822	7.49	<.0001	4.01334
TEAM_PITCHING_SO	Strikeouts by pitchers	1	0.00174	0.00058774	2.95	0.0032	1.59556
TEAM_FIELDING_E	Errors	1	-0.06141	0.00356	-17.24	<.0001	9.25392
TEAM_FIELDING_DP	Double Plays	1	-0.09303	0.01344	-6.92	<.0001	1.71896
M_TEAM_BATTING_SO		1	7.29980	1.47619	4.95	<.0001	1.47527
M_TEAM_BASERUN_SB		1	35.93190	1.84814	19.44	<.0001	2.93016
M_TEAM_FIELDING_DP		1	4.76373	1.50811	3.16	0.0016	3.95193
QUANT_TEAM_PITCHING_H		1	-0.94875	0.16535	-5.74	<.0001	3.55679
QUANT_TEAM_PITCHING_BB		1	0.37448	0.22170	1.69	0.0913	6.41061
QUANT_TEAM_PITCHING_SO		1	0.92921	0.18737	4.96	<.0001	3.33740

Examination of the Variance Inflation values (VIF) for the parameters in the table above shows that one variable has a value slightly above 10. This is a good indication that the effects of multicollinearity are not very high.

The following are the plots created for model 4. Examining the Residual Plot shows that the points are randomly distributed. The density of the points falls off at the outer edges, which is most likely caused by some additional outliers, but the majority of the points are random in their distribution.



As you can see above in the QQ plot, the points follow fairly closely to the 45 degree line. There is some deviation at the ends, which is most likely due to outlier influences. The deviation appears to be smaller than previous. Outliers may be having some additional impacts to the model. Looking at the normal distribution histogram you can see that the residuals appear to follow a normal distribution.

Looking at the Cook's Distance plot for this model you can see all the points are below 1, which indicates there are no extremely high influencers. However, there are a couple points that appear further away from the rest. This is an indication that there are still outliers, but the numbers of spikes are fewer in this model.

In addition to the plots in the above diagram, there are also some metrics that were produced that assist in evaluating the model. The following are the metrics and their values:

Metric	Value
MSE	143.91
Adj R-Square	0.42
AIC	11328
SBC	11431

These metrics will be discussed further in the model selection section of this document.

The resulting equation for this model is below:

WINS_TARGET =		17.83319		
	+	0.05915	* TEAM_BATTING_H	Base Hits by batters
	-	0.02334	* TEAM_BATTING_2B	Doubles by batters
	+	0.06546	* TEAM_BATTING_3B	Triples by batters
	+	0.01501	* TEAM_BATTING_BB	Walks by batters
	-	0.02436	* TEAM_BATTING_SO	Strikeouts by batters
	+	0.05747	* TEAM_BASERUN_SB	Stolen bases
	-	0.04586	* TEAM_BASERUN_CS	Caught stealing
	+	0.06158	* TEAM_PITCHING_HR	Homeruns allowed
	+	0.00174	* TEAM_PITCHING_SO	Strikeouts by pitchers
	-	0.06141	* TEAM_FIELDING_E	Errors
	-	0.09303	* TEAM_FIELDING_DP	Double Plays
	+	7.29980	* M_TEAM_BATTING_SO	Missing flag
	+	35.93190	* M_TEAM_BASERUN_SB	Missing flag
	+	4.76373	* M_TEAM_FIELDING_DP	Missing flag
	-	0.94875	* QUANT_TEAM_PITCHING_H	Percentile bin
	+	0.37448	* QUANT_TEAM_PITCHING_BB	Percentile bin
	+	0.92921	* QUANT_TEAM_PITCHING_SO	Percentile bin

Model - Bingo Bonus

These models are produced for the Bingo Bonus using PROC GLM and PROC GENMOD and are not to be considered for model selection. Below are the parameter coefficients using these two methods. The variables selected for these models are the final variable selection from Model 4, which is the chosen model.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	17.83319181	5.50613618	3.24	0.0012
TEAM_BATTING_H	0.05914575	0.00351610	16.82	<.0001
TEAM_BATTING_2B	-0.02334484	0.00876148	-2.66	0.0078
TEAM_BATTING_3B	0.06546323	0.01521505	4.30	<.0001
TEAM_BATTING_BB	0.01500515	0.00680061	2.21	0.0275
TEAM_BATTING_SO	-0.02436310	0.00281133	-8.67	<.0001
TEAM_BASERUN_SB	0.05747349	0.00517717	11.10	<.0001
TEAM_BASERUN_CS	-0.04585903	0.01693005	-2.71	0.0068
TEAM_PITCHING_HR	0.06158347	0.00821984	7.49	<.0001
TEAM_PITCHING_SO	0.00173542	0.00058774	2.95	0.0032
TEAM_FIELDING_E	-0.06140688	0.00356153	-17.24	<.0001
TEAM_FIELDING_DP	-0.09303329	0.01343873	-6.92	<.0001
M_TEAM_BATTING_SO	7.29979504	1.47619439	4.95	<.0001
M_TEAM_BASERUN_SB	35.93189971	1.84813795	19.44	<.0001
M_TEAM_FIELDING_DP	4.76373348	1.50811074	3.16	0.0016
QUANT_TEAM_PITCHING_	-0.94874596	0.16535490	-5.74	<.0001
QUANT_TEAM_PITCHING_	0.37447812	0.22169614	1.69	0.0913
QUANT_TEAM_PITCHING_	0.92921334	0.18736563	4.96	<.0001

PROC GLM Output

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	17.8332	5.4843	7.0841	28.5823	10.57	0.0011
TEAM_BATTING_H	1	0.0591	0.0035	0.0523	0.0660	285.21	<.0001
TEAM_BATTING_2B	1	-0.0233	0.0087	-0.0404	-0.0062	7.16	0.0075
TEAM_BATTING_3B	1	0.0655	0.0152	0.0358	0.0952	18.66	<.0001
TEAM_BATTING_BB	1	0.0150	0.0068	0.0017	0.0283	4.91	0.0267
TEAM_BATTING_SO	1	-0.0244	0.0028	-0.0299	-0.0189	75.70	<.0001
TEAM_BASERUN_SB	1	0.0575	0.0052	0.0474	0.0676	124.22	<.0001
TEAM_BASERUN_CS	1	-0.0459	0.0169	-0.0789	-0.0128	7.40	0.0065
TEAM_PITCHING_HR	1	0.0616	0.0082	0.0455	0.0776	56.58	<.0001
TEAM_PITCHING_SO	1	0.0017	0.0006	0.0006	0.0029	8.79	0.0030
TEAM_FIELDING_E	1	-0.0614	0.0035	-0.0684	-0.0545	299.65	<.0001
TEAM_FIELDING_DP	1	-0.0930	0.0134	-0.1193	-0.0668	48.31	<.0001
M_TEAM_BATTING_SO	1	7.2998	1.4703	4.4180	10.1816	24.65	<.0001
M_TEAM_BASERUN_SB	1	35.9319	1.8408	32.3240	39.5398	381.01	<.0001
M_TEAM_FIELDING_DP	1	4.7637	1.5021	1.8196	7.7079	10.06	0.0015
QUANT_TEAM_PITCHING_	1	-0.9487	0.1647	-1.2716	-0.6259	33.18	<.0001
QUANT_TEAM_PITCHING_	1	0.3745	0.2208	-0.0583	0.8073	2.88	0.0899
QUANT_TEAM_PITCHING_	1	0.9292	0.1866	0.5634	1.2950	24.79	<.0001
Scale	1	11.9489	0.1771	11.6068	12.3011		

Note: The scale parameter was estimated by maximum likelihood.

Proc GENMOD Output

Comparing the parameter coefficients between the generalized linear model method and the PROC REG it is shown that the values are the same. Using the identity function and having the normal distribution causes the GLM approach to be the same as the OLS method.

Select Models

There are 4 models that were created to compare and select the model that is desired for production. Work was done on additional models to reduce the variable selection and perform additional transformations to the data prior to building the models. However, every time this was done the models did not perform as well.

The following are the summary metrics for each of the four models:

Model	R-Square	Adj R-Square	MSE	AIC	SBC
Model 1	0.4219	0.4173	144.58	11339	11448
Model 2	0.4213	0.4175	144.54	11336	11427
Model 3	0.4245	0.4201	143.88	11327	11431
Model 4	0.4243	0.42	143.91	11328	11431

R-Square was provided for informational purposes. Since each model contains different number and selection of variables it is not a metric that can be used for model selection. The more appropriate metric is the adjusted R-Square since it

takes into account the number of variables used. From the above table model 3 and 4 would be the models that are slightly better than model 1 and model 2.

MSE gives an indication of the mean of the square of the error. The smaller the value the better fit of the model to the actual data. AIC and SBC are criterion metrics that are used to compare the relative quality of the models. The lower the values, the better the model. Using the AIC, model 3 would be the recommended model. However, using SBC, model 3 and 4 are tied as the recommended model.

Based on the metrics alone both model 3 and model 4 are the best choices. To further the selection down to one I analyzed the signs of the parameter estimates for any counterintuitive observations. Both models have a few sign issues. The following is a summary of these issues.

Variable	Expected	Model 3	Model 4
TEAM_BATTING_2B	+	-	-
TEAM_FIELDING_DP	+	-	-
TEAM_PITCHING_H	-	+	not used
TEAM_PITCHING_HR	-	+	+

Not shown above are the binned variables that are used. These are omitted since they are a transformation of a variable and not as understood with respect to the sign.

Based on the above chart it is expected that a 2 base hit by the batter will contribute positively to the number of wins. Also a double play from the fielders would also be a positive impact. However, in both models the influence is negative. This may be do to risk by the team to achieve both of these. A batter running to second base versus staying at first may be taking an additional risk and may get out if they are pushing it. With this risk, there can be some negative influence to number of wins based on the other variables. Similarly with making a double play the fielders may be rushing the play and risk the win. Therefore, even though there may be some multicollinearity amongst these variables it is believed it still provides a level of accuracy that is required for the model. Additional investigation by a baseball expert may further explain this.

As was shown in the data exploration step, TEAM_PITCHING_H appeared to be heavily influenced by outliers. Therefore it was seen that not using the variable in the model would be more beneficial. In my opinion this gave an advantage to Model 4.

Based on the above analysis and performance of the models, model 4 is the chosen model and is the model that was used to score the test data being submitted.

Conclusion

Several models were developed in order to predict the number of wins in a season for baseball teams. The data used was baseball statistical data from 1871 to 2000 inclusive. The best model was created using backward selection on imputed data that had outliers corrected using percentile range and binning. Model 4 was the model chosen as the best model to use to perform the prediction of number of wins. A number of metrics and reasoning of signs was used during the model selection.

There were sign issues with a few variables. These anomalies were discussed within this analysis. Some of the sign issues may be due to complex multicollinearity or complexities of understanding baseball statistics. Further investigation can be done to address this further.

Code

The following is the SAS code for both the main analysis program and the standalone scoring program used to produce all the necessary output for this analysis.

Main Analysis Program

```
*******;
* Main analysis code for Moneyball Data;
*;
* Jeffrey Vagg;
* PREDICT 411-Sec55;
*******;

* Setup some variables to use;
%let ME = jeffreyvagg2015;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT01/HW;

* Configure the location of the data files for the assignment.;
libname HW "&PATH." access=readonly;

%let MBFILE = HW.moneyball;

* Display information on the moneyball training data;
proc contents data=&MBFILE.;
run;

* Print the first 10 records of the data to become familiar on how it looks.;
proc print data=&MBFILE.(obs=10);
run;

* Use proc means to detemrine if there is any data missing and ;
* detemrine the mean and median for the data.  This will be used ;
* to impute the data.;
proc means data=&MBFILE. n nmiss mean median stddev min max;
run;

* Show plots for data prior to any work.;
Title "Explore Data";
proc univariate data=&MBFILE. noprint;
histogram
        TARGET_WINS
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
```

```
        TEAM_BATTING_HR
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_BATTING_HBP
        TEAM_PITCHING_H
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
;
run;

* Show the scatterplots and correlation values for the variables.;
ods graphics on;
Title2 "Scatterplot Matrix";
proc corr data=&MBFILE. plot(MAXPOINTS=NONE)=matrix(histogram nvar=ALL);
run;
ods graphics off;

* Show scatter plots between two sets of strongly correlated variables;
ods graphics on;
PROC SGPLOT DATA=&MBFILE.;
        LOESS X= TEAM_BATTING_H Y= TARGET_WINS / NOMARKERS;;
        REG X= TEAM_BATTING_H Y= TARGET_WINS;
        Title2 "TEAM_BATTING_H and TARGET_WINS Scatter Plot with LOESS and Regression";
run;
ods graphics off;

ods graphics on;
PROC SGPLOT DATA=&MBFILE.;
        LOESS X= TEAM_BATTING_HR Y= TEAM_PITCHING_HR / NOMARKERS;;
        REG X= TEAM_BATTING_HR Y= TEAM_PITCHING_HR;
        Title2 "TEAM_BATTING_HR and TEAM_PITCHING_HR Scatter Plot with LOESS and
Regression";
run;
ods graphics off;

*****;
* Create an imputed dataset with missing values filled in.  The;
* median value for the attributes kept was used.;
*****;
data IMP_MB;
        set &MBFILE.;

        * Drop TEAM_BATTING_HBP since the number of missing values is very;
        * high.;
        drop TEAM_BATTING_HBP;

        * Impute missing values using the median values.  Also add a flag;
        * variable that indicates the variable was imputed.  This is done;
        * for the 5 variables that were kept and had missing values.;
        * The median values were obtained from the output of the proc means;
        * done above.;
        M_TEAM_BATTING_SO = 0;
        if missing(TEAM_BATTING_SO) then do;
                TEAM_BATTING_SO = 750;
                M_TEAM_BATTING_SO = 1;
        end;

        M_TEAM_BASERUN_SB = 0;
        if missing(TEAM_BASERUN_SB) then do;
                TEAM_BASERUN_SB = 101;
                M_TEAM_BASERUN_SB = 1;
        end;

        M_TEAM_BASERUN_CS = 0;
        if missing(TEAM_BASERUN_CS) then do;
                TEAM_BASERUN_CS = 49;
                M_TEAM_BASERUN_CS = 1;
```

```
        end;

        M_TEAM_PITCHING_SO = 0;
        if missing(TEAM_PITCHING_SO) then do;
                TEAM_PITCHING_SO=814;
                M_TEAM_PITCHING_SO = 1;
        end;

        M_TEAM_FIELDING_DP = 0;
        if missing(TEAM_FIELDING_DP) then do;
                TEAM_FIELDING_DP = 149;
                M_TEAM_FIELDING_DP = 1;
        end;
run;

* Run a proc means to verify that all missing data has been corrected.;
* Also get the 1 and 99 percentiles so they can be used to constrain;
* the data to take care of any outliers that exist.;
Title "Proc Means - Missing Data Imputed.";
proc means data=IMP_MB n nmiss mean median stddev min p1 p99 max;
run;

* Run univariate on the data to identify any variables that should be ;
* addressed for outliers.;
Title "Look for Outliers";
proc univariate data=IMP_MB noprint;
histogram
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_HR
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_H
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
;
run;


* Based on the graphs from the above univariate command, the following;
* variables have strong outliers.  The 1st and 99th percentile values;
* are given.;
*       TEAM_BASERUN_SB               (24 - 438);
*       TEAM_BASERUN_CS               (18 - 125);
*       TEAM_PITCHING_H               (1244 - 7093);
*       TEAM_PITCHING_BB      (237 - 924);
*       TEAM_PITCHING_SO      (208 - 1464);
*       TEAM_FIELDING_E               (86 - 1237);
* Create a new data set with the outliers constrained;
data IMP_OUTLIERCOR_MB;
        set IMP_MB;

        if TEAM_BASERUN_SB < 24 then TEAM_BASERUN_SB = 24;
        if TEAM_BASERUN_SB > 438 then TEAM_BASERUN_SB = 438;

        if TEAM_BASERUN_CS < 18 then TEAM_BASERUN_CS = 18;
        if TEAM_BASERUN_CS > 125 then TEAM_BASERUN_CS = 125;

        if TEAM_PITCHING_H < 1244 then TEAM_PITCHING_H = 1244;
        if TEAM_PITCHING_H > 7093 then TEAM_PITCHING_H = 7093;

        if TEAM_PITCHING_BB < 237 then TEAM_PITCHING_BB = 237;
        if TEAM_PITCHING_BB > 924 then TEAM_PITCHING_BB = 924;

        if TEAM_PITCHING_SO < 208 then TEAM_PITCHING_SO = 208;
        if TEAM_PITCHING_SO > 1464 then TEAM_PITCHING_SO = 1464;
```

```
        if TEAM_FIELDING_E < 86 then TEAM_FIELDING_E = 86;
        if TEAM_FIELDING_E > 1237 then TEAM_FIELDING_E = 1237;
run;

* Show the proc means for the new dataset to verify the ranges are;
* more reasonable.;
Title "Verify Outliers corrected";
proc means data=IMP_OUTLIERCOR_MB n nmiss mean median stddev min p1 p99 max;
run;

* Run univariate to validate the outliers have been constrained.;
proc univariate data=IMP_OUTLIERCOR_MB noprint;
histogram
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_HR
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_H
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
;
run;

* MODEL 1;
* Run a linear regression on the data with all the variables.;
ods graphics on;
Title "MODEL 1 - OLS on all Vars";
proc reg data=IMP_OUTLIERCOR_MB plots=diagnostics(stats=(default aic sbc));
model TARGET_WINS =
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_HR
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_H
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
        M_TEAM_BATTING_SO
        M_TEAM_BASERUN_SB
        M_TEAM_BASERUN_CS
        M_TEAM_PITCHING_SO
        M_TEAM_FIELDING_DP
        /vif
;
run;
ods graphics off;

* MODEL 2;
* Run a linear regression on the data using stepwise selection.;
ods graphics on;
Title "Model 2 - Stepwise";
proc reg data=IMP_OUTLIERCOR_MB plots=diagnostics(stats=(default aic sbc));
model TARGET_WINS =
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_HR
        TEAM_BATTING_BB
```

```
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_H
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
        M_TEAM_BATTING_SO
        M_TEAM_BASERUN_SB
        M_TEAM_BASERUN_CS
        M_TEAM_PITCHING_SO
        M_TEAM_FIELDING_DP
        /selection=stepwise vif
;
run;
ods graphics off;

* Run the following ranking to determine the boundaries for binning.;
Title "Determine Binning";
proc rank data=IMP_MB out=RANKFILE  groups=10;
var TEAM_PITCHING_H;
ranks QUANT_TEAM_PITCHING_H;
run;

* Determine the upper bounds of the bins.;
proc means data=RANKFILE max;
class QUANT_TEAM_PITCHING_H;
var TEAM_PITCHING_H;
run;

* Run the following ranking to determine the boundaries for binning.;
proc rank data=IMP_MB out=RANKFILE  groups=10;
var TEAM_PITCHING_BB;
ranks QUANT_TEAM_PITCHING_BB;
run;

* Determine the upper bounds of the bins.;
proc means data=RANKFILE max;
class QUANT_TEAM_PITCHING_BB;
var TEAM_PITCHING_BB;
run;

* Run the following ranking to determine the boundaries for binning.;
proc rank data=IMP_MB out=RANKFILE  groups=10;
var TEAM_PITCHING_SO;
ranks QUANT_TEAM_PITCHING_SO;
run;

* Determine the upper bounds of the bins.;
proc means data=RANKFILE max;
class QUANT_TEAM_PITCHING_SO;
var TEAM_PITCHING_SO;
run;

* Create another data set that removes some outliers using binning.;
* The bounderies used were obtained above.  I only binned the 3 variables;
* that appeared to have the largest outlier influences.;
data IMP_OUTLIERCOR_MB2;
        set IMP_MB;

        if TEAM_BASERUN_SB < 24 then TEAM_BASERUN_SB = 24;
        if TEAM_BASERUN_SB > 438 then TEAM_BASERUN_SB = 438;

        if TEAM_BASERUN_CS < 18 then TEAM_BASERUN_CS = 18;
        if TEAM_BASERUN_CS > 125 then TEAM_BASERUN_CS = 125;

        if TEAM_PITCHING_H < 1355 then QUANT_TEAM_PITCHING_H = 1;
        else if TEAM_PITCHING_H <= 1399 then QUANT_TEAM_PITCHING_H = 2;
        else if TEAM_PITCHING_H <= 1438 then QUANT_TEAM_PITCHING_H = 3;
        else if TEAM_PITCHING_H <= 1478 then QUANT_TEAM_PITCHING_H = 4;
```

```
        else if TEAM_PITCHING_H <= 1517 then QUANT_TEAM_PITCHING_H = 5;
        else if TEAM_PITCHING_H <= 1561 then QUANT_TEAM_PITCHING_H = 6;
        else if TEAM_PITCHING_H <= 1636 then QUANT_TEAM_PITCHING_H = 7;
        else if TEAM_PITCHING_H <= 1749 then QUANT_TEAM_PITCHING_H = 8;
        else if TEAM_PITCHING_H <= 2059 then QUANT_TEAM_PITCHING_H = 9;
        else QUANT_TEAM_PITCHING_H = 10;

        if TEAM_PITCHING_BB < 417 then QUANT_TEAM_PITCHING_BB = 1;
        else if TEAM_PITCHING_BB <= 459 then QUANT_TEAM_PITCHING_BB = 2;
        else if TEAM_PITCHING_BB <= 488 then QUANT_TEAM_PITCHING_BB = 3;
        else if TEAM_PITCHING_BB <= 513 then QUANT_TEAM_PITCHING_BB = 4;
        else if TEAM_PITCHING_BB <= 536 then QUANT_TEAM_PITCHING_BB = 5;
        else if TEAM_PITCHING_BB <= 562 then QUANT_TEAM_PITCHING_BB = 6;
        else if TEAM_PITCHING_BB <= 594 then QUANT_TEAM_PITCHING_BB = 7;
        else if TEAM_PITCHING_BB <= 632 then QUANT_TEAM_PITCHING_BB = 8;
        else if TEAM_PITCHING_BB <= 693 then QUANT_TEAM_PITCHING_BB = 9;
        else QUANT_TEAM_PITCHING_BB = 10;

        if TEAM_PITCHING_SO < 494 then QUANT_TEAM_PITCHING_SO = 1;
        else if TEAM_PITCHING_SO <= 583 then QUANT_TEAM_PITCHING_SO = 2;
        else if TEAM_PITCHING_SO <= 660 then QUANT_TEAM_PITCHING_SO = 3;
        else if TEAM_PITCHING_SO <= 743 then QUANT_TEAM_PITCHING_SO = 4;
        else if TEAM_PITCHING_SO <= 813 then QUANT_TEAM_PITCHING_SO = 5;
        else if TEAM_PITCHING_SO <= 862 then QUANT_TEAM_PITCHING_SO = 6;
        else if TEAM_PITCHING_SO <= 924 then QUANT_TEAM_PITCHING_SO = 7;
        else if TEAM_PITCHING_SO <= 994 then QUANT_TEAM_PITCHING_SO = 8;
        else if TEAM_PITCHING_SO <= 1092 then QUANT_TEAM_PITCHING_SO = 9;
        else QUANT_TEAM_PITCHING_SO = 5;

        if TEAM_FIELDING_E < 86 then TEAM_FIELDING_E = 86;
        if TEAM_FIELDING_E > 1237 then TEAM_FIELDING_E = 1237;
run;


* MODEL 3;
* Run linear regression on the transformed data using stepwise selection;
ods graphics on;
Title "MODEL 3 - backward using binning data";
proc reg data=IMP_OUTLIERCOR_MB2 plots=diagnostics(stats=(default aic sbc));
model TARGET_WINS =
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_HR
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_H
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
        M_TEAM_BATTING_SO
        M_TEAM_BASERUN_SB
        M_TEAM_BASERUN_CS
        M_TEAM_PITCHING_SO
        M_TEAM_FIELDING_DP
        QUANT_TEAM_PITCHING_H
        QUANT_TEAM_PITCHING_BB
        QUANT_TEAM_PITCHING_SO
        /selection=backward vif
;
run;
ods graphics off;


* MODEL 4;
* Run linear regression on the transformed data using backward selection;
ods graphics on;
Title "MODEL 4 - backward using bining data - Reduced Set 1";
```

```
proc reg data=IMP_OUTLIERCOR_MB2 plots=diagnostics(stats=(default aic sbc));
model TARGET_WINS =
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_HR
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
        M_TEAM_BATTING_SO
        M_TEAM_BASERUN_SB
        M_TEAM_BASERUN_CS
        M_TEAM_FIELDING_DP
        QUANT_TEAM_PITCHING_H
        QUANT_TEAM_PITCHING_BB
        QUANT_TEAM_PITCHING_SO
        /selection=backward vif
;
run;
ods graphics off;

* Look at the relationship between TEAM_BATTING_2B and;
* TARGET_WINS.;
ods graphics on;
PROC SGPLOT DATA=&MBFILE.;
        LOESS X= TEAM_BATTING_2B Y= TARGET_WINS / NOMARKERS;
        REG X= TEAM_BATTING_2B Y= TARGET_WINS;
        Title2 "TEAM_FIELDING_DP and TARGET_WINS Scatter Plot with LOESS and Regression";
run;
ods graphics off;

* MODEL 5;
* Run linear regression on the data using backward selection;
* and reduce the number of variables to select from.;
ods graphics on;
Title "MODEL 5 - backward using transformed data - Reduced Set 2";
proc reg data=IMP_OUTLIERCOR_MB2 plots=diagnostics(stats=(default aic sbc));
model TARGET_WINS =
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_HR
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_HR
        TEAM_PITCHING_BB
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
        M_TEAM_BATTING_SO
        M_TEAM_BASERUN_SB
        M_TEAM_BASERUN_CS
        M_TEAM_FIELDING_DP
        QUANT_TEAM_PITCHING_H
        /selection=backward vif
;
run;

ods graphics off;
* Run the GLM using the variables chosen in Model 4.;
ods graphics on;
Title "MODEL BB - Using GLM";
proc glm data=IMP_OUTLIERCOR_MB2;
model TARGET_WINS =
```

```
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_HR
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
        M_TEAM_BATTING_SO
        M_TEAM_BASERUN_SB
        M_TEAM_FIELDING_DP
        QUANT_TEAM_PITCHING_H
        QUANT_TEAM_PITCHING_BB
        QUANT_TEAM_PITCHING_SO
;
run;
ods graphics off;

* Run GENMOD using the variables chosen in Model 4.;
ods graphics on;
Title "MODEL BB - Using GENMOD";
proc genmod data=IMP_OUTLIERCOR_MB2;
model TARGET_WINS =
        TEAM_BATTING_H
        TEAM_BATTING_2B
        TEAM_BATTING_3B
        TEAM_BATTING_BB
        TEAM_BATTING_SO
        TEAM_BASERUN_SB
        TEAM_BASERUN_CS
        TEAM_PITCHING_HR
        TEAM_PITCHING_SO
        TEAM_FIELDING_E
        TEAM_FIELDING_DP
        M_TEAM_BATTING_SO
        M_TEAM_BASERUN_SB
        M_TEAM_FIELDING_DP
        QUANT_TEAM_PITCHING_H
        QUANT_TEAM_PITCHING_BB
        QUANT_TEAM_PITCHING_SO
        / link=identity dist=normal
;
run;
ods graphics off;
```

Scoring Program

```
*******;
* Score code for Moneyball Test Data;
*;
* Jeffrey Vagg;
* PREDICT 411-Sec55;
*******;

* Setup some variables to use;
%let ME = jeffreyvagg2015;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT01/HW;

* Configure the location of the data files for the assignment.;
libname HW "&PATH." access=readonly;

%let MBFILE = HW.moneyball_test;

*****;
* Create an imputed dataset with missing values filled in.  The;
* median value for the attributes kept was used.;
```

```
*****;
data SCOREFILE;
	set &MBFILE.;

	* Drop TEAM_BATTING_HBP since the number of missing values is very;
	* high.;
	drop TEAM_BATTING_HBP;

	* Impute missing values using the median values.  Also add a flag;
	* variable that indicates the variable was imputed.  This is done;
	* for the 5 variables that were kept and had missing values.;
	* The median values were obtained from the output of the proc means;
	* done above.;
	M_TEAM_BATTING_SO = 0;
	if missing(TEAM_BATTING_SO) then do;
		TEAM_BATTING_SO = 750;
		M_TEAM_BATTING_SO = 1;
	end;

	M_TEAM_BASERUN_SB = 0;
	if missing(TEAM_BASERUN_SB) then do;
		TEAM_BASERUN_SB = 101;
		M_TEAM_BASERUN_SB = 1;
	end;

	M_TEAM_BASERUN_CS = 0;
	if missing(TEAM_BASERUN_CS) then do;
		TEAM_BASERUN_CS = 49;
		M_TEAM_BASERUN_CS = 1;
	end;

	M_TEAM_PITCHING_SO = 0;
	if missing(TEAM_PITCHING_SO) then do;
		TEAM_PITCHING_SO=814;
		M_TEAM_PITCHING_SO = 1;
	end;

	M_TEAM_FIELDING_DP = 0;
	if missing(TEAM_FIELDING_DP) then do;
		TEAM_FIELDING_DP = 149;
		M_TEAM_FIELDING_DP = 1;
	end;

	* Adjust for outliers based on the 1st and 99th percentiles;
	* and binning.;
	if TEAM_BASERUN_SB < 24 then TEAM_BASERUN_SB = 24;
	if TEAM_BASERUN_SB > 438 then TEAM_BASERUN_SB = 438;

	if TEAM_BASERUN_CS < 18 then TEAM_BASERUN_CS = 18;
	if TEAM_BASERUN_CS > 125 then TEAM_BASERUN_CS = 125;

	if TEAM_PITCHING_H < 1355 then QUANT_TEAM_PITCHING_H = 1;
	else if TEAM_PITCHING_H <= 1399 then QUANT_TEAM_PITCHING_H = 2;
	else if TEAM_PITCHING_H <= 1438 then QUANT_TEAM_PITCHING_H = 3;
	else if TEAM_PITCHING_H <= 1478 then QUANT_TEAM_PITCHING_H = 4;
	else if TEAM_PITCHING_H <= 1517 then QUANT_TEAM_PITCHING_H = 5;
	else if TEAM_PITCHING_H <= 1561 then QUANT_TEAM_PITCHING_H = 6;
	else if TEAM_PITCHING_H <= 1636 then QUANT_TEAM_PITCHING_H = 7;
	else if TEAM_PITCHING_H <= 1749 then QUANT_TEAM_PITCHING_H = 8;
	else if TEAM_PITCHING_H <= 2059 then QUANT_TEAM_PITCHING_H = 9;
	else QUANT_TEAM_PITCHING_H = 10;

	if TEAM_PITCHING_BB < 417 then QUANT_TEAM_PITCHING_BB = 1;
	else if TEAM_PITCHING_BB <= 459 then QUANT_TEAM_PITCHING_BB = 2;
	else if TEAM_PITCHING_BB <= 488 then QUANT_TEAM_PITCHING_BB = 3;
	else if TEAM_PITCHING_BB <= 513 then QUANT_TEAM_PITCHING_BB = 4;
	else if TEAM_PITCHING_BB <= 536 then QUANT_TEAM_PITCHING_BB = 5;
	else if TEAM_PITCHING_BB <= 562 then QUANT_TEAM_PITCHING_BB = 6;
	else if TEAM_PITCHING_BB <= 594 then QUANT_TEAM_PITCHING_BB = 7;
	else if TEAM_PITCHING_BB <= 632 then QUANT_TEAM_PITCHING_BB = 8;
	else if TEAM_PITCHING_BB <= 693 then QUANT_TEAM_PITCHING_BB = 9;
	else QUANT_TEAM_PITCHING_BB = 10;
```

```
        if TEAM_PITCHING_SO < 494 then QUANT_TEAM_PITCHING_SO = 1;
        else if TEAM_PITCHING_SO <= 583 then QUANT_TEAM_PITCHING_SO = 2;
        else if TEAM_PITCHING_SO <= 660 then QUANT_TEAM_PITCHING_SO = 3;
        else if TEAM_PITCHING_SO <= 743 then QUANT_TEAM_PITCHING_SO = 4;
        else if TEAM_PITCHING_SO <= 813 then QUANT_TEAM_PITCHING_SO = 5;
        else if TEAM_PITCHING_SO <= 862 then QUANT_TEAM_PITCHING_SO = 6;
        else if TEAM_PITCHING_SO <= 924 then QUANT_TEAM_PITCHING_SO = 7;
        else if TEAM_PITCHING_SO <= 994 then QUANT_TEAM_PITCHING_SO = 8;
        else if TEAM_PITCHING_SO <= 1092 then QUANT_TEAM_PITCHING_SO = 9;
        else QUANT_TEAM_PITCHING_SO = 5;

        if TEAM_FIELDING_E < 86 then TEAM_FIELDING_E = 86;
        if TEAM_FIELDING_E > 1237 then TEAM_FIELDING_E = 1237;

        * Model chosen;
        P_TARGET_WINS =
                17.83319
        +       0.05915         * TEAM_BATTING_H
        +       -0.02334        * TEAM_BATTING_2B
        +       0.06546         * TEAM_BATTING_3B
        +       0.01501         * TEAM_BATTING_BB
        +       -0.02436        * TEAM_BATTING_SO
        +       0.05747         * TEAM_BASERUN_SB
        +       -0.04586        * TEAM_BASERUN_CS
        +       0.06158         * TEAM_PITCHING_HR
        +       0.00174         * TEAM_PITCHING_SO
        +       -0.06141        * TEAM_FIELDING_E
        +       -0.09303        * TEAM_FIELDING_DP
        +       7.29980         * M_TEAM_BATTING_SO
        +       35.93190        * M_TEAM_BASERUN_SB
        +       4.76373         * M_TEAM_FIELDING_DP
        +       -0.94875        * QUANT_TEAM_PITCHING_H
        +       0.37448         * QUANT_TEAM_PITCHING_BB
        +       0.92921         * QUANT_TEAM_PITCHING_SO;

        * Verify there are no missing values and the;
        * predicted wins are within the 1st and 99th;
        * percentiles to ensure no outliers.;
        if missing(P_TARGET_WINS) then P_TARGET_WINS = 81;
        if P_TARGET_WINS < 38 then P_TARGET_WINS = 38;
        if P_TARGET_WINS > 114 then P_TARGET_WINS = 114;

        * Just keep the two columns required for submission.;
        keep INDEX;
        keep P_TARGET_WINS;
run;

* Set the destnation for the output file.;
libname scorelib "/home/jeffreyvagg2015/MoneyBall";

* Write the score file.;
data scorelib.JeffreyVaggMoneyballScore;
set SCOREFILE;
run;
```