Assignment 1: Moneyball Baseball Problem

MSPA PREDICT 411-DL-SEC56

Darryl Buswell

1 Introduction

This document presents results of the first assignment for the Masters of Science in Predictive Analytics course: PREDICT 411. This assessment required the student to build an Ordinary Least Squares (OLS) regression model which is able to predict the number of wins for a baseball team. In order to find the 'best' predictive model, this assessment manually specified one regression model and employed two automated variable selection techniques with the resultant model specifications assessed over a range of performance criteria. Automated variable selection techniques included the adjusted R Squared and stepwise variable selection methods, which were assessed using metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE), Adjusted R-square, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). A final model was selected and used to generate a range of prediction labels based on a test set of data.

As a bonus, this assessment employed Principal Component Analysis (PCA) as a method of dimensionality reduction for regression analysis. PCA is initially performed on the dataset in order to derive a set of components. A regression model is then fitted which makes use of the principle components for its set of predictor variables. Finally, an assessment of the model is made, with a focus on whether using principle components for predictor variables results in an improvement in either goodness-of-fit or model performance.

2 Data

The dataset contains variables which focus on performance statistics for baseball teams over the years 1871 to 2006. Each record details the performance of a team for any given year, with all statistics adjusted to match the performance of a 162 game season. For this assessment, two datasets have been provided, one is a training set of data, which is comprised of 2,276 observations. The second is a testing data set of data, which is comprised of 259 observations.

At a first pass, it seems the dataset has quite a large amount of scope. There are variables tracking a number of gameplay attributes, including those related to batter performance, pitcher performance and gameplay outcomes. The below table shows a list of variables included in the original dataset. A proposed effect on the response variable is also shown for each.

Table 2.1: Variable Descriptions

Variable	Definition	Proposed Effect on Wins
TARGET_WINS	Target wins	
TEAM_BATTING_H	Base Hits by batters	Positive
TEAM_BATTING_2B	Doubles by batters	Positive
TEAM_BATTING_3B	Triples by batters	Positive
TEAM_BATTING_HR	Homeruns by batters	Positive
TEAM_BATTING_BB	Walks by batters	Positive
TEAM_BATTING_SO	Strikeouts by batters	Negative
TEAM_BASERUN_SB	Stolen bases	Positive
TEAM_BASERUN_CS	Caught stealing	Negative
TEAM_BATTING_HBP	Batters hit by pitch	Positive
TEAM_PITCHING_H	Hits allowed	Negative
TEAM_PITCHING_HR	Homeruns allowed	Negative
TEAM_PITCHING_BB	Walks allowed	Negative

Variable	Definition	Proposed Effect on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive
TEAM_FIELDING_E	Errors	Negative
TEAM_FIELDING_DP	Double Plays	Positive

It is worth noting the dataset has quite a long span (from 1871 to 2006). As such, the dataset may indeed suffer from consistency issues. For example, there may have been changes as to how the data for each variable was collected over this period, or even changes to the nature of what each variable is to represent. Unfortunately, the dataset does not include a variable to allow observations to be identified by time, which limits our ability to deal with this issue.

3 Data Exploration

Prior to performing any regression analysis, a number of data exploration routines are conducted. These routines allow us to gain an understanding of any potential limitations of the dataset, including identifying variables which have missing observations, outlier observations, or those variables which may benefit from transformation.

3.1 Univariate Data Analysis

A review of the dataset shows that each variable is of continuous type with varying statistical properties. Summary statistics for each variable is shown in the table below.

Table 3.1.1: Data Statistics

Variable	Minimum	Maximum	Mean	Std Dev	N Miss	N
TARGET_WINS	0	146	80.7908612	15.7521525	0	2276
TEAM_BATTING_H	891	2554	1469.27	144.5911954	0	2276
TEAM_BATTING_2B	69	458	241.2469244	46.8014146	0	2276
TEAM_BATTING_3B	0	223	55.25	27.938557	0	2276
TEAM_BATTING_HR	0	264	99.6120387	60.546872	0	2276
TEAM_BATTING_BB	0	878	501.5588752	122.6708615	0	2276
TEAM_BATTING_SO	0	1399	735.6053358	248.5264177	102	2174
TEAM_BASERUN_SB	0	697	124.7617716	87.791166	131	2145
TEAM_BASERUN_CS	0	201	52.8038564	22.9563376	772	1504
TEAM_BATTING_HBP	29	95	59.3560209	12.9671225	2085	191
TEAM_PITCHING_H	1137	30132	1779.21	1406.84	0	2276
TEAM_PITCHING_HR	0	343	105.698594	61.2987469	0	2276
TEAM_PITCHING_BB	0	3645	553.0079086	166.3573617	0	2276
TEAM_PITCHING_SO	0	19278	817.7304508	553.0850315	102	2174
TEAM_FIELDING_E	65	1898	246.4806678	227.7709724	0	2276
TEAM_FIELDING_DP	52	228	146.3879397	26.2263853	286	1990

What becomes immediately obvious is that a number of variables suffer from a large amount of missing observations. TEAM_BATTING_HBP in particular, includes only 191 observations, which is less than 10% of the total number of observations recorded for the wider dataset. While the remaining variables with missing observations may benefit from some form of imputation, this variable will instead be excluded from the analysis completely.

Figure 3.1.1 Histogram: Team Batting 2B

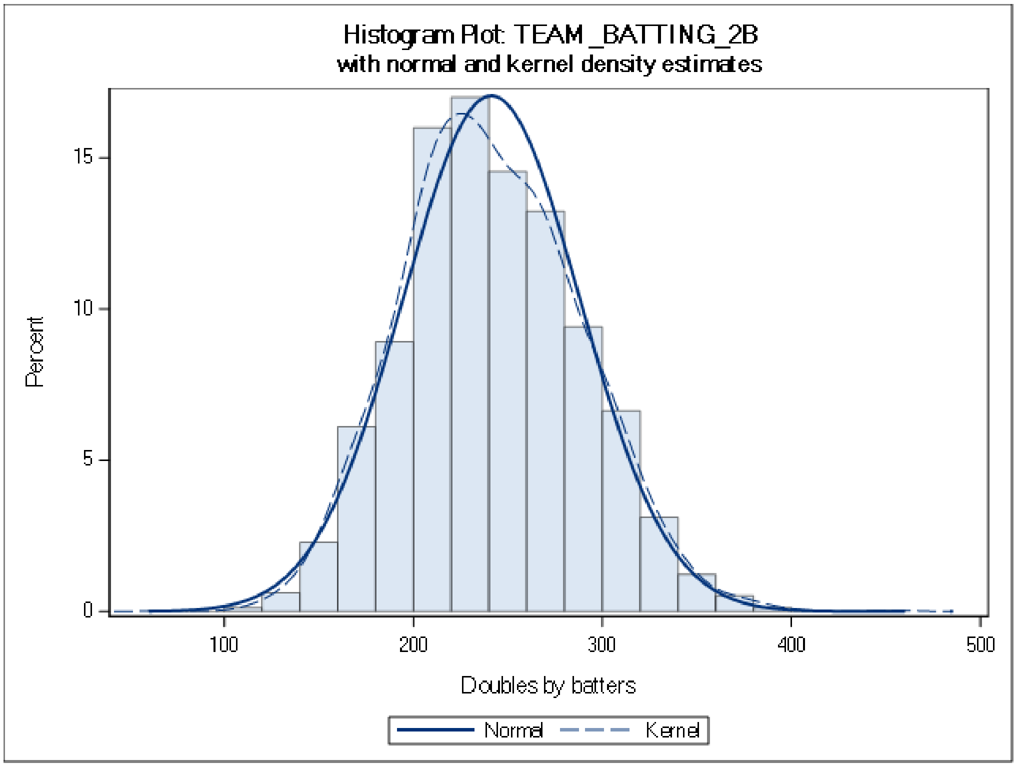


Figure 3.1.2 Box Plot: Team Batting 2B

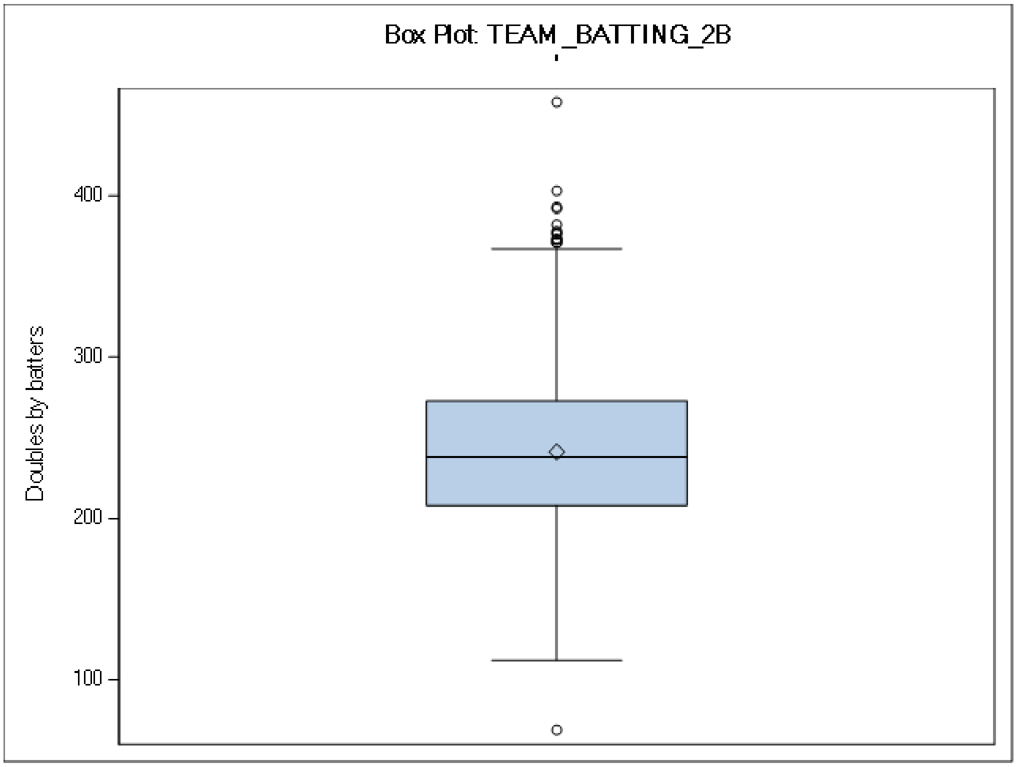


Figure 3.1.3 Histogram: Team Batting 3B

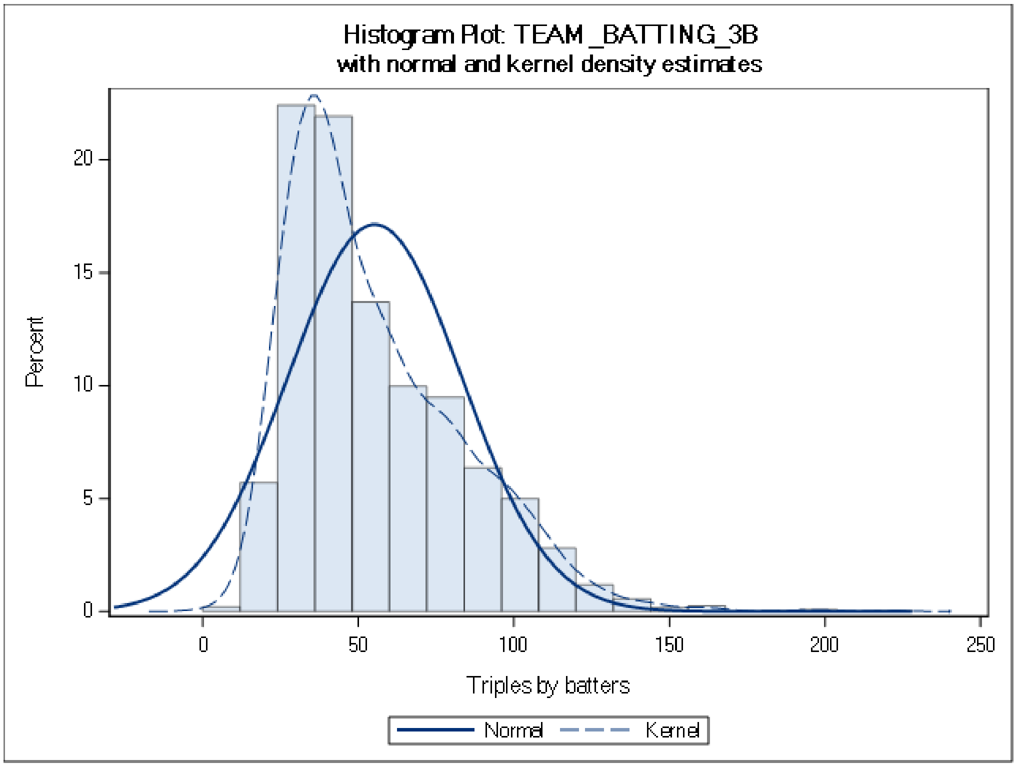


Figure 3.1.4 Box Plot: Team Batting 3B

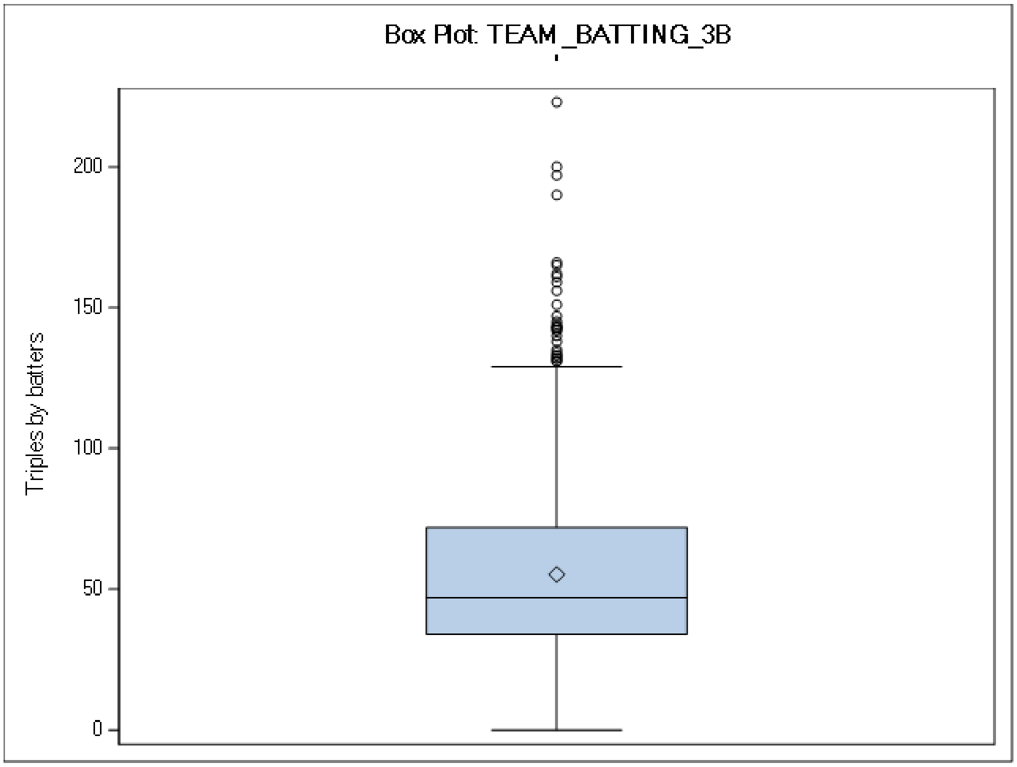


Figure 3.1.5 Histogram: Team Batting HR

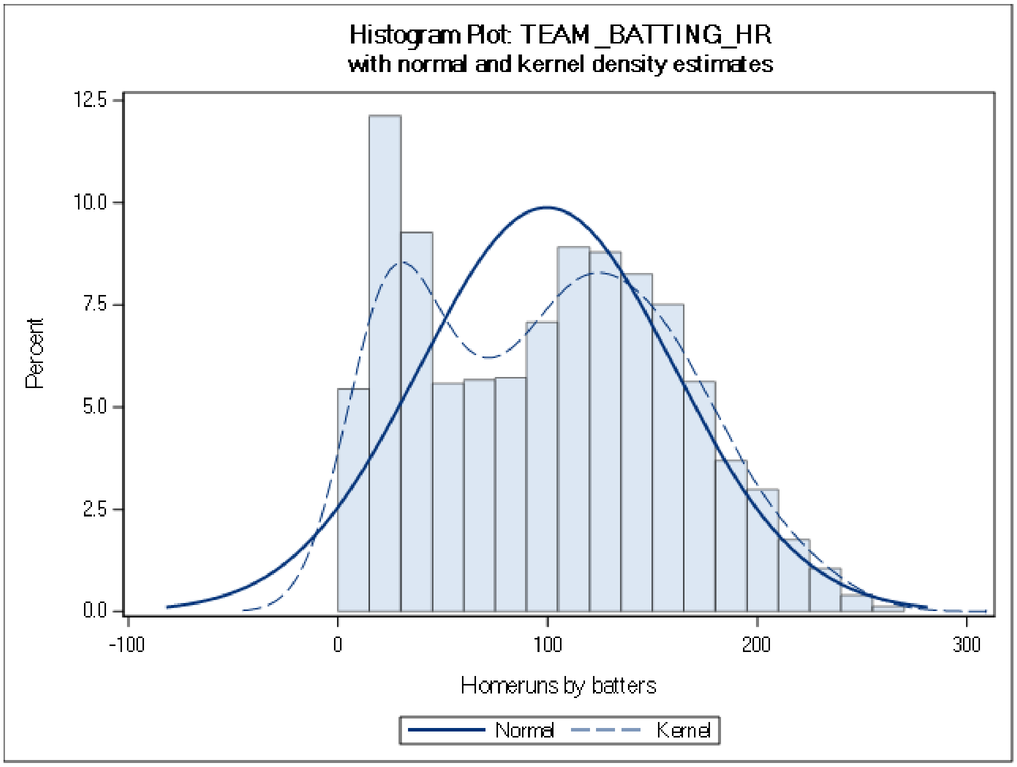
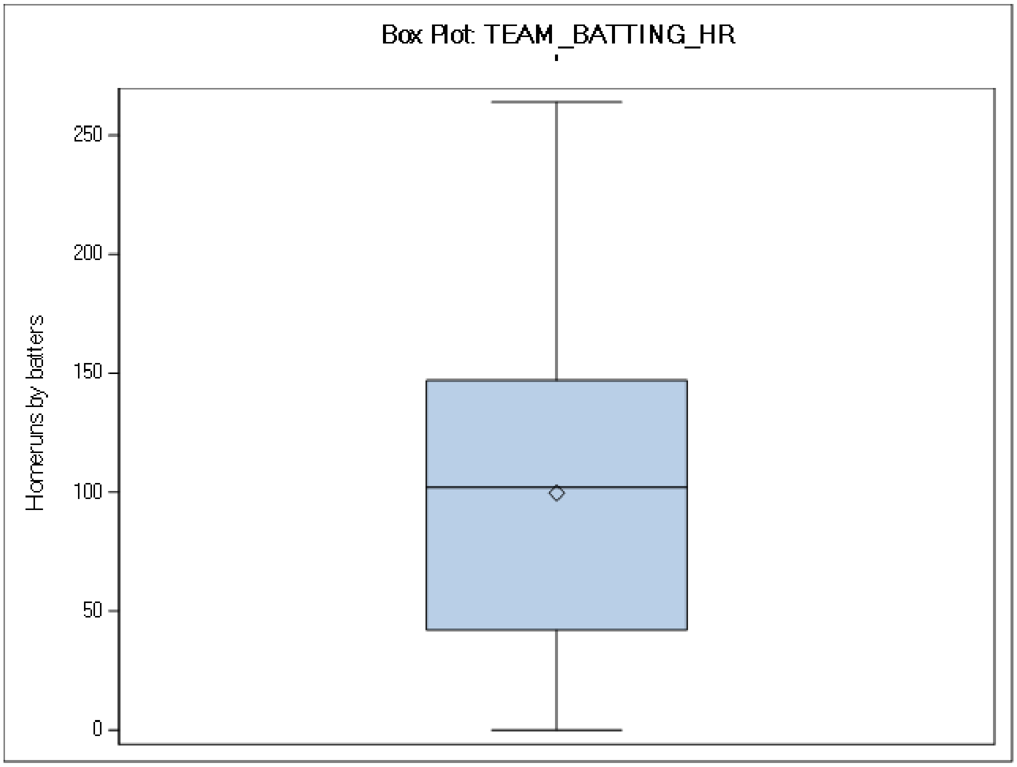


Figure 3.1.6 Box Plot: Team Batting HR



We can also use a combination of histogram and box plots in order to gain a greater understanding of each variable included within the dataset. Histogram and box plots were generated and reviewed for all variables, however with plots for only three variables were selected for further discussion. We can see that TEAM_BATTING_2B, has a normal shaped distribution, and minimal outliers. However, there are variables within the dataset such as TEAM_BATTING_3B which have a heavy skew. In this case, the positive skewed distribution also carries a number of outlier observations. Even more concerning are variables such as TEAM_BATTING_HR, which seem to have quite a non-normal shaped distribution. In this case, observations seem to collect around both a low and high value, resulting in a bimodal shaped distribution.

3.2 Bivariate Data Analysis

Since we intend on building a prediction model for target wins, we have an interest in those variables which have explanatory power over this variable. As such, we can use the SAS procedure 'corr' to see if there are any variables that have a high Pearson correlation coefficient and low *p*-value in relation to our response variable.

Table 3.2.1: Data Correlations

Variable	Correlation	Proposed Effect on Wins
TEAM_BATTING_H	0.38877	Positive
TEAM_BATTING_2B	0.28910	Positive
TEAM_BATTING_3B	0.14261	Positive
TEAM_BATTING_HR	0.17615	Positive
TEAM_BATTING_BB	0.23256	Positive
TEAM_BATTING_SO	-0.03175	Negative
TEAM_BASERUN_SB	0.13514	Positive
TEAM_BASERUN_CS	0.02240	Negative
TEAM_PITCHING_H	-0.10994	Negative
TEAM_PITCHING_HR	0.18901	Negative
TEAM_PITCHING_BB	0.12417	Negative
TEAM_PITCHING_SO	-0.07844	Positive
TEAM_FIELDING_E	-0.17648	Negative
TEAM_FIELDING_DP	-0.03485	Positive

None of the variables are reported to have a particularly strong positive or negative correlation coefficient with the response variable, with the greatest absolute correlation being reported by 'base hits by batters' and 'doubles by batters' at 0.39 and 0.29 respectively. We also observe that a number of variables polarity of correlation coefficient does not match the proposed effect on wins. We will take note of these variables, as they may have less justification for inclusion in any regressions which are manually specified.

We can also use scatter plots with a Locally Estimated Scatter Plot Smoother (LOESS) overlay to further explore the relationship between the response variable and each predictor variable. Scatter plots for a collection of variables against the response variable were generated and reviewed, however only two of these plots were selected for further discussion below. We can see from the plots that both TEAM_BASERUN_CS and TEAM_BATTING_HR do indeed have little positive or negative relation with TARGET_WINS. It is worth noting that while many variables such as TEAM_BASERUN_CS are prone to outlier observations which results in the appearance of a tight collection of observations on scatterplots, there are a selection of variables such as TEAM_BATTING_HR who suffer less from outliers which results in a more uniform pattern over scatter plots.

Figure 3.2.1 Scatter: Team Baserun CS

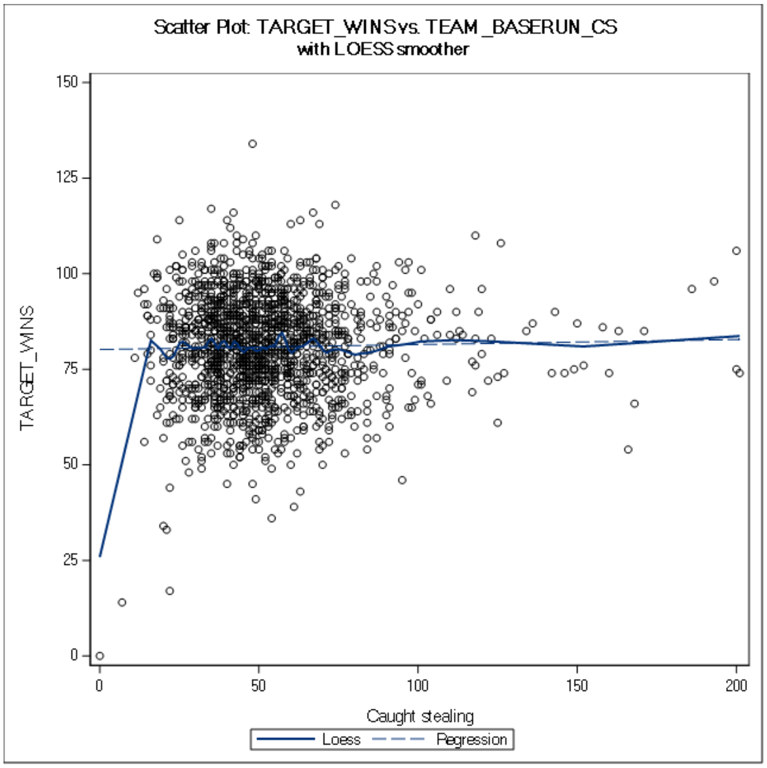
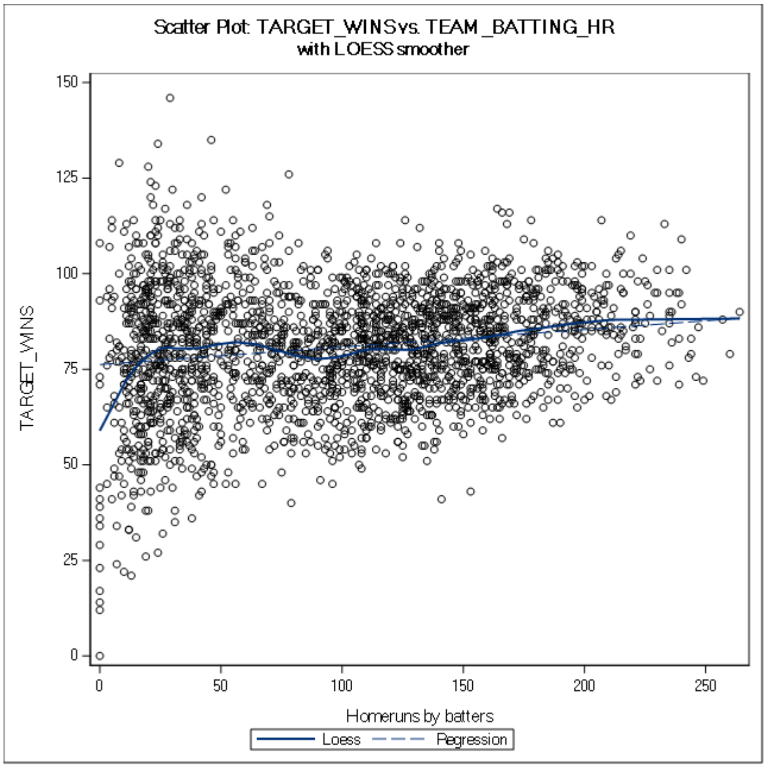


Figure 3.2.3 Scatter: Team Batting HR



4 Data Preparation

From an observation of each univariate and bivariate plot, we have identified that the majority of variables do in-fact suffer from outlier observations. In fact, only TEAM_BATTING_HR, TEAM_BATTING_SO and TEAM_PITCHING_HR were shown to have minimal outlier observations, with variables such as TEAM_PITCHING_H, TEAM_FIELDING_E and TEAM_BASERUN_SB having the greatest number of outliers. A review of the percentiles for each variable confirms this, with a rather large gap between min, max and the 1st and 99th percentile respectively. A summary of percentiles for each variable can be found in the table below.

Table 4.1: Quantiles Summary

Variable	Max	0.99	0.95	0.9	Q3	Med	Q1	0.1	0.05	0.01	Min
TARGET_WINS	146	114	104	100	92	82	71	61	54	38	0
TEAM_BATTING_H	2554	1950	1696	1636	1537.5	1454	1383	1315	1280	1188	891
TEAM_BATTING_2B	458	352	320	303	273	238	208	182	167	141	69
TEAM_BATTING_3B	223	134	108	96	72	47	34	27	23	17	0
TEAM_BATTING_HR	264	235	199	180	147	102	42	20	14	4	0
TEAM_BATTING_BB	878	755	671	635	580	512	451	363	246	79	0
TEAM_BATTING_SO	1399	1193	1104	1049	930	750	548	421	359	67	0
TEAM_BASERUN_SB	697	439	302	231	156	101	66	44	35	23	0
TEAM_BASERUN_CS	201	143	91	77	62	49	38	30	24	16	0
TEAM_PITCHING_H	30132	7093	2563	2059	1683	1518	1419	1356	1316	1244	1137
TEAM_PITCHING_HR	343	244	210	187	150	107	50	25	18	8	0
TEAM_PITCHING_BB	3645	924	757	694	611	536.5	476	417	377	237	0
TEAM_PITCHING_SO	19278	1474	1173	1095	968	813.5	615	490	420	205	0
TEAM_FIELDING_E	1898	1237	716	542	249.5	159	127	109	100	86	65
TEAM_FIELDING_DP	228	204	186	178	164	149	131	109	98	79	52

We do not know at this point whether these outlier observations have predictive power over TARGET_WINS. To understand this, one option may be to conduct a simple Ordinary Least Squares (OLS) for each variable against our response variable and observe the residuals of each. For this assessment however, we have elected to instead generate a trimmed copy of all variables, and will rely on automated variable selection routines to identify those variables which demonstrate the greatest predictive power. Such a method will help reduce the amount of selective bias introduced by the analyst.

From a review of both the univariate plots in the previous section as well as the quantile summaries above, we have elected to generate trimmed copies of each variable by their 99th percentile. Although the inclusion of additional trimmed variables is an option, at this stage we prefer to minimize the amount of adjusted variable copies introduced into the dataset. Trimmed variables include the suffix '_T99'.

Following introducing copies of trimmed variables, we then look towards imputing values for missing observations. By recalculating statistical measures for each variable after trimming, we are able to avoid imputing skewed values for those variables which have been trimmed into each variable. For this assessment, we generate new imputed variables based on that variable's median value. Note that in order to simplify the SAS logic used for this assessment, all variables will include the suffix '_IME', however only those variables shown to have missing observations in the previous section have actually received imputation.

As a final step in our data preparation routine, we perform a natural logarithm transformation of each variable. Variables which have been transformed include the suffix '_LN'. Such a transformation will help penalize extreme values and may provide an improved fit within subsequent regression models.

5 Model Development

For this section, we build three linear regression models. For the first ‘Subjective Model’, variables have been retained based our opinion of which would likely have the greatest relevance towards predicting wins. The remaining two models are specified according to automated variable selection techniques. The ‘Adjusted R-squared Selection Model’, which uses the maximum improvement technique for variable selection. And finally, the ‘Stepwise Selection Model’ which uses the stepwise selection technique for variable selection.

5.1 Model 1: Subjective Model

For the Subjective Model (Model_Subj), we include only those variables which were shown to produce a correlation coefficient against the response variable which was aligned with our proposed effect. We build on this by including missing variable flags for those variables which were shown to have missing variables, and outlier flags for those variables which were shown to have a large amount of outliers. Finally, we removed any variables from this specification which were found to be highly insignificant.

Parameter estimates for the Model_Subj are shown below.

Table 5.1.1: Subjective Model Parameter Estimates (Training Set)

Variable	DF	Par. Est.	Par. S.E.	t Value	$Pr >	t
Intercept	1	3.16454	5.617	0.56	0.5733	0
TEAM_BATTING_H_IME	1	0.04876	0.00413	11.82	<.0001	3.71652
TEAM_BATTING_2B_IME	1	-0.03895	0.01065	-3.66	0.0003	2.68623
TEAM_BATTING_3B_IME	1	0.08737	0.01949	4.48	<.0001	3.02755
TEAM_BATTING_HR_IME	1	0.06472	0.0113	5.73	<.0001	4.96995
TEAM_BATTING_BB_IME	1	0.02307	0.00387	5.96	<.0001	2.37525
TEAM_BATTING_SO_IME	1	-0.00891	0.00264	-3.37	0.0008	4.23806
TEAM_BASERUN_SB_IME	1	0.06231	0.00529	11.78	<.0001	2.02626
TEAM_PITCHING_H_IME	1	0.0012	0.00036487	3.28	0.0011	3.2279
TEAM_FIELDING_E_IME	1	-0.05495	0.00357	-15.4	<.0001	7.04413
TEAM_BATTING_SO_MF	1	9.90691	1.61891	6.12	<.0001	1.30858
TEAM_BASERUN_SB_MF	1	36.73709	2.23367	16.45	<.0001	3.02158
TEAM_BATTING_3B_OF	1	-4.02422	2.56594	-1.57	0.117	1.16292
TEAM_PITCHING_H_OF	1	3.98635	2.71559	1.47	0.1423	1.63274
TEAM_FIELDING_E_OF	1	4.8098	2.73651	1.76	0.079	1.41879

For Model_Subj, the majority of coefficient estimates have significant p-values at the 95% level, allowing us to reject the null hypothesis and conclude that each have non-zero coefficients. The only exceptions are the outlier flag variables, TEAM_BATTING_3B_OF, TEAM_PITCHING_H_OF and TEAM_FIELDING_E_OF. We also note that the polarity of estimate for the majority of coefficients seems reasonable, with the only exceptions being TEAM_BATTING_2B_IME and TEAM_PITCHING_H_IME. We also note the magnitude of the coefficient estimate for each of the non-flag variables is quite small, suggesting only a marginal expected change in wins based on a unit change in each. Finally, we note that the VIF for the majority of predictor variables is less than five suggesting little to moderate correlation between predictors.

Goodness-of-fit information for Model_Subj is shown below.

Table 5.1.2: Subjective Model Analysis of Variance (Training Set)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	166986	11928	79.40	<.0001
Error	1547	232393	150.22183		

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Corrected Total	1561	399379			

The model has reported a large F-value suggesting that the observations and regression differ from the grand mean. Likewise, the F-value has a highly significant p-value under the null hypothesis that there is no linear relationship between the predictor and response variable.

Model performance statistics over the training set for Model_Subj are shown below.

Table 5.1.3: Subjective Model Performance Metrics (Training Set)

Measure	Value	Measure	Value
MSE	150.22183	R-Square	0.4181
MAE	9.71342	Adj R-Sq	0.4128
Root MSE	12.25650	C(p)	15.0000
Dependent Mean	80.57875	AIC	7843.8481
Coeff Var	15.21059	BIC	7846.1388

The R-square value above suggests that Model_Subj explains ~42% of the variability in TARGET_WINS using each of the included predictor variables. The adjusted R-squared value also indicates a similar level of explanatory power. The AIC and BIC are also reported above. Both AIC and BIC form a model selection criteria which look to assess goodness of fit of the model. These metrics will be used to assess the above model against models formed from alternative automated selection techniques below.

Model performance statistics over the training set for Model_Subj are shown below.

Table 5.1.4: Subjective Model Performance Metrics (Test Set)

Measure	Value	Measure	Value
MSE	146.95376	R-Square	0.3646
MAE	9.29372	Adj R-Sq	0.3770
Root MSE	12.12245	C(p)	15.0000
Dependent Mean	81.25490	AIC	3577.7844
Coeff Var	14.91904	BIC	3580.4273

From the table above, we can see that applying the same model specification to the test set of data shows a slight reduction in MSE, MAE and likewise, a reduction in reported R-Square and adjusted R-Square values. Although we can also see a reduction in AIC and BIC, these metrics are better suited to compare alternative model specifications which have been assessed against the test set of data. Generally, performance metrics for Model_Subj have remained fairly consistent between the training and test sets, suggesting that the model is able to generalize over the test set of data.

5.2 Model 2: AdjR2 Selection Model

The adjusted R-squared selection technique finds subsets of independent variables that 'best' predict the dependent variable. 'Best' is defined by this technique as the model which produces the highest adjusted R-squared value (Inc 2016). This technique is applied to the training set of data, with the resulting model (Model_AdjR2) shown below.

Parameter estimates for Model_AdjR2 are shown below.

Table 5.2.1: Model AdjR2 Parameter Estimates (Training Set)

Variable	DF	Par. Est.	Par. S.E.	t Value	$Pr >	t
Intercept	1	4.03808	86.73687	0.05	0.9629	0
TEAM_BASERUN_CS_OF	1	3.39374	1.01873	3.33	0.0009	2.6933
TEAM_BASERUN_SB_MF	1	33.98348	2.12377	16	<.0001	2.98107
TEAM_BATTING_SO_MF	1	7.44002	1.73068	4.3	<.0001	1.63211
TEAM_FIELDING_DP_MF	1	6.18056	1.7548	3.52	0.0004	3.69955
TEAM_BASERUN_SB_IME	1	0.0485	0.00553	8.76	<.0001	2.41932
TEAM_BATTING_2B_T99_IME	1	-0.25625	0.06227	-4.12	<.0001	85.18334
TEAM_BATTING_3B_T99_IME	1	0.1575	0.01908	8.26	<.0001	2.71637
TEAM_BATTING_BB_IME	1	0.02836	0.00365	7.77	<.0001	2.30183
TEAM_BATTING_H_IME	1	0.06498	0.00412	15.77	<.0001	4.04579
TEAM_PITCHING_HR_IME	1	0.04503	0.00987	4.56	<.0001	4.25924
TEAM_PITCHING_SO_T99_IME	1	-0.04225	0.00806	-5.24	<.0001	35.47802
TEAM_BATTING_2B_T99_IME_LN	1	49.84735	14.86247	3.35	0.0008	86.28656
TEAM_BATTING_H_T99_IME_LN	1	-26.9484	7.31153	-3.69	0.0002	3.9306
TEAM_FIELDING_DP_IME_LN	1	-16.05367	2.26834	-7.08	<.0001	1.96833
TEAM_FIELDING_E_IME_LN	1	-23.64552	1.3426	-17.61	<.0001	7.65474
TEAM_PITCHING_SO_T99_IME_LN	1	25.16199	6.22459	4.04	<.0001	37.675

For Model_AdjR2, all coefficient estimates have significant p-values at the 95% level, allowing us to reject the null hypothesis for each estimate and conclude that each have non-zero coefficients. We also note that the polarity of estimate for many of coefficients seems reasonable, with the exceptions being TEAM_BATTING_2B_T99_IME, TEAM_BATTING_BB_IME, TEAM_PITCHING_H, TEAM_FIELDING_DP_IME_LN and TEAM_PITCHING_SO_T99_IME_LN. Unfortunately, a number of variables are included which have VIF's greater than 10, suggesting a high amount of correlation with the remaining predictors.

Goodness-of-fit information for Model_AdjR2 is shown below.

Table 5.2.2: Model AdjR2 Analysis of Variance (Training Set)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	186712	11669	84.78	<.0001
Error	1545	212667	137.64865		
Corrected Total	1561	399379			

The model has reported a large F-value suggesting that the observations and regression differ from the grand mean. Likewise, the F-value has a highly significant p-value under the null hypothesis that there is no linear relationship between the predictor and response variable.

Model performance statistics over the training set for Model_AdjR2 are shown below.

Table 5.2.3: Model AdjR2 Performance Metrics (Training Set)

Measure	Value	Measure	Value
MSE	137.64865	R-Square	0.4675
MAE	9.25776	Adj R-Sq	0.462
Root MSE	11.73238	C(p)	17
Dependent Mean	80.57875	AIC	7709.2951
Coeff Var	14.56014	BIC	7711.669

Measure	Value	Measure	Value

The R-square value above suggests that Model_AdjR2 explains ~47% of the variability in TARGET_WINS using each of the included predictor variables. The adjusted R-squared value for this specification is slightly higher (superior) than that reported for Model_Subj (0.4128), and its AIC and BIC are slightly lower (superior) than those reported for Model_Subj (7843.8481 and 7846.1388 respectively). These metrics suggest that Model_AdjR2 may have a slightly better fit over the training set of data compared to Model_Subj.

Model performance statistics over the test set for Model_AdjR2 are shown below.

Table 5.2.4: Model AdjR2 Performance Metrics (Test Set)

Measure	Value	Measure	Value
MSE	137.76292	R-Square	0.4177
MAE	9.03843	Adj R-Sq	0.4043
Root MSE	11.73725	C(p)	17
Dependent Mean	81.2549	AIC	3533.6258
Coeff Var	14.44497	BIC	3536.4539

Performance metrics for Model_AdjR2 have remained fairly consistent between the training and test sets. This suggests that the model is able to generalize over the test set of data.

5.3 Model 3: Stepwise Selection

The stepwise technique is a modification of the forward-selection technique. It differs in that variables already in the model do not necessarily remain in the model (D. Montgomery 2012). For this technique, we elected to use a SLENTRY value of 0.02, which indicates that variables should only be added to the specification if they have a significance level (p-value) of less than 2%, and also elected to use a SLSTAY value of 0.02, which indicates that variables should not be removed from the specification if they have a significance level (p-value) less than 2%. This technique is applied to the training set of data, with the resulting model (Model_S) shown below.

Parameter estimates for Model_S are shown below.

Table 5.3.1: Stepwise Selection Model Parameter Estimates (Training Set)

Variable	DF	Par. Est.	Par. S.E.	t Value	$Pr >	t
Intercept	1	266.4722	21.59628	12.34	<.0001	0
TEAM_BASERUN_CS_OF	1	3.44889	1.03252	3.34	0.0009	2.69668
TEAM_BASERUN_SB_IME	1	0.05721	0.00546	10.47	<.0001	2.2965
TEAM_BASERUN_SB_MF	1	34.88631	2.05252	17	<.0001	2.71393
TEAM_BATTING_3B_T99_IME	1	0.15324	0.01927	7.95	<.0001	2.70108
TEAM_BATTING_BB_IME	1	0.05204	0.00626	8.32	<.0001	6.6007
TEAM_BATTING_H_IME	1	0.05862	0.00394	14.89	<.0001	3.59982
TEAM_BATTING_H_T99_IME	1	-0.02893	0.00476	-6.08	<.0001	3.59958
TEAM_BATTING_SO_IME	1	-0.01576	0.00244	-6.46	<.0001	3.83262
TEAM_BATTING_SO_MF	1	7.6899	1.69006	4.55	<.0001	1.51701
TEAM_FIELDING_DP_IME_LN	1	-14.46299	2.21211	-6.54	<.0001	1.82458
TEAM_FIELDING_DP_OF	1	4.45505	1.50274	2.96	0.0031	2.89485
TEAM_FIELDING_E_IME_LN	1	-22.62866	1.41822	-15.96	<.0001	8.32515
TEAM_PITCHING_BB_T99_IME_LN	1	-12.41998	3.06748	-4.05	<.0001	4.06087
TEAM_PITCHING_HR_T99_IME	1	0.05559	0.00997	5.58	<.0001	3.80833

For Model_S, all coefficient estimates have significant p-values at the 95% level, allowing us to reject the null hypothesis for each estimate and conclude that each have non-zero coefficients. We also note that the polarity of estimate for the majority of coefficients seems reasonable, with the only exceptions being TEAM_BATTING_H_IME, TEAM_PITCHING_BB_T99_IME_LN and TEAM_PITCHING_HR_T99_IME. Finally, we note that the VIF for the majority of predictor variables is less than five suggesting little to moderate correlation between predictors.

Goodness-of-fit information for Model_S is shown below.

Table 5.3.2: Stepwise Selection Model Analysis of Variance (Training Set)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	180908	12922	91.5	<.0001
Error	1547	218470	141.22202		
Corrected Total	1561	399379			

The model has reported a large F-value suggesting that the observations and regression differ from the grand mean. Likewise, the F-value has a highly significant p-value under the null hypothesis that there is no linear relationship between the predictor and response variable.

Model performance statistics over the training set for Model_S are shown below.

Table 5.3.3: Stepwise Selection Model Performance Metrics (Training Set)

Measure	Value	Measure	Value
MSE	141.22202	R-Square	0.453
MAE	9.3488	Adj R-Sq	0.448
Root MSE	11.88369	C(p)	15
Dependent Mean	80.57875	AIC	7747.3481
Coeff Var	14.74792	BIC	7749.6388

The R-square value above suggests that Model_S explains ~45% of the variability in sale price using each of the included predictor variables. The adjusted R-squared value for this specification is slightly higher (superior) than the adjusted R-square reported for Model_Subj (0.4128), and its AIC and BIC are slightly lower (superior) than those reported for Model_Subj (7843.8481 and 7846.1388 respectively). These metrics suggest that Model_S may have a slightly better fit over the training set of data compared to Model_Subj.

Model performance statistics over the test set for Model_S are shown below.

Table 5.3.4: Stepwise Selection Model Performance Metrics (Test Set)

Measure	Value	Measure	Value
MSE	138.00705	R-Square	0.415
MAE	9.07674	Adj R-Sq	0.4033
Root MSE	11.74764	C(p)	15
Dependent Mean	81.2549	AIC	3532.9358
Coeff Var	14.45776	BIC	3535.5787

Again, performance metrics for Model_S have remained fairly consistent between the training and test sets. This suggests that the model is able to generalize over the test set of data.

6 Model Selection

The subjective model resulted in a specification with the lowest adjusted R-square score. This specification included 14 predictor variables which were hand picked, however three of those variables were were found to not be significantly different from zero at the 95% confidence level. Both the AIC and BIC value for this specification were also found to be higher (inferior) over both the training and test sets of data when compared to the other two models.

The adjusted R-square selection technique resulted in a specification with the highest adjusted R-square score. This specification included 16 predictor variables, all of which were significantly different from zero at the 95% confidence level. By maximizing adjusted R-square, this selection technique avoids including too many insignificant coefficient estimates, however it does suffer from including variables with a high VIF. Both the AIC and BIC value for this specification was less (superior) than Model_Subj.

The stepwise selection technique resulted in a specification which includes 14 predictor variables, all of which were found to be significantly different from zero at the 95% confidence level. This specification had a lower adjusted R-square score and a higher (inferior) AIC and BIC score than Model_AdjR2 over the training set. However, its adjusted R-square score, AIC and BIC were more comparable over the test set of data.

A summary of performance metrics over each model using the training set of data is shown below.

Table 6.1: Model Performance Metric Summary (Training Set)

Model	Pred.	MSE	MAE	R-square	Adj R-Square	C(p)	AIC	BIC
Model_Subj	14	150.22183	9.71342	0.4181	0.4128	15	7843.8481	7846.1388
Model_AdjR2	16	137.64865	9.25776	0.4675	0.462	17	7709.2951	7711.669
Model_S	14	141.22202	9.3488	0.453	0.448	15	7747.3481	7749.6388

A summary of performance metrics over each model using the test set of data is shown below.

Table 6.2: Model Performance Metric Summary (Test Set)

Model	Pred.	MSE	MAE	R-square	Adj R-Square	C(p)	AIC	BIC
Model_Subj	14	146.95376	9.29372	0.3646	0.3770	15	3577.7844	3580.4273
Model_AdjR2	16	137.76292	9.03843	0.4177	0.4043	17	3533.6258	3536.4539
Model_S	14	138.00705	9.07674	0.415	0.4033	15	3532.9358	3535.5787

Based on the results above, this assessment concludes that Model_S is the superior model. Its performance metrics were among the most favorable over the test set of data, yet it avoided the multicollinearity issues which came with Model_AdjR2.

7 Conclusion

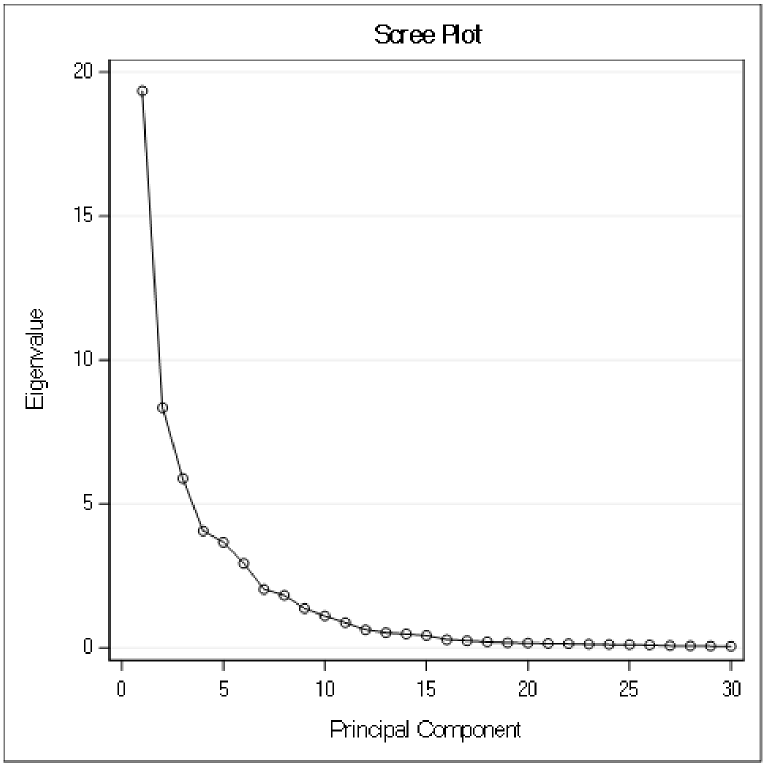
Each of the models fitted as part of this assessment has its own definition of a ‘best’ model specification. However this assessment ultimately found Model_S to be the superior model for predicting wins, due to its relatively low AIC and BIC scores and since it retained only significant variables with low VIF values. It is important to note however, that no single statistical method can be relied on to identify the ‘true’ or ‘best’ model. Effective model building requires substantive theory to suggest relevant predictors and plausible functional forms for the model. This may be particularly relevant when considering the short-comings of the dataset used for this assessment That is, we found little statistical relationship between many of the variables within the dataset, yet, we had a fundamental basis for including many variables and an expectation of appropriate coefficient polarity and model specification based on those fundamentals.

8 Bonus

8.1 PCA Analysis

We perform PCA on each of the original and transformed variables within the dataset. Note that the response variable is excluded from the dataset prior to performing the PCA routine. This plot is referred to as a Scree plot, and can be used evaluate the amount of explained variability per component.

Figure 8.1.1 Scree Plot



There are a number of options available for determining how many components should be retained. One possibility is to simply set an arbitrary threshold on the desired amount of retained variability. Alternatively, we can take a more informed approach and leverage the Scree plot above to assess any substantial drop-offs in explained variance over each component. Such a method may motivate us to retain the first four components, as the amount of explained variance does seem to level off beyond this level. For this assessment however, we allow the first fifteen components to be retained in an attempt to capture the maximum amount of explained variance.

8.2 Model 4: PCA Model

Parameter estimates for the PCA Model (Model_PCA) are shown below.

Table 8.2.1: PCA Model Parameter Estimates (Training Set)

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr >	t
Intercept	1	80.64409	0.33936	237.63	<.0001	0
Prin1	1	0.24426	0.07745	3.15	0.0016	1.0086
Prin2	1	1.74522	0.11696	14.92	<.0001	1.01017
Prin3	1	1.449	0.14212	10.2	<.0001	1.01549

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr >	t
Prin4	1	1.19577	0.16954	7.05	<.0001	1.00413
Prin5	1	0.01219	0.17527	0.07	0.9445	1.00323
Prin6	1	-2.26033	0.19259	-11.74	<.0001	1.00547
Prin7	1	0.54104	0.23257	2.33	0.0201	1.00847
Prin8	1	-1.3535	0.25583	-5.29	<.0001	1.01993
Prin9	1	2.1215	0.29347	7.23	<.0001	1.02084
Prin10	1	-0.27843	0.31722	-0.88	0.3802	1.0171
Prin11	1	-0.10821	0.35374	-0.31	0.7597	1.01481
Prin12	1	-1.41118	0.42144	-3.35	0.0008	1.01014
Prin13	1	-1.17669	0.46832	-2.51	0.0121	1.01585
Prin14	1	-1.6239	0.47863	-3.39	0.0007	1.02133
Prin15	1	-0.9432	0.5163	-1.83	0.0679	1.04349

For Model_PCA, the majority of coefficient estimates have significant p-values at the 95% level, allowing us to reject the null hypothesis and conclude that each have non-zero coefficients. The only exceptions to this are components five, ten, eleven and fifteen. The greatest improvement however is in the reported VIF value for each coefficient estimate. The VIF for all variables are close to one, which suggests no correlation between predictors.

Goodness-of-fit information for Model_PCA is shown below.

Table 8.2.2: PCA Model Analysis of Variance (Training Set)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	122014	8134.27504	45.34	<.0001
Error	1546	277365	179.40795		
Corrected Total	1561	399379			

The model has reported a large F-value suggesting that the observations and regression differ from the grand mean. Likewise, the F-value has a highly significant p-value under the null hypothesis that there is no linear relationship between the predictor and response variable.

Model performance statistics over the training set for Model_PCA are shown below.

Table 8.2.3: PCA Model Performance Metrics (Training Set)

Measure	Value	Measure	Value
MSE	179.40795	R-Square	0.3055
MAE	10.4695	Adj R-Sq	0.2988
Root MSE	13.39433	C(p)	16.0000
Dependent Mean	80.57875	AIC	8122.1699
Coeff Var	16.62265	BIC	8124.5009

The R-square value above suggests that Model_PCA explains ~30% of the variability in TARGET_WINS. The adjusted R-squared value also indicates a similar level of explanatory power. The adjusted R-squared value for this specification is lower (inferior) than the other three models and its AIC and BIC are also higher (inferior).

Model performance statistics over the training set for Model_PCA are shown below.

Table 8.2.4: PCA Model Performance Metrics (Test Set)

Measure	Value	Measure	Value
MSE	166.34445	R-Square	0.2959
MAE	9.90133	Adj R-Sq	0.2807
Root MSE	12.89746	C(p)	16.0000
Dependent Mean	81.25490	AIC	3667.2573
Coeff Var	15.87284	BIC	3669.9898

Again, performance metrics for Model_PCA have remained fairly consistent between the training and test sets. This suggests that the model is able to generalize over the test set of data.

While its performance metrics were found to be generally inferior to the other three models, we have elected to generate predictions using this model over the final test set of data. We find that the trade-off in performance to be acceptable in eliminating the multicollinearity issues found in the other three specifications.

References

D. Montgomery, & G. Vining, E. Peck. 2012. "Introduction to Linear Regression Analysis." John Wiley & Sons Inc.

Inc, SAS Institute. 2016. "SAS/STAT(R) 9.22 User's Guide." https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect030.htm.