Homework 1: MoneyBall Data Prediction
Jack Doig
PREDICT 411 Section 59

Contents

Introduction 2

1. Data Exploration 3

2. Data Preparation 7

3. Build Models 10

4. Select Models 11

Conclusion 13

Introduction

The purpose of this paper is to analyse and model historical baseball data so as to build a model to predict the number of winning games a team has within a regular season.

The baseball data includes records from 1871 to 2006 inclusive, with each record representing the performance statistics for a given baseball team and the corresponding number of wins for the season. The data has been normalized to a 164 game season and contains approximately 2200 records.

This task will involve firstly exploring and optimizing the data in the context of real-world baseball data, such that it offers as much predictive intelligence as possible, and then developing and scoring multiple regression models.

In order to determine the best model, a variety of different methods will be employed, including Stepwise, Forward and backward variable selection models, while each model judged on a balance of in-sample and out-of-sample measures, interpretability and simplicity.

The final model will then be analysed so as to understand its accuracy in predicting baseball wins, and further analysis undertaken if required.

1. Data Exploration

The following dimensions of the raw data were explored in order to both get an understanding of the data set and to judge the quality of the data, and what kind of manipulations or data cleansing, if any, might be required in order to improve the predictive power of the data set.

a) Data summary

There are a total of 2276 records in the data set. There are 15 possible predictor variables, all of which are continuous, and a response variable, number of wins, which is also continuous. There are no categorical variables in the data set.

The 15 predictor variables can be divided into two categories. Firstly there are 9 variables which are associated with the performance of the team while batting, for example the number of homeruns by batters, and secondly there are 6 variables related to performance while fielding/pitching, for example the number of walks allowed.

b) Descriptive statistics

The following table outlines the basic descriptive statistics for the predictor and response variables.

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
TARGET_WINS		2276	80.7908612	15.7521525	0	146
TEAM_BATTING_H	Base Hits by batters	2276	1469.27	144.5911954	891	2554.00
TEAM_BATTING_2B	Doubles by batters	2276	241.2469244	46.8014146	69	458
TEAM_BATTING_3B	Triples by batters	2276	55.2500000	27.9385570	0	223
TEAM_BATTING_HR	Homeruns by batters	2276	99.6120387	60.5468720	0	264
TEAM_BATTING_BB	Walks by batters	2276	501.5588752	122.6708615	0	878
TEAM_BATTING_SO	Strikeouts by batters	2174	735.6053358	248.5264177	0	1399.00
TEAM_BASERUN_SB	Stolen bases	2145	124.7617716	87.7911660	0	697
TEAM_BASERUN_CS	Caught stealing	1504	52.8038564	22.9563376	0	201
TEAM_BATTING_HBP	Batters hit by pitch	191	59.3560209	12.9671225	29	95
TEAM_PITCHING_H	Hits allowed	2276	1779.21	1406.84	1137.00	30132.00
TEAM_PITCHING_HR	Homeruns allowed	2276	105.6985940	61.2987469	0	343
TEAM_PITCHING_BB	Walks allowed	2276	553.0079086	166.3573617	0	3645.00
TEAM_PITCHING_SO	Strikeouts by pitchers	2174	817.7304508	553.0850315	0	19278.00
TEAM_FIELDING_E	Errors	2276	246.4806678	227.7709724	65	1898.00
TEAM_FIELDING_DP	Double Plays	1990	146.3879397	26.2263853	52	228

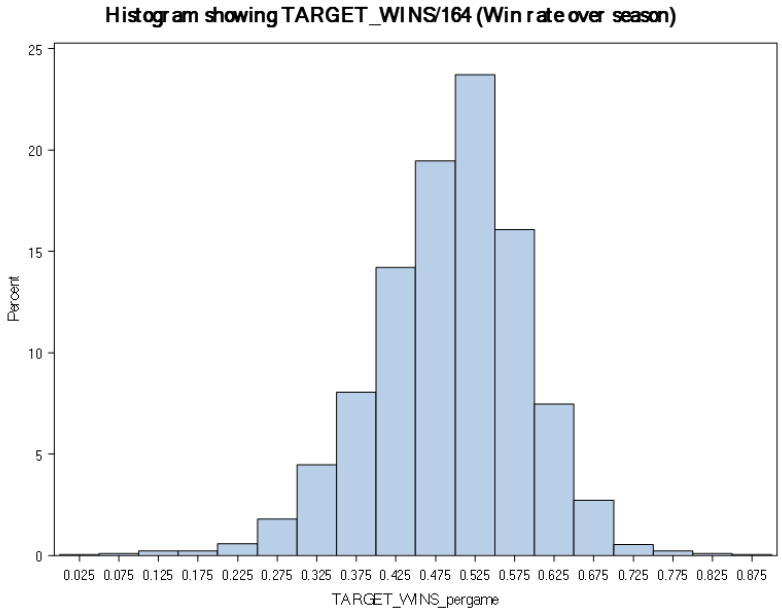
As seen in the table above, there are a number of missing variables and a number of outlier values which do not seem realistic and will need to be corrected in the data preparation step.

Furthermore there are a number of variables where we would expect to see correlation, I would expect that a team would have a general batting competence which would mean that hits, doubles, triples and home runs would be correlated, for example. There may be an opportunity to reduce dimensionality and therefore multi-colinearity.

These factors are explored further through this section and the specific adjustments made in the data preparation section.

c) Descriptive statistics – ‘Number of Wins’ response variable

The response variable has an average of 80.79, a maximum of 146, a minimum of 0, and approximates a normal distribution. In a 164-game season the maximum and minimum are perfectly possible, although the extreme values will be investigated further given the fact that winning percentages of more than 77% and less that 13% are unheard of in major league baseball. The following histogram outlines the distribution of the response variable:



While this data is from an unspecified professional league, the maximum range of values do seem to fall into the general pattern when considering the relative batting performance.

The minimum records of 0 and 12 wins do exhibit irregular data however, also claiming to have not conceded a single homerun, and having zeros in a number of other categories. This will be addressed during data preparation.

d) Predictor Variables - Missing Data

As can be seen in the table above, there are 6 predictor variables which have missing data. There is one variable which has less than 10% of the data available (Batters hit by Pitch), one variable has 66% of the data points available (Caught Stealing), while the remaining 4 variables have between 87% and 95% of the data points available.

Each of these cases will be addressed during data preparation taking into account the best logical solution for each.

a) Predictor Variables - Outliers

There are some outliers in the data which require further investigation, For example, in a regular 9-innings game with 3 outs, or 27 outs per team per game, it can be safely assumed that an average 'Strikeouts by pitchers' of 117.5 per game (19278/164) is bad data as it is virtually impossible unless the game went to 39 innings.

The full set of data which I consider to be outliers which are <u>outside the boundary of realistic</u>, even under extreme considerations is outlined below. These values will need to be dealt with during data preparation:

Predictor Variable	Description of outlier	Rationale
TEAM_BATTING_3B	There are two teams with zero values	With values of 135 and 338 for 2-base hits, I don't think 0 for 3-base hits is realistic given the general correlation
TEAM_BATTING_3B	The maximum value is 223	While this is possible, the maximum record is again uncorrelated with 2-base and base hits
TEAM_BATTING_BB	There is a value of '0'	I don't think this value is realistic over a 164 game season
TEAM_BATTING_SO	There are 20 values of '0'	Given the next value is 66, I conclude this is bad data. Not having a single strikeout in 164 games is unrealistic
TEAM_BASERUN_SB	There is a value of '0', however TEAM_BASERUN_CS is 54	If the team was attempting so many steals, I would not expect a 0 value here given the general correlation between these variables
TEAM_PITCHING_H	There is a maximum value of 30132	this is ~10,000 more than the next realistic value
TEAM_PITCHING_HR	There are 15 values of '0'	Given the general correlation with base hits, I conclude this is bad data
TEAM_PITCHING_BB	There is one value of 0	The next value is 119, so I conclude this is bad data. This is consistent with my baseball research, which gives a minimum of 282
TEAM_PITCHING_BB	There is one value in excess of 200	These are well outside normal boundaries, and does not correlate with the high number of wins (108)
TEAM_PITCHING_SO	There are 18 values of '0'	A 0-strikeout season is unrealistic given the next lowest value is 181, and baseball research confirms this
TEAM_PITCHING_SO	There are 4 values in excess of 4000 (19278,12758,5456,4224)	With normally or 27 outs per team per game, a value over 4428 is highly suspicious

b) Correlations in the data

The table below outlines the correlations in the raw data set. There are two reasons identifying the correlations in the data might be useful. Firstly there may be an opportunity to reduce dimensionality so as to avoid multi-colinearity issues, and secondly there may be an opportunity to impute missing data by using a regression with a correlated predictor variable. The correlations of greater than abs(0.4) are highlighted in red.

	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
TARGET_WINS	1.00000	0.38877	0.28910	0.14261	0.17615	0.23256	-0.03175	0.13514	0.02240	0.07350	-0.10994	0.18901	0.12417	-0.07844	-0.17648	-0.03485
TEAM_BATTING_H Base Hits by batters	0.38877	1.00000	0.56285	0.42770	-0.00654	-0.07246	-0.46385	0.12357	0.01671	-0.02911	0.30269	0.07285	0.09419	-0.25266	0.26490	0.15538
TEAM_BATTING_2B Doubles by batters	0.28910	0.56285	1.00000	-0.10731	0.43540	0.25573	0.16269	-0.19976	-0.09981	0.04608	0.02369	0.45455	0.17805	0.06479	-0.23515	0.29088
TEAM_BATTING_3B Triples by batters	0.14261	0.42770	-0.10731	1.00000	-0.63557	-0.28724	-0.66978	0.53351	0.34876	-0.17425	0.19488	-0.56784	-0.00222	-0.25882	0.50978	-0.32307
TEAM_BATTING_HR Homeruns by batters	0.17615	-0.00654	0.43540	-0.63557	1.00000	0.51373	0.72707	-0.45358	-0.43379	0.10618	-0.25015	0.96937	0.13693	0.18471	-0.58734	0.44899
TEAM_BATTING_BB Walks by batters	0.23256	-0.07246	0.25573	-0.28724	0.51373	1.00000	0.37975	-0.10512	-0.13699	0.04746	-0.44978	0.45955	0.48936	-0.02076	-0.65597	0.43088
TEAM_BATTING_SO Strikeouts by batters	-0.03175	-0.46385	0.16269	-0.66978	0.72707	0.37975	1.00000	-0.25449	-0.21788	0.22094	-0.37569	0.66718	0.03701	0.41623	-0.58466	0.15489
TEAM_BASERUN_SB Stolen bases	0.13514	0.12357	-0.19976	0.53351	-0.45358	-0.10512	-0.25449	1.00000	0.65524	-0.06400	0.07329	-0.41651	0.14642	-0.13713	0.50963	-0.49708
TEAM_BASERUN_CS Caught stealing	0.02240	0.01671	-0.09981	0.34876	-0.43379	-0.13699	-0.21788	0.65524	1.00000	-0.07051	-0.05201	-0.42257	-0.10696	-0.21022	0.04832	-0.21425
TEAM_BATTING_HBP Batters hit by pitch	0.07350	-0.02911	0.04608	-0.17425	0.10618	0.04746	0.22094	-0.06400	-0.07051	1.00000	-0.02770	0.10676	0.04785	0.22157	0.04179	-0.07121
TEAM_PITCHING_H Hits allowed	-0.10994	0.30269	0.02369	0.19488	-0.25015	-0.44978	-0.37569	0.07329	-0.05201	-0.02770	1.00000	-0.14161	0.32068	0.26725	0.66776	-0.22865
TEAM_PITCHING_HR Homeruns allowed	0.18901	0.07285	0.45455	-0.56784	0.96937	0.45955	0.66718	-0.41651	-0.42257	0.10676	-0.14161	1.00000	0.22194	0.20588	-0.49314	0.43917
TEAM_PITCHING_BB Walks allowed	0.12417	0.09419	0.17805	-0.00222	0.13693	0.48936	0.03701	0.14642	-0.10696	0.04785	0.32068	0.22194	1.00000	0.48850	-0.02284	0.32446
TEAM_PITCHING_SO Strikeouts by pitchers	-0.07844	-0.25266	0.06479	-0.25882	0.18471	-0.02076	0.41623	-0.13713	-0.21022	0.22157	0.26725	0.20588	0.48850	1.00000	-0.02329	0.02616
TEAM_FIELDING_E Errors	-0.17648	0.26490	-0.23515	0.50978	-0.58734	-0.65597	-0.58466	0.50963	0.04832	0.04179	0.66776	-0.49314	-0.02284	-0.02329	1.00000	-0.49768
TEAM_FIELDING_DP Double Plays	-0.03485	0.15538	0.29088	-0.32307	0.44899	0.43088	0.15489	-0.49708	-0.21425	-0.07121	-0.22865	0.43917	0.32446	0.02616	-0.49768	1.00000

There is no close correlation with the target variable. The best is the TEAM_BATTING_H variable with a correlation of 0.38. The strongest correlation among the predictor variables is the correlation of 0.97 between home runs scored when batting, and homeruns conceded while fielding. This would indicate that pitching weakness (conceding the biggest hits) is correlated to batting strength, or you cannot be both good at scoring and preventing homeruns.

Other notable correlations are:

- Correlation of 0.66 between hits allowed and errors, and a correlation of correlation of 0.66 between errors and walks, suggesting a general fielding weakness/strength factor in the data
- Correlation of 0.72 between homeruns by batters and strikeouts, suggesting a high risk/low risk game plan factor in the data
- Correlation of -0.66 between strikeouts and 3-base hits, which seems counter-intuitive given b)

- Correlation of 0.66 between stolen bases and caught stealing, suggesting there is a constant probability of success when stealing bases

2. Data Preparation

a) Severe Missing data – data removal

There are 3 rows which contain too much unrealistic data (as defined in the data exploration section) across the columns, such that I determine they cannot be salvaged so I have removed them from the data set.

b) Missing and outlier data – data correction

As defined above, there are a number of data points which sit outside the realms of realistic in a 164-game season. In order to correct this data, there were a number of options:

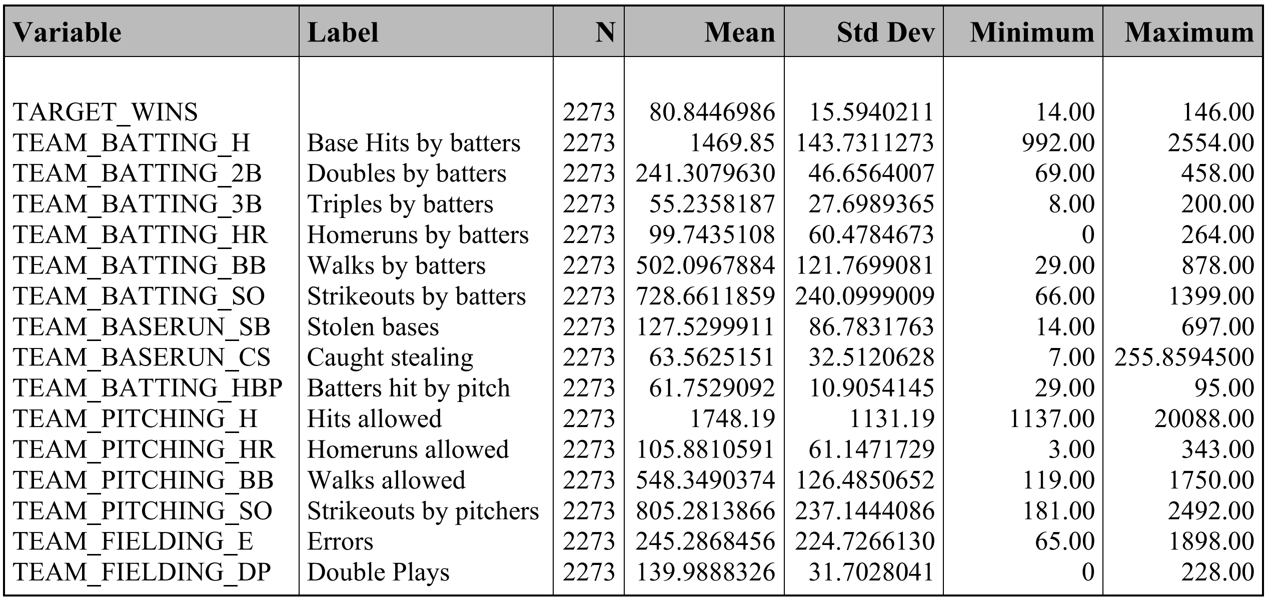
- Remove the row or column of data – this would have the effect of significantly reducing the sample size, or number of predictors and the strength of the analysis
- Imputation using an average value, zero, or a regression imputation.
- Flagging the data as missing

I have created the most of the missing data points based on a regression (regression imputation) with the most correlated predictor variable from the table of correlations above. For example, a team's execution of double plays is negatively correlated with the errors they make while fielding, this is consistent with the idea that skilled fielding teams would be able to execute more double plays.

This has the effect of making the model perhaps more precise than is warranted, but means that the sample size is not reduced and the predictive power of the remaining data in the row is preserved. This covers about 60 missing and unrealistic data points as defined in the data exploration section.

The exceptions to this are where there is no reasonably correlated variable. This specifically refers to the 'team batting hit by pitch' variable. I decided that while this is likely a non-random event, a random imputation, within the range and distribution of the available data is likely to be a better model of the real-world and therefore create a better model that imputing a mean or median.

Following the above data corrections, the following are the descriptive statistics of the original predictor variables:



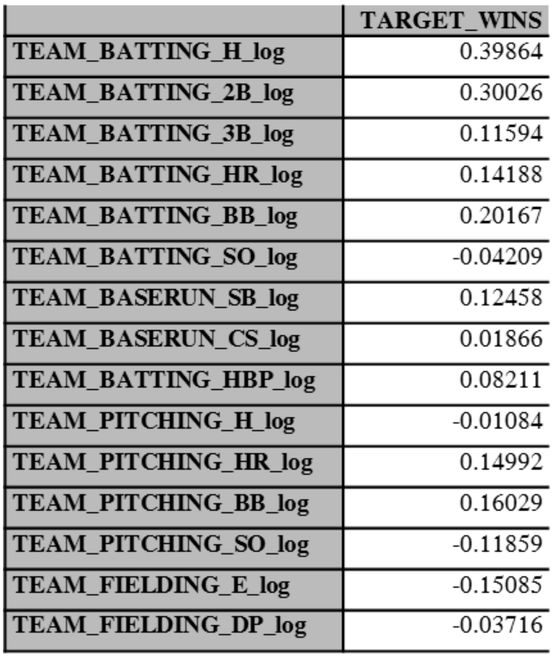
Variable	Label	N	Mean	Std Dev	Minimum	Maximum
TARGET_WINS		2273	80.8446986	15.5940211	14.00	146.00
TEAM_BATTING_H	Base Hits by batters	2273	1469.85	143.7311273	992.00	2554.00
TEAM_BATTING_2B	Doubles by batters	2273	241.3079630	46.6564007	69.00	458.00
TEAM_BATTING_3B	Triples by batters	2273	55.2358187	27.6989365	8.00	200.00
TEAM_BATTING_HR	Homeruns by batters	2273	99.7435108	60.4784673	0	264.00
TEAM_BATTING_BB	Walks by batters	2273	502.0967884	121.7699081	29.00	878.00
TEAM_BATTING_SO	Strikeouts by batters	2273	728.6611859	240.0999009	66.00	1399.00
TEAM_BASERUN_SB	Stolen bases	2273	127.5299911	86.7831763	14.00	697.00
TEAM_BASERUN_CS	Caught stealing	2273	63.5625151	32.5120628	7.00	255.8594500
TEAM_BATTING_HBP	Batters hit by pitch	2273	61.7529092	10.9054145	29.00	95.00
TEAM_PITCHING_H	Hits allowed	2273	1748.19	1131.19	1137.00	20088.00
TEAM_PITCHING_HR	Homeruns allowed	2273	105.8810591	61.1471729	3.00	343.00
TEAM_PITCHING_BB	Walks allowed	2273	548.3490374	126.4850652	119.00	1750.00
TEAM_PITCHING_SO	Strikeouts by pitchers	2273	805.2813866	237.1444086	181.00	2492.00
TEAM_FIELDING_E	Errors	2273	245.2868456	224.7266130	65.00	1898.00
TEAM_FIELDING_DP	Double Plays	2273	139.9888326	31.7028041	0	228.00

Generally, the changing of data has generally had the effect of a marginal increase in the averages, given many of the corrected values were 0, and a reducing of the variance or standard deviation for the same reason.

c) Mathematical transformations

I have explored square roots, squares and natural logs of the predictor and the response variables in order to see if there was a noticeable increase in correlation with the response variable. The most promising transformation was the log transformation.

Below are the correlations of the log-transforms with the response variable:



	TARGET_WINS
TEAM_BATTING_H_log	0.39864
TEAM_BATTING_2B_log	0.30026
TEAM_BATTING_3B_log	0.11594
TEAM_BATTING_HR_log	0.14188
TEAM_BATTING_BB_log	0.20167
TEAM_BATTING_SO_log	-0.04209
TEAM_BASERUN_SB_log	0.12458
TEAM_BASERUN_CS_log	0.01866
TEAM_BATTING_HBP_log	0.08211
TEAM_PITCHING_H_log	-0.01084
TEAM_PITCHING_HR_log	0.14992
TEAM_PITCHING_BB_log	0.16029
TEAM_PITCHING_SO_log	-0.11859
TEAM_FIELDING_E_log	-0.15085
TEAM_FIELDING_DP_log	-0.03716

I will retain these log transformations to see if they can create a positive impact when creating models in terms of linearizing the relationship, increasing the R-squared, reducing heteroscedacity, or to create a more normal distribution of residuals.

d) Combinations of variables

Using my knowledge of baseball, I have created 7 additional variables that may offer some predictive power, and possibly help in reducing the dimensionality of the data set. Since all teams are measured on a 164 game season, using the absolute values is appropriate, as while a team may, for example, have a higher percentage of base steals, they also need to be attempting many steals to gain benefit. There is a low value in attempting one base hit only, even if it is successful and the team has a 100% success rate.

- **Batting competence:** This measure adds a single base hit, with a 2-base hit, a 3-base hit and a home run, scaling each by the number of bases gained (1-4). Walks by batters are added also, which are a positive reflection of batting, while strikeouts by batters are subtracted from the total.
- **Base-stealing competence:** This measure takes bases successfully stolen minus caught stealing
- **Pitching competence** (high is less competent): This measure takes hits allowed, combined with homeruns allowed *4, walks allowed, minus strikeouts by pitchers
- **Fielding competence** (high is less competent): Errors - double plays
- **Offensive competence**: batting competence + base-stealing competence
- **Defensive competence:** pitching competence + fielding competence
- **Total competence** : offensive – defensive competence

The correlations of these variables with the target variable are as follows:

	TARGET_ WINS	PITCHING_ COMPETENCE	BASESTEALING_ COMPETENCE	BATTING_ COMPETENCE	FIELDING_ COMPETENCE
TARGET_WINS	1.00000	0.02092	0.15183	0.44622	-0.15146
PITCHING_COMPETENCE	0.02092	1.00000	0.07582	0.32526	0.57564
BASESTEALING_COMPETENCE	0.15183	0.07582	1.00000	-0.01655	0.43648
BATTING_COMPETENCE	0.44622	0.32526	-0.01655	1.00000	-0.09918
FIELDING_COMPETENCE	-0.15146	0.57564	0.43648	-0.09918	1.00000

There is, as would be expected, lower multicolinearity between these variables, so they may be useful to explore during the modelling stage.

3. Build Models

I have chosen to use the following ways of deriving a set of possible models:

a) The full set of original predictor variables, modelled as-is (ie no predictors removed), and with stepwise, forward and backward selection processes
b) A model which uses the derived variables (competences) above only modelled as-is, and with stepwise, forward and backward selection processes
c) A model using the mathematical transformations which gave a better correlation with the target variable, modelled as-is, and with stepwise, forward and backward selection processes

During modelling I took note of the major in-sample metrics: R-squared, adjusted R-squared, AIC, BIC and RMSE. For out-of-sample testing I used 10% of the dataset or 228 rows of data were used in the prediction of the out-of-sample RMSE.

Below are the measures for the methods mentioned above:

Model type	Selection method	Number or variables in model	R-Squared	Adjusted R-Squared	AIC	BIC	In-Sample RMSE	Out of Sample RMSE
Original predictor variables	As-is	15	0.3281	0.3231	10419.91	10422.17	12.75	13.14
	Forward	13	0.3278	0.3235	10416.87	10419.09	12.72	13.13
	Backward	11	0.3269	0.3233	10415.50	10417.65	12.72	13.18
	Stepwise	12	0.3269	0.3233	10415.50	10417.65	12.72	13.18
Derived predictor variables	As-is	4	0.2581	0.2567	10600.49	10602.52	13.33	13.91
	Forward	4	0.2581	0.2567	10600.49	10602.52	13.33	13.91
	Backward	3	0.2574	0.2563	10600.56	10602.57	13.34	13.86
	Stepwise	3	0.2574	0.2563	10600.56	10602.57	13.34	13.86
Original predictors with log transformation replacement	As-is	15	0.3177	0.3127	10451.21	10453.47	12.82	12.97
	Forward	13	0.3176	0.3132	10447.72	10449.94	12.82	13.01
	Backward	13	0.3196	0.3153	10441.59	10443.81	12.80	12.98
	Stepwise	**12**	**0.3173**	**0.3133**	**10446.47**	**10448.65**	**12.81**	**12.98**

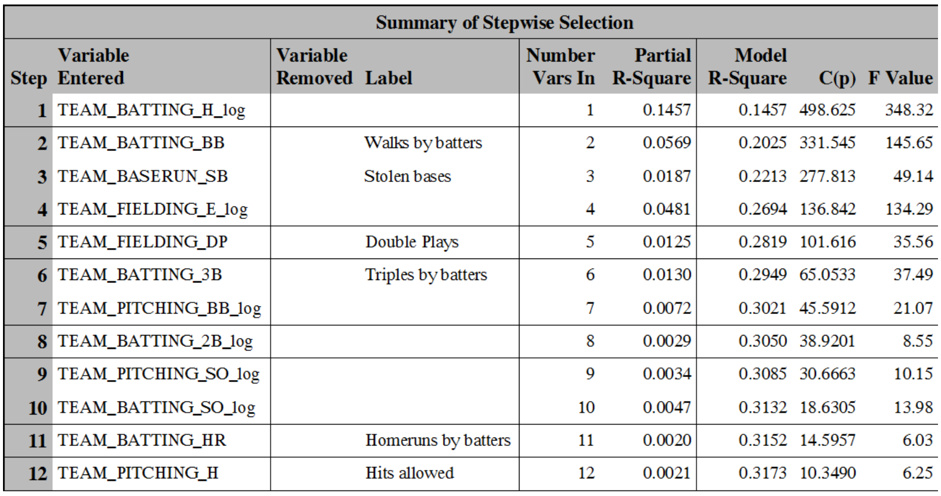
As can be seen above, the set of derived variables underperform the other two categories of model, and the largest different in percentage terms is among the RMSE values. This will lead to this metric being influential during model selection.

4. Select Models

My objective in choosing the model are as follows:

- The model should be meaningful, the predictor variables should contribute as one would expect for the real-world meaning of the predictors
- The model should perform well in in-sample testing based on Adjusted R-squared
- In a choice between similarly performing complex versus parsimonious, the parsimonious model is preferred
- The model should perform well on out-of-sample testing

According to these criteria, which I weigh equally, and studying the output from the regression model tests above, the best model is the final model, which contains the predictor variables below:



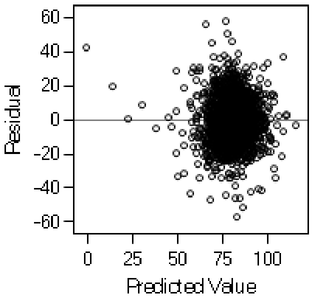
Summary of Stepwise Selection

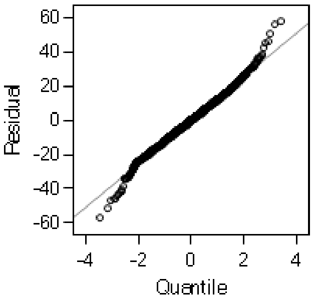
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value
1	TEAM_BATTING_H_log			1	0.1457	0.1457	498.625	348.32
2	TEAM_BATTING_BB		Walks by batters	2	0.0569	0.2025	331.545	145.65
3	TEAM_BASERUN_SB		Stolen bases	3	0.0187	0.2213	277.813	49.14
4	TEAM_FIELDING_E_log			4	0.0481	0.2694	136.842	134.29
5	TEAM_FIELDING_DP		Double Plays	5	0.0125	0.2819	101.616	35.56
6	TEAM_BATTING_3B		Triples by batters	6	0.0130	0.2949	65.0533	37.49
7	TEAM_PITCHING_BB_log			7	0.0072	0.3021	45.5912	21.07
8	TEAM_BATTING_2B_log			8	0.0029	0.3050	38.9201	8.55
9	TEAM_PITCHING_SO_log			9	0.0034	0.3085	30.6663	10.15
10	TEAM_BATTING_SO_log			10	0.0047	0.3132	18.6305	13.98
11	TEAM_BATTING_HR		Homeruns by batters	11	0.0020	0.3152	14.5957	6.03
12	TEAM_PITCHING_H		Hits allowed	12	0.0021	0.3173	10.3490	6.25

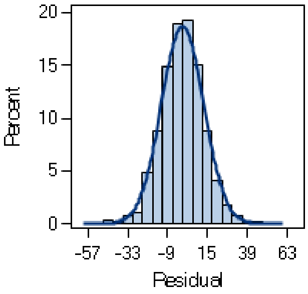
The model has 12 variables, which is the lowest among the top-performing models. The shortcoming in terms of understanding this model is the use of the log variables. This can be explained however, as being used to 'linearize' the predictor variables against the target variable

The strongest aspect of this model is that it performs the second-best in out-of-sample RMSE testing, and is 0.09 from the best model in in-sample RMSE testing. The differences between the out of sample tests in particular are significant between the top performing models.

The model has an adjusted r-squared of 0.3133. While this was not the lowest, the difference with the other models was marginal (this model is less than 0.01 from the best performing model on this metric). The differences between AIC and BIC among these models is also negligible.







In terms of the in-sample testing, the charts above show the distribution of the residuals. The residuals follow a reasonable normal distribution, and there is no discernible pattern when plotted against the predictions, so there is no major hetroscedacity issue.

The final model, with the beta values and errors, is as below:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-304.30066	46.95399	6902.61680	42.00	<.0001
TEAM_BATTING_H_log	75.07852	6.31657	23218	141.28	<.0001
TEAM_BATTING_2B_log	-5.88131	2.33921	1038.86814	6.32	0.0120
TEAM_BATTING_3B	0.11961	0.01846	6895.70688	41.96	<.0001
TEAM_BATTING_HR	0.02753	0.01011	1219.08330	7.42	0.0065
TEAM_BATTING_BB	0.04044	0.00683	5759.25269	35.04	<.0001
TEAM_BATTING_SO_log	-11.24385	2.39138	3633.17416	22.11	<.0001
TEAM_BASERUN_SB	0.03965	0.00473	11539	70.21	<.0001
TEAM_PITCHING_H	-0.00103	0.00041167	1027.96224	6.25	0.0125
TEAM_PITCHING_BB_log	-13.57495	2.82175	3803.58330	23.14	<.0001
TEAM_PITCHING_SO_log	11.44466	2.40890	3709.53796	22.57	<.0001
TEAM_FIELDING_E_log	-12.92534	1.09888	22737	138.35	<.0001
TEAM_FIELDING_DP	-0.08670	0.01448	5891.44681	35.85	<.0001

The interpretability of the model is counter-intuitive in the case of ‘doubles’ or hits to second base in particular, which one would expect a positive effect on the number of wins in a season. The interpretability is also affected generally by the variables with log transforms.

I believe these small interpretability issues are a good trade-off against the improved performance of the model.

Conclusion

In exploring this data, I have used techniques in exploratory data analysis to reconcile a data-set to a real world domain, specifically this exercise explored a set of baseball data from a professional league from 1871 to 2006. It is important to note that a dramatic change to the rules or the composition of the game in the future could weaken the predictive strength of this analysis.

In some cases, the rows of data contained so much error that I elected to remove it. The data was found to contain a lot of missing values, which required action in order to improve the predictive capability of the data set. Similarly, there were a number of data points which I judged to be erroneous so were corrected. Therefore, there was a reasonable degree of error in the data set used however there was reasonable rigor in the measuring process such that I am confident the output I a powerful model.

Only where there was an extremely low-probability of valid data or was it missing did I take corrective action. The most frequent action I took was doing a regression replacement with the best other single predictor of that variable in the data set. Where there was no correlation I tried to match the distribution of the existing variables. I believe these to be reasonable approaches to data correction, however the model should be reviewed going forward to ensure it remains useful.

I explored a number of derived variables data transformations in order to better linearize some of the predictor variables. This appeared to make a marginal improvement in the case of the natural log, so those derived predictors were retained for the modelling process. While this reduced the interpretability of the model I judged this to be a reasonable trade-off against the performance improvement.

In modelling, I explored the original predictors, the derived competencies and the transformed variables, each time using a full set of predictors, and forward, backward and stepwise selection methods. The derived variables performed poorly, while the transformed variable set of models and the original variables set of models performed similarly well. The stepwise process seemed to product slightly better results, as did the log transformed variables. The sample error was the strongest differentiator in selecting a model.

The trade-offs to selecting a model with log-transforms as part of it is the real-world understanding of the model, however I believe that those transforms can be explained and the performance improvements justify these transformations.