Keys to Success in Major League Baseball

Major League Baseball is deeply rooted in statistics and is perhaps the most statistically influenced sport. Statistics in baseball has evolved from simple to much more advanced methods as teams realize the tremendous benefits. This concept of eliminating subjectivity in baseball picks was introduced to the public via the mainstream 2011 sports drama film *MoneyBall.* While this did not introduce the idea of statistical analysis in baseball, it did bring the concept to the general public. As a result, there is an abundance of player and team statistics widely available through various sports websites. However, this abundance of statistics also challenges even the most experienced statisticians and data scientists in providing meaningful insight. The following paper attempts to reduce the amount of available statistics into just a few meaningful aspects of the game of baseball that can be used to improve overall team success.

Section I - Descriptive Statistics

Problem Statement

Determine which of the 18 individual statistics are most related to a MLB team's success as measured by total wins per season. Specifically, the intent is to develop and select several combinations of variables that account for a large percent of the variance.

Constraints and Limitations

The scope of the study is the 2015 MLB regular season of play, consisting of 30 individual teams playing a total of 162 games each. The reduced model will be developed using the 2015 season and validated using the 2012 season of play. Other methods could be utilized such as combining or averaging several seasons of play, but the proposed model will be solely based on the 2015 season of play.

Several considerations or limitations need to be understood. The team statistics included in this evaluation is certainly not a comprehensive list, as there are many more available through various sources on the Internet. Also, the data included in this evaluation are team based statistics and does not include the individual player statistics. The inclusion of the detailed player statistics would undoubtedly complicate the analysis, but may also lead to additional clarity and insight. Additional studies are certainly merited and should be performed to evaluate the need to expand the study in both width and depth.

Lastly, the study is purely observational and no casual inferences can be made about the relationships between the explanatory variables and the single response variable.

Data Set Description

The data for the analysis was retrieved from the Internet and can be found at the following location:

http://www.baseball-reference.com/leagues/MLB/2015.shtml

The data was supplemented with the number of wins in the season which is also available at the above noted website. In addition, most of the calculated and estimated variables were removed from the dataset. The intent was to include the most basic statistics. All variables, both response and explanatory, are listed in Figure 1 and the entire dataset is shown in Figure 2.

Variable	Usage	Description
Tm	Identifier	Team name
Wins	Reponse	The number of wins in the regular season
AB	Explanatory	At Bat - Plate appearances, not including bases on balls, being hit by pitch, sacrifices, interference, or obstruction.
R	Explanatory	Runs scored - Number of times a player crosses home plate
H	Explanatory	Hits - Times reached base because of a batted, fair ball without error by the defense
B2	Explanatory	Double - Hits on which the batter reaches second base safely without the contribution of a fielding error.
B3	Explanatory	Triple - Hits on which the batter reaches third base safely without the contribution of a fielding error.
HR	Explanatory	Home Runs - Hits on which the batter successfully touched all four bases, without the contribution of a fielding error.
RBI	Explanatory	Run Batted In - Number of runners who score due to a batters' action, except when batter grounded into double play or reached on an error
SB	Explanatory	Stolen Bases - Number of bases advanced by the runner while the ball is in the possession of the defense.
CS	Explanatory	Caught Stealing - Times tagged out while attempting to steal a base
BB	Explanatory	Base on Balls (walk) - Hitter not swinging at four pitches called out of the strike zone and awarded first base.
SO	Explanatory	Strikeout
BA	Explanatory	Batting Average - Hits divided by at bats (H/AB)
GDP	Explanatory	Ground into Double Play - Number of ground balls hit that became double plays
HBP	Explanatory	Hit by Pitch - Times touched by a pitch and awarded first base as a result
SH	Explanatory	Sacrifice Hit - Number of sacrifice bunts which allow runners to advance on the basepaths
SF	Explanatory	Sacrifice Fly - Fly balls hit to the outfield which although caught for an out, allow a baserunner to advance
IBB	Explanatory	Intentional Base on Balls - Times awarded first base on balls deliberately thrown by the pitcher
LOB	Explanatory	Runner Left on Base - The number of baserunners a pitcher does not allow to score

Figure 1 List of Variables and Descriptions

Tm	Wins	AB	R	H	B2	B3	HR	RBI	SB	CS	BB	SO	BA	GDP	HBP	SH	SF	IBB	LOB
ARI	79	5649	720	1494	289	48	154	680	132	44	490	1312	0.264	134	33	46	57	40	1153
ATL	67	5420	573	1361	251	18	100	548	69	33	471	1107	0.251	148	44	67	31	39	1145
BAL	81	5485	713	1370	246	20	217	686	44	25	418	1331	0.25	127	51	20	32	23	990
BOS	78	5640	748	1495	294	33	161	706	71	27	478	1148	0.265	127	46	30	42	28	1142
CHC	97	5491	689	1341	272	30	171	657	95	37	567	1518	0.244	101	74	32	35	47	1165
CHW	76	5533	622	1381	260	27	136	595	68	42	404	1231	0.25	125	65	30	37	22	1065
CIN	64	5571	640	1382	257	27	167	613	134	38	496	1255	0.248	112	42	47	40	38	1148
CLE	81	5439	669	1395	303	29	141	640	86	28	533	1157	0.256	134	39	47	50	34	1147
COL	68	5572	737	1479	274	49	186	702	97	43	388	1283	0.265	114	33	44	34	47	1016
DET	74	5605	689	1515	289	49	151	660	83	51	455	1259	0.27	152	41	23	35	36	1111
HOU	86	5459	729	1363	278	26	230	691	121	48	486	1392	0.25	102	56	28	43	22	1036
KCR	95	5575	724	1497	300	42	139	689	104	34	383	973	0.269	133	77	34	47	28	1079
LAA	85	5417	661	1331	243	21	176	621	52	34	435	1150	0.246	116	58	37	40	34	1013
LAD	92	5385	667	1346	263	26	187	638	59	34	563	1258	0.25	135	60	49	30	31	1121
MIA	71	5463	613	1420	236	40	120	575	112	45	375	1150	0.26	133	39	71	40	30	1059
MIL	68	5480	655	1378	274	34	145	624	84	29	412	1299	0.251	130	41	55	34	35	1026
MIN	83	5467	696	1349	277	44	156	661	70	38	439	1264	0.247	133	40	30	41	31	993
NYM	90	5527	683	1351	295	17	177	654	51	25	488	1290	0.244	130	68	29	32	42	1098
NYY	87	5567	764	1397	272	19	212	737	63	25	554	1227	0.251	105	63	24	54	23	1151
OAK	68	5600	694	1405	277	46	146	661	78	29	475	1119	0.251	124	40	14	38	21	1102
PHI	63	5529	626	1374	272	37	130	586	88	32	387	1274	0.249	119	54	53	29	20	1066
PIT	98	5631	697	1462	292	27	140	661	98	45	461	1322	0.26	115	89	63	41	46	1166
SDP	74	5457	650	1324	260	36	148	623	82	29	426	1327	0.243	108	40	52	42	22	1028
SEA	76	5544	656	1379	262	22	198	624	69	45	478	1336	0.249	123	36	38	35	31	1080
SFG	84	5565	696	1486	288	39	136	663	93	36	457	1159	0.267	142	49	45	37	30	1130
STL	100	5484	647	1386	288	39	137	619	69	38	506	1267	0.253	128	66	39	42	47	1152
TBR	80	5485	644	1383	278	32	167	612	87	45	436	1310	0.252	121	84	19	47	22	1075
TEX	88	5511	751	1419	279	32	172	707	101	39	503	1233	0.257	99	76	43	54	32	1130
TOR	93	5509	891	1480	308	17	232	852	88	23	570	1151	0.269	140	54	36	62	12	1057
WSN	83	5428	703	1363	265	13	177	665	57	23	539	1344	0.251	129	44	55	51	38	1114

Figure 2 Complete 2015 Dataset

Section II - Exploratory Analysis

Suitability of PCA

The ability to interpret datasets grows increasingly difficult as the number of explanatory variables increase. While the 18 explanatory variables might work in a traditional multiple regression analysis, reducing the amount of variables will simplify the process. The larger concern with the data is the likelihood of multicollinearity. Therefore, principal components analysis (PCA) will be utilized to develop theme based linear combinations. The selected principal components will then be analyzed using traditional multiple regression.

The 2015 MLB data is summarized in Figure 3 indicating generalized statistics for the response and explanatory variables. While all variables are continuous the scales range dramatically, therefore, standardization is required. The PCA analysis will be performed using the correlation matrix in lieu of the covariance matrix. This standardization will level all variables and ensure the results are not dominated by a select few.

Variable	N	Mean	Median	Std Dev	Variance	Minimum	Maximum
Wins	30	80.97	81.00	10.45	109.27	63.00	100.00
AB	30	5516.27	5510.00	70.47	4965.65	5385.00	5649.00
R	30	688.23	689.00	58.76	3452.94	573.00	891.00
H	30	1403.53	1382.50	57.14	3265.09	1324.00	1515.00
B2	30	274.73	275.50	18.10	327.44	236.00	308.00
B3	30	31.30	31.00	10.45	109.25	13.00	49.00
HR	30	163.63	158.50	31.82	1012.72	100.00	232.00
RBI	30	655.00	658.50	56.67	3211.10	548.00	852.00
SB	30	83.50	83.50	22.82	520.53	44.00	134.00
CS	30	35.47	35.00	8.06	65.02	23.00	51.00
BB	30	469.10	473.00	57.05	3255.13	375.00	570.00
SO	30	1248.20	1261.50	103.76	10766.03	973.00	1518.00
BA	30	0.25	0.25	0.01	0.00	0.24	0.27
GDP	30	124.63	127.00	13.52	182.86	99.00	152.00
HBP	30	53.40	50.00	15.70	246.52	33.00	89.00
SH	30	40.00	38.50	14.39	207.03	14.00	71.00
SF	30	41.07	40.00	8.45	71.44	29.00	62.00
IBB	30	31.70	31.00	9.20	84.70	12.00	47.00
LOB	30	1091.93	1100.00	54.43	2962.69	990.00	1166.00

Figure 3 Summary Statistics

Linearity of the data is also a concern as nonlinearity requires an alternate dimension reduction technique. The scatterplot shown in Figure 4 does not indicate visual evidence on nonlinear relationships amongst the variables. In fact, there are several highly correlated variables as highlighted by the red boxes.

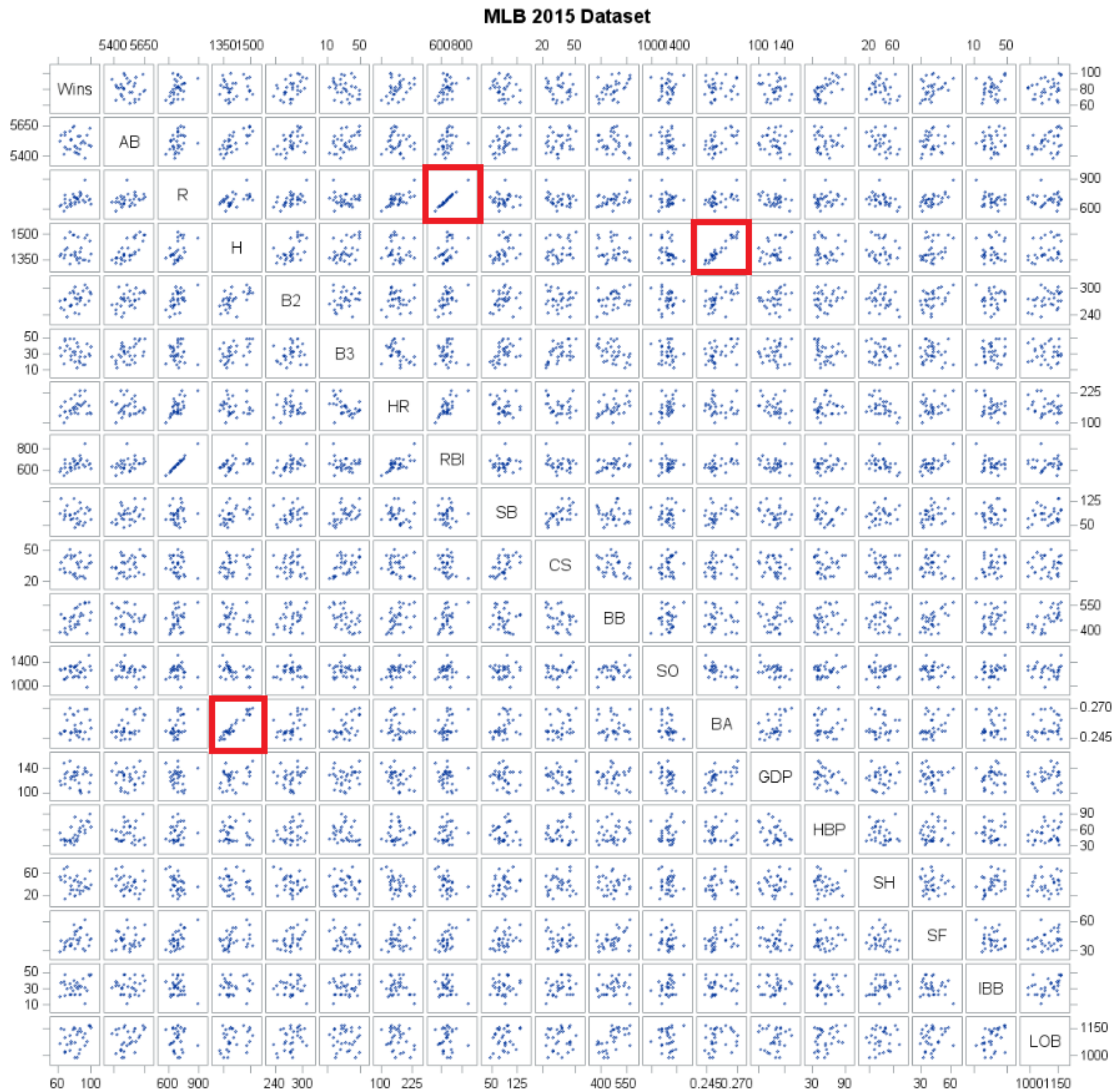

Figure 4 Scatterplot Matrix of all Variables

The correlations amongst the variables are quantified by using Pearson Coefficients and are presented in Figure 5. The red coloration indicates high correlation, blue moderate, cyan low, and white very low. A large percentage of the data are correlated and PCA will generate independent linear combinations.

Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0																			
	Wins	AB	R	H	B2	B3	HR	RBI	SB	CS	BB	SO	BA	GDP	HBP	SH	SF	IBB	LOB
Wins	1.00000	-0.08795 0.6440	0.43075 0.0175	0.03761 0.8436	0.42780 0.0184	-0.25112 0.1807	0.30741 0.0984	0.43671 0.0158	-0.15723 0.4067	-0.06608 0.7286	0.48434 0.0067	0.11185 0.5562	0.08788 0.6442	-0.11962 0.5289	0.67343 <.0001	-0.18937 0.3162	0.34658 0.0606	0.17982 0.3417	0.26641 0.1547
AB	-0.08795 0.6440	1.00000	0.31946 0.0853	0.73912 <.0001	0.45337 0.0119	0.43542 0.0162	-0.06698 0.7251	0.31318 0.0920	0.37262 0.0426	0.22462 0.2327	-0.13641 0.4723	-0.10602 0.5771	0.53350 0.0024	-0.00981 0.9590	0.00395 0.9835	-0.25581 0.1724	0.13932 0.4628	0.07473 0.6947	0.32294 0.0817
R	0.43075 0.0175	0.31946 0.0853	1.00000	0.48286 0.0069	0.56008 0.0013	-0.07007 0.7129	0.67128 <.0001	0.99642 <.0001	0.08137 0.6691	-0.27235 0.1454	0.40245 0.0275	-0.05473 0.7739	0.48311 0.0068	-0.10820 0.5693	0.09326 0.6240	-0.36983 0.0443	0.62154 0.0002	-0.26442 0.1579	-0.03057 0.8726
H	0.03761 0.8436	0.73912 <.0001	0.48286 0.0069	1.00000	0.56685 0.0011	0.47869 0.0075	-0.09085 0.6330	0.46824 0.0091	0.41344 0.0231	0.24013 0.2012	-0.11828 0.5336	-0.39883 0.0290	0.96340 <.0001	0.35857 0.0517	-0.06443 0.7352	-0.03926 0.8368	0.32577 0.0790	0.03179 0.8676	0.25399 0.1756
B2	0.42780 0.0184	0.45337 0.0119	0.56008 0.0013	0.56685 0.0011	1.00000	0.22049 0.2417	0.05629 0.7676	0.56186 0.0012	0.19503 0.3017	-0.08751 0.6457	0.30270 0.1040	-0.15075 0.4265	0.52920 0.0026	0.19476 0.3024	0.22698 0.2277	-0.25441 0.1749	0.42374 0.0196	0.03574 0.8513	0.34332 0.0632
B3	-0.25112 0.1807	0.43542 0.0162	-0.07007 0.7129	0.47869 0.0075	0.22049 0.2417	1.00000	-0.43091 0.0174	-0.08843 0.6421	0.45744 0.0110	0.46839 0.0090	-0.45495 0.0115	-0.14120 0.4567	0.42465 0.0193	0.13645 0.4722	-0.26466 0.1575	-0.06512 0.7325	-0.07791 0.6824	0.12428 0.5129	-0.07039 0.7117
HR	0.30741 0.0984	-0.06698 0.7251	0.67128 <.0001	-0.09085 0.6330	0.05629 0.7676	-0.43091 0.0174	1.00000	0.68445 <.0001	-0.13657 0.4718	-0.16554 0.3820	0.42569 0.0190	0.35992 0.0507	-0.08506 0.6549	-0.36171 0.0495	0.06269 0.7421	-0.48129 0.0071	0.28110 0.1324	-0.28826 0.1224	-0.26940 0.1500
RBI	0.43671 0.0158	0.31318 0.0920	0.99642 <.0001	0.46824 0.0091	0.56186 0.0012	-0.08843 0.6421	0.68445 <.0001	1.00000	0.05489 0.7733	-0.29236 0.1169	0.42361 0.0197	-0.04864 0.7986	0.46718 0.0092	-0.09477 0.6184	0.09201 0.6287	-0.39390 0.0313	0.60748 0.0004	-0.26461 0.1576	-0.02340 0.9023
SB	-0.15723 0.4067	0.37262 0.0426	0.08137 0.6691	0.41344 0.0231	0.19503 0.3017	0.45744 0.0110	-0.13657 0.4718	0.05489 0.7733	1.00000	0.51865 0.0033	-0.09835 0.6051	0.03097 0.8710	0.36986 0.0442	-0.20839 0.2691	-0.06998 0.7133	0.26208 0.1618	0.30469 0.1016	0.09336 0.6236	0.21323 0.2579
CS	-0.06608 0.7286	0.22462 0.2327	-0.27235 0.1454	0.24013 0.2012	-0.08751 0.6457	0.46839 0.0090	-0.16554 0.3820	-0.29236 0.1169	0.51865 0.0033	1.00000	-0.30180 0.1050	0.22789 0.2258	0.21245 0.2597	-0.07143 0.7076	0.07338 0.7000	0.04666 0.8066	-0.13253 0.4851	0.25892 0.1671	0.02608 0.8912
BB	0.48434 0.0067	-0.13641 0.4723	0.40245 0.0275	-0.11828 0.5336	0.30270 0.1040	-0.45495 0.0115	0.42569 0.0190	0.42361 0.0197	-0.09835 0.6051	-0.30180 0.1050	1.00000	0.23365 0.2140	-0.10378 0.5852	-0.08706 0.6473	0.12279 0.5180	-0.11959 0.5291	0.40184 0.0277	0.11590 0.5419	0.59158 0.0006
SO	0.11185 0.5562	-0.10602 0.5771	-0.05473 0.7739	-0.39883 0.0290	-0.15075 0.4265	-0.14120 0.4567	0.35992 0.0507	-0.04864 0.7986	0.03097 0.8710	0.22789 0.2258	0.23365 0.2140	1.00000	-0.46953 0.0089	-0.47060 0.0087	0.08040 0.6728	-0.06416 0.7362	-0.16849 0.3734	0.28342 0.1291	-0.02473 0.8968
BA	0.08788 0.6442	0.53350 0.0024	0.48311 0.0068	0.96340 <.0001	0.52920 0.0026	0.42465 0.0193	-0.08506 0.6549	0.46718 0.0092	0.36986 0.0442	0.21245 0.2597	-0.10378 0.5852	-0.46953 0.0089	1.00000	0.45260 0.0120	-0.07369 0.6988	0.04963 0.7945	0.35505 0.0542	-0.01120 0.9531	0.18087 0.3388
GDP	-0.11962 0.5289	-0.00981 0.9590	-0.10820 0.5693	0.35857 0.0517	0.19476 0.3024	0.13645 0.4722	-0.36171 0.0495	-0.09477 0.6184	-0.20839 0.2691	-0.07143 0.7076	-0.08706 0.6473	-0.47060 0.0087	0.45260 0.0120	1.00000	-0.34343 0.0632	0.13026 0.4927	-0.11985 0.5281	0.00213 0.9911	0.03107 0.8705
HBP	0.67343 <.0001	0.00395 0.9835	0.09326 0.6240	-0.06443 0.7352	0.22698 0.2277	-0.26466 0.1575	0.06269 0.7421	0.09201 0.6287	-0.06998 0.7133	0.07338 0.7000	0.12279 0.5180	0.08040 0.6728	-0.07369 0.6988	-0.34343 0.0632	1.00000	-0.18774 0.3205	0.11048 0.5611	0.02926 0.8780	0.24644 0.1892
SH	-0.18937 0.3162	-0.25581 0.1724	-0.36983 0.0443	-0.03926 0.8368	-0.25441 0.1749	-0.06512 0.7325	-0.48129 0.0071	-0.39390 0.0313	0.26208 0.1618	0.04666 0.8066	-0.11959 0.5291	-0.06416 0.7362	0.04963 0.7945	0.13026 0.4927	-0.18774 0.3205	1.00000	-0.08024 0.6734	0.34633 0.0608	0.16815 0.3744
SF	0.34658 0.0606	0.13932 0.4628	0.62154 0.0002	0.32577 0.0790	0.42374 0.0196	-0.07791 0.6824	0.28110 0.1324	0.60748 0.0004	0.30469 0.1016	-0.13253 0.4851	0.40184 0.0277	-0.16849 0.3734	0.35505 0.0542	-0.11985 0.5281	0.11048 0.5611	-0.08024 0.6734	1.00000	-0.22891 0.2237	0.20942 0.2667
IBB	0.17982 0.3417	0.07473 0.6947	-0.26442 0.1579	0.03179 0.8676	0.03574 0.8513	0.12428 0.5129	-0.28826 0.1224	-0.26461 0.1576	0.09336 0.6236	0.25892 0.1671	0.11590 0.5419	0.28342 0.1291	-0.01120 0.9531	0.00213 0.9911	0.02926 0.8780	0.34633 0.0608	-0.22891 0.2237	1.00000	0.41029 0.0243
LOB	0.26641 0.1547	0.32294 0.0817	-0.03057 0.8726	0.25399 0.1756	0.34332 0.0632	-0.07039 0.7117	-0.26940 0.1500	-0.02340 0.9023	0.21323 0.2579	0.02608 0.8912	0.59158 0.0006	-0.02473 0.8968	0.18087 0.3388	0.03107 0.8705	0.24644 0.1892	0.16815 0.3744	0.20942 0.2667	0.41029 0.0243	1.00000

Figure 5 Correlation Matrix

Section III – Principal Components Analysis

PCA Results

The dimension reduction process using PCA yields interesting initial results. The scree plot shown in Figure 6 does not indicate any pronounced inflections. However, closer examination reveals several points of interest at the third, fifth, and seventh principal component. Anything beyond the seventh point appears to have minimal benefits on the overall explanation of variability.

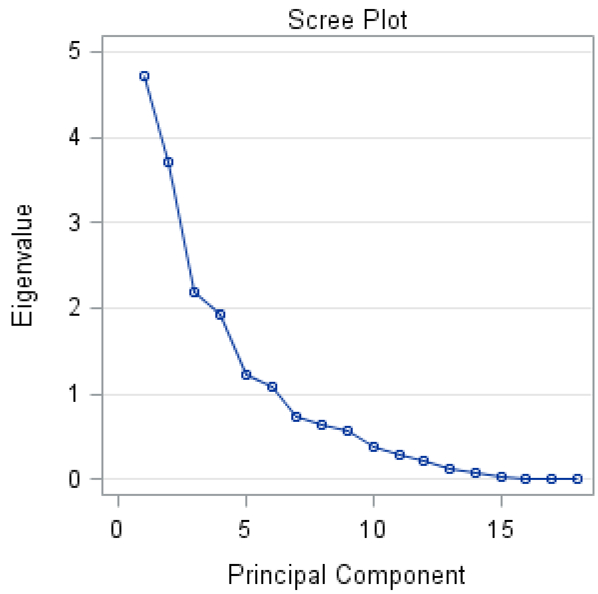

Figure 6 Scree Plot

In addition, Figure 7 shows each of the 18 eigenvalues and the cumulative percent of variation explained. Seven principal components, highlighted by the red rectangle, explain 86% of the variation and will be used in the initial regression analysis. However, it is unknown at this time if each of the seven principal components is significant and the final model may be altered accordingly.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.72757243	1.01198746	0.2626	0.2626
2	3.71558498	1.52375443	0.2064	0.4691
3	2.19183055	0.25624270	0.1218	0.5908
4	1.93558785	0.70604633	0.1075	0.6984
5	1.22954153	0.14497452	0.0683	0.7667
6	1.08456701	0.34582530	0.0603	0.8269
7	0.73874171	0.09220715	0.0410	0.8680
8	0.64653455	0.07024051	0.0359	0.9039
9	0.57629405	0.19445986	0.0320	0.9359
10	0.38183418	0.09677902	0.0212	0.9571
11	0.28505516	0.06136147	0.0158	0.9730
12	0.22369369	0.09453009	0.0124	0.9854
13	0.12916360	0.04360322	0.0072	0.9926
14	0.08556039	0.04481830	0.0048	0.9973
15	0.04074209	0.03576009	0.0023	0.9996
16	0.00498200	0.00253570	0.0003	0.9998
17	0.00244630	0.00217837	0.0001	1.0000
18	0.00026793		0.0000	1.0000

Figure 7 Eigenvalues

Explanation and themes can now be applied to each of the seven principal components. The eigenvector loadings are shown in Figure 8 with the selected first seven highlighted by the red rectangle. Principal component themes for the 2015 MLB data are as follows:

1. Moderate associations with several measures of offense: runs (R), hits (H), doubles (B2), runs batted in (RBI), and batting average (BA). The overall theme is more general and an indication of overall offensive performance.
2. This is an interesting triple (B3) and homeruns (HR) association. There is almost a perfect contrast between these two statistics. This would relate to the more dramatic and memorable events during a game and perhaps the need for high profile players.
3. Strong associations between intentional base on balls (IBB) and runners left on base (LOB). This is simply indicating more players on base (walked or left on base) is preferred.
4. Moderate associations between ground into double play (GDP) and runners left on base (LOB). There is nothing remarkable about this principal component and is later removed from the model due to statistical insignificance.
5. Stolen bases (SB) and sacrifice hits (SH) contrast with hit by pitch (HBP). This is a more tactical theme and is one of the more interesting of the principal components.
6. This is a strong single association with hit by pitch (HBP). This is another indication of players on base and perhaps more tactical. Note: This principal component is also later removed from the model due to statistical insignificance.
7. Moderate associations between hit by pitch (HBP), batting average (BA), and caught stealing (CS). No singular theme can be applied in this case.

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15	Prin16	Prin17	Prin18
AB	0.289586	0.197527	0.103112	-.160518	-.269707	0.084084	-.286697	-.385965	-.376431	0.133848	0.292985	-.136275	-.041713	0.386759	0.148151	-.200587	-.025601	0.227072
R	0.385895	-.237494	-.064789	-.069253	0.099710	0.078471	0.125125	-.152073	0.085840	-.048309	-.110088	0.127386	0.013229	0.272612	-.244635	0.211231	0.715786	0.038075
H	0.379805	0.257894	-.048400	0.008502	-.028322	0.059051	0.158351	-.100767	-.210247	0.006798	0.044044	0.123845	0.020564	-.261195	-.032241	-.210325	0.026385	-.755752
B2	0.351631	0.015456	0.095734	0.136391	-.241465	0.024303	-.039078	0.090254	0.530601	0.410054	-.009594	-.354701	-.422277	-.153468	0.046413	0.012849	-.001663	0.001446
B3	0.111016	0.371004	-.043325	-.274342	-.062486	0.116883	-.279656	0.091836	0.441840	-.219282	-.205816	0.456599	0.057482	0.124278	0.394541	0.010977	-.000388	-.000634
HR	0.139224	-.407769	-.020213	-.251537	0.106449	0.215152	0.196106	0.062685	-.234595	-.026354	-.167742	-.207802	-.042824	-.160031	0.700412	0.071531	0.021591	0.001146
RBI	0.382901	-.248230	-.067565	-.058436	0.078982	0.098866	0.114078	-.139117	0.075682	-.042251	-.120502	0.109840	0.051195	0.265672	-.224738	0.313375	-.690983	-.044024
SB	0.158371	0.237162	0.267569	-.243752	0.420177	-.206041	-.158510	0.091269	0.008591	0.249892	-.317708	-.376828	0.471789	0.000952	-.082624	-.016722	0.006540	0.003922
CS	-.008635	0.279269	0.243546	-.376105	0.006300	0.000441	0.307129	0.515643	-.213138	-.123965	0.050140	-.053515	-.434903	0.253059	-.158463	0.127680	-.007337	-.009462
BB	0.125382	-.301423	0.312179	0.318455	0.086106	0.200094	-.178081	0.321324	-.070913	-.050555	-.257733	0.162975	-.061153	0.225142	-.060068	-.593445	-.036320	-.006072
SO	-.124977	-.172968	0.375176	-.306187	0.028106	0.380100	0.080077	0.029477	0.122672	0.436745	0.398620	0.343780	0.213749	-.172929	-.093621	0.009701	0.002491	0.008523
BA	0.363002	0.244084	-.110377	0.070130	0.069205	0.023733	0.316320	0.026533	-.125485	-.031980	-.063946	0.215607	0.026405	-.443694	-.091041	-.190295	-.054092	0.610822
GDP	0.054926	0.213566	-.318650	0.388929	-.074243	0.296541	0.199658	0.411172	-.027568	0.215497	0.179897	-.101370	0.398265	0.317664	0.165339	0.135906	0.031991	-.001002
HBP	0.037996	-.128712	0.266825	-.002719	-.453912	-.573810	0.434329	-.002103	0.094441	0.055499	-.054166	0.148340	0.277344	0.166609	0.179603	-.113394	-.005079	-.009425
SH	-.140186	0.206764	0.130135	0.275925	0.541195	-.094978	0.291709	-.331827	0.053994	0.277876	-.004396	0.198037	-.294818	0.255071	0.274414	-.011844	-.021172	-.013114
SF	0.299285	-.137184	0.071600	0.075421	0.352791	-.317749	-.135098	0.195220	0.162478	-.333935	0.660773	-.038468	0.043014	-.032784	0.140103	-.017552	-.011236	0.002489
IBB	-.070545	0.167434	0.421284	0.143900	-.060052	0.395811	0.287950	-.277347	0.240706	-.500514	0.027862	-.336590	0.159002	-.005615	-.013844	-.021044	0.002795	0.014626
LOB	0.118664	0.076969	0.461772	0.392503	-.121015	-.039907	-.271044	0.068121	-.301573	-.009866	-.097010	0.194821	-.012205	-.190668	0.104635	0.574473	0.048911	0.009299

Figure 8 Eigenvector Loadings

Section IV – Regression Analysis

Model Selection

The initial regression model is represented by the following equation where P1 thru P7 represent the numerical principal component:

$$Wins=ß_0 + ß_1P_1 + ß_2P_2 + ß_3P_3 + ß_4P_4 + ß_5P_5 + ß_6P_6 + ß_7P_7$$

Model Fit

The model fit well with most of the parameters being significant at the α = 0.05 level. The overall fit was significant with a p-value of 0.0002. Additionally, the model explains about 68% (R^2=0.6843) of the variation in wins. The initial model fit statistics are presented in Figure 9.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	2168.45933	309.77990	6.81	0.0002
Error	22	1000.50734	45.47761		
Corrected Total	29	3168.96667			

Root MSE	6.74371	R-Square	0.6843
Dependent Mean	80.96667	Adj R-Sq	0.5838
Coeff Var	8.32899		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	80.96667	1.23123	65.76	<.0001
Prin1	1	1.57024	0.57594	2.73	0.0123
Prin2	1	-2.18947	0.64966	-3.37	0.0028
Prin3	1	2.47703	0.84586	2.93	0.0078
Prin4	1	1.37348	0.90011	1.53	0.1413
Prin5	1	-2.46886	1.12935	-2.19	0.0397
Prin6	1	-1.54813	1.20246	-1.29	0.2113
Prin7	1	4.95160	1.45698	3.40	0.0026

Figure 9 Initial Model Fit

Two principal components as shown in Figure 9 are not significant and will be removed from the model. Principal components four and six are considerably above the α = 0.05 level of significance, 0.1413 and 0.2113 respectively, and do not belong in the model. Therefore, the final model will be reduced from seven to five components as represented by the following equation:

$$Wins=ß_0 + ß_1P_1 + ß_2P_2 + ß_3P_3 + ß_5P_5 + ß_7P_7$$

The refit model is presented in Figure 10. The overall model is significant and each parameter is now significant. The R^2 slightly reduced and now explains about 63% (R^2=0.6271) of the variation in wins. While this is not spectacular, results explaining 68% of the variation with five independent principal components is far less complicated than interpreting 18 highly correlated individual variables.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1987.18699	397.43740	8.07	0.0001
Error	24	1181.77968	49.24082		
Corrected Total	29	3168.96667			

Root MSE	7.01718	R-Square	0.6271
Dependent Mean	80.96667	Adj R-Sq	0.5494
Coeff Var	8.66675		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	80.96667	1.28116	63.20	<.0001
Prin1	1	1.57024	0.59930	2.62	0.0150
Prin2	1	-2.18947	0.67601	-3.24	0.0035
Prin3	1	2.47703	0.88016	2.81	0.0096
Prin5	1	-2.46886	1.17515	-2.10	0.0463
Prin7	1	4.95160	1.51606	3.27	0.0033

Figure 10 Final Model Fit

Assumptions

The final model has been determined and the assumptions must be validated. One of the main purposes of PCA is to create independent linear combinations of the original variables. As expected, the scatterplot shown in Figure 11 indicate all five principal components are uncorrelated.

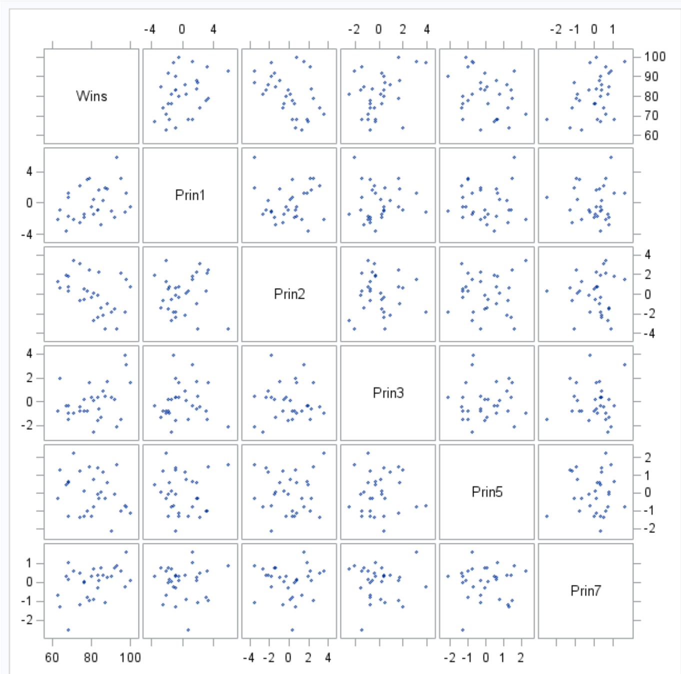

Figure 11 Scatterplot Matrix

The three plots contained in Figure 12 indicate three of four assumptions have been met. The QQ plot of the residuals indicates evidence of linearity. The histogram of the residuals indicates the residuals are normally distributed and scatter plot indicates fairly constant variance.

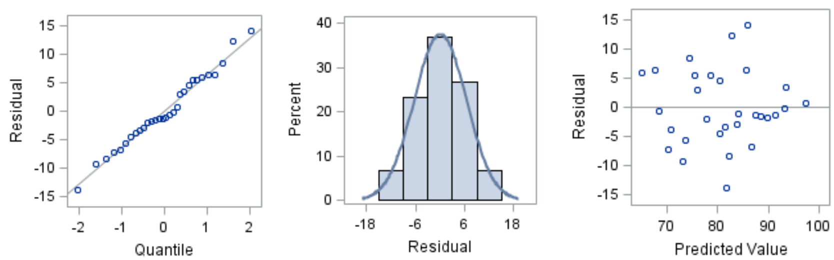

Figure 12 QQ Plot, Histogram, and Residual Plot

In addition it is important to determine if there are any high leverage data points or outliers in the dataset. The studentized residual plot shown in Figure 13 indicates there are no significant outliers or leverage points. There are three teams above 2 and below -2, but nothing significant. In addition, the two teams indicated as leverage are minor. The Toronto Blue Jays have the highest leverage but the value is less than 0.6 and is nothing of great concern. Lastly, the Cook's D plot also shown in Figure 9 indicates the Kansas City Royals as an outlier but it is not significant.

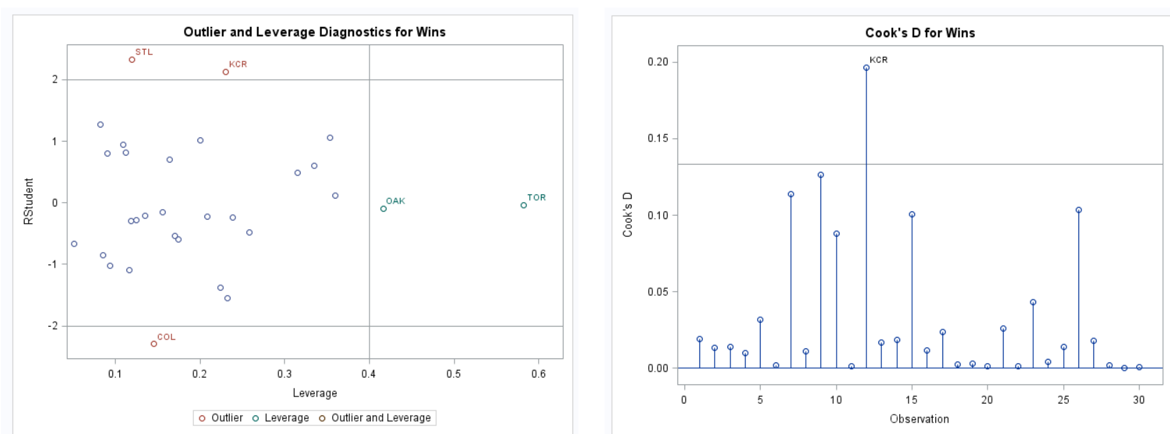

Figure 13 Studentized Residual and Cook's D Plots

Model Validation

Validation of the model is an important step in the process. The model was developed using the 2015 MLB regular season will be validated using the 2012 regular season. This is sometimes known as training and test dataset scenario. The training dataset is the original 2015 data and the test dataset is the 2012 data. The intent is to see if the model holds true for other years of play. The average square error (ASE) for each dataset is plotted on Figure 14. The zero step is the intercept and one through five are each of the principal components.

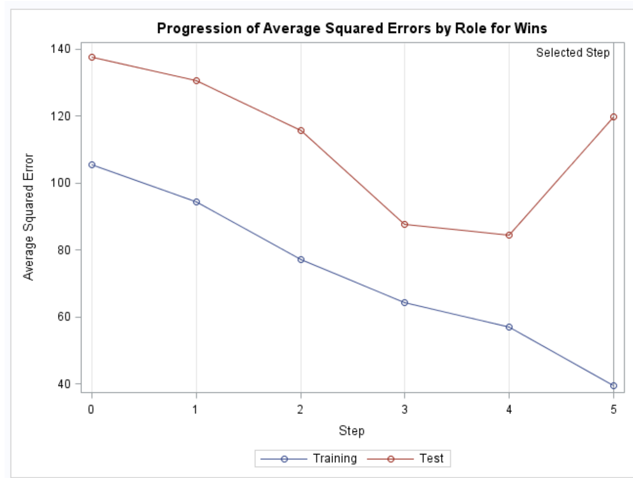

Figure 14 Model Validation - ASE

Ideally, the lines for the training and test datasets would be identical indicating the model validated extremely well. However, in this case the slopes of the lines are nearly identical but they diverge at the fifth principal component. The overall ASE value for the training dataset is 39.3 while the test dataset is considerably higher at 119.7. This is a threefold increase in ASE indicating the model did not validate as well as expected. The model may need refinement such as dropping the fifth principal component or trying different methods entirely. Further research and analysis is suggested and required in order to better validate the model.

Conclusions

Major League Baseball is a simple yet complicated sport. This analysis indicates some of the complexities involved in winning games. Principal components analysis is a more generic theme like approach to variable reduction, therefore, the following is a general summarization MLB coaches may find useful.

- The first principal component shows strong offensive play results in more wins. Not all that intriguing but does indicate team play is the most important.
- The second reveals the importance of the more exciting events in the game of baseball and perhaps the need for high profile players. Homeruns and triples are indeed important and exciting!
- The third indicates the need to get players on the bases regardless of method (walk or otherwise). Get players on base is the theme!
- The fifth principal component is an interesting tactical aspect of baseball. Indicating the importance of stolen bases and sacrifice hits.
- The seventh is a more general unassignable theme.

The above note five principal components explain about 63% (R^2=0.6271) of the variation in wins. While this is certainly not a guaranteed path to success, coaches certainly have a better understanding of what is required to win the game of baseball.

Appendix

```
ods graphics on;

* Means plots - all estimated and combined stats variables were removed ;
proc means data =MLB2015 n mean median std var min max maxdec=2;
var Wins AB R H B2 B3 HR RBI SB CS BB SO BA GDP HBP SH SF IBB LOB;
run;

* Custom template to color the correlation matrix;
proc template;
      edit Base.Corr.StackedMatrix;
         column (RowName RowLabel) (Matrix) * (Matrix2);
         edit matrix;
            cellstyle _val_  = -1.00 as {backgroundcolor=CXEEEEEE},
                      _val_ <= -0.75 as {backgroundcolor=red},
                      _val_ <= -0.50 as {backgroundcolor=blue},
                      _val_ <= -0.25 as {backgroundcolor=cyan},
                      _val_ <=  0.25 as {backgroundcolor=white},
                      _val_ <=  0.50 as {backgroundcolor=cyan},
                      _val_ <=  0.75 as {backgroundcolor=blue},
                      _val_ <   1.00 as {backgroundcolor=red},
                      _val_  =  1.00 as {backgroundcolor=CXEEEEEE};
            end;
         end;
 run;

 * Correlation plots - all estimated and combined stats variables were
removed ;
proc corr data=MLB2015 plots=matrix(histogram);
var Wins AB R H B2 B3 HR RBI SB CS BB SO BA GDP HBP SH SF IBB LOB;
run;

ods graphics / reset width=12in height=12in;
proc sgscatter data=MLB2015;
title "MLB 2015 Dataset";
matrix Wins AB R H B2 B3 HR RBI SB CS BB SO BA GDP HBP SH SF IBB LOB;
run;
ods graphics / reset;

* 2015 Principal components using only a subset of the data - all estimated
and combined stats variables were removed ;
proc princomp plots=all data=MLB2015 out=pca15;
var AB R H B2 B3 HR RBI SB CS BB SO BA GDP HBP SH SF IBB LOB;
id Tm;
run;

* 2014 Principal components ;
proc princomp plots=all data=MLB2014 out=pca14;
var AB R H B2 B3 HR RBI SB CS BB SO BA GDP HBP SH SF IBB LOB;
id Tm;
run;

* 2013 Principal components ;
proc princomp plots=all data=MLB2013 out=pca13;
var AB R H B2 B3 HR RBI SB CS BB SO BA GDP HBP SH SF IBB LOB;
```

```
id Tm;
run;

* 2012 Principal components ;
proc princomp plots=all data=MLB2012 out=pca12;
var AB R H B2 B3 HR RBI SB CS BB SO BA GDP HBP SH SF IBB LOB;
id Tm;
run;

* PCA regression analysis ;
proc corr data=pca15 plots=matrix(histogram);
var wins prin1 - prin3 prin5 prin7;
run;

proc sgscatter data=pca15;
matrix Wins prin1 - prin3 prin5 prin7;
run;

* Regression using prin comp 1 thru 7;
proc reg data=pca15;
model wins= prin1-prin7;
run;

* Final regression using prin comp 1, 2, 3, 5, and 7;
proc reg data=pca15;
model wins= prin1-prin3 prin5 prin7;
run;

* Validating the 2015 model data to the 2014 data;
proc glmselect data=pca15 testdata=pca14 seed=1
plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);
model wins= prin1-prin3 prin5 prin7 prin9 / selection=none  ;
run;

* Validating the 2015 model data to the 2013 data;
proc glmselect data=pca15 testdata=pca13 seed=1
plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);
model wins= prin1-prin3 prin5 prin7 prin9 / selection=none  ;
run;

* Validating the 2015 model data to the 2012 data - this one used in the
final analysis;
proc glmselect data=pca15 testdata=pca12 seed=1
plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);
model wins= prin1-prin3 prin5 prin7 / selection=stepwise  ;
run;
```