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.



Figure 1 List of Variables and Descriptions



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.

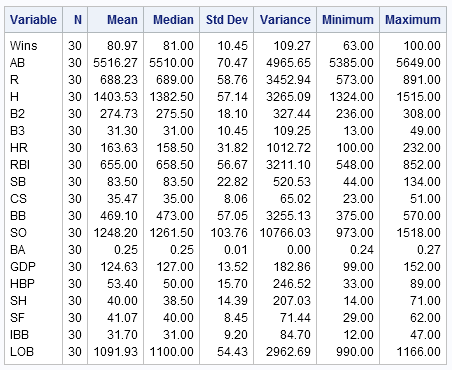


Figure 3 Summary Statistics

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.

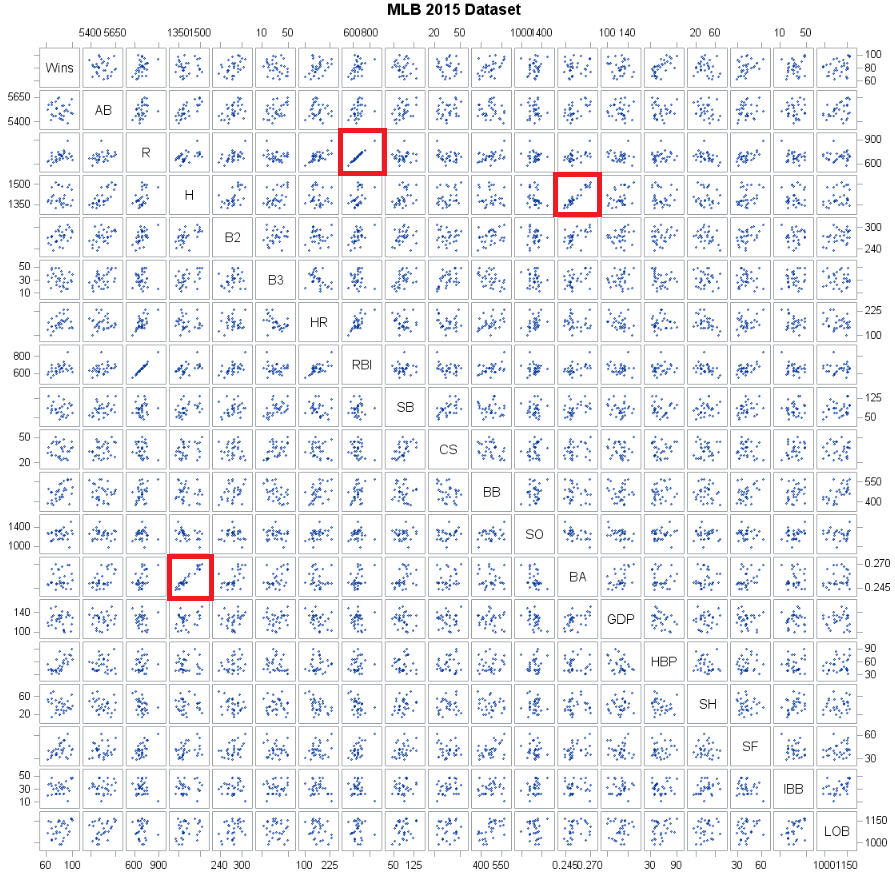


Figure 4 Scatterplot Matrix of all Variables

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.

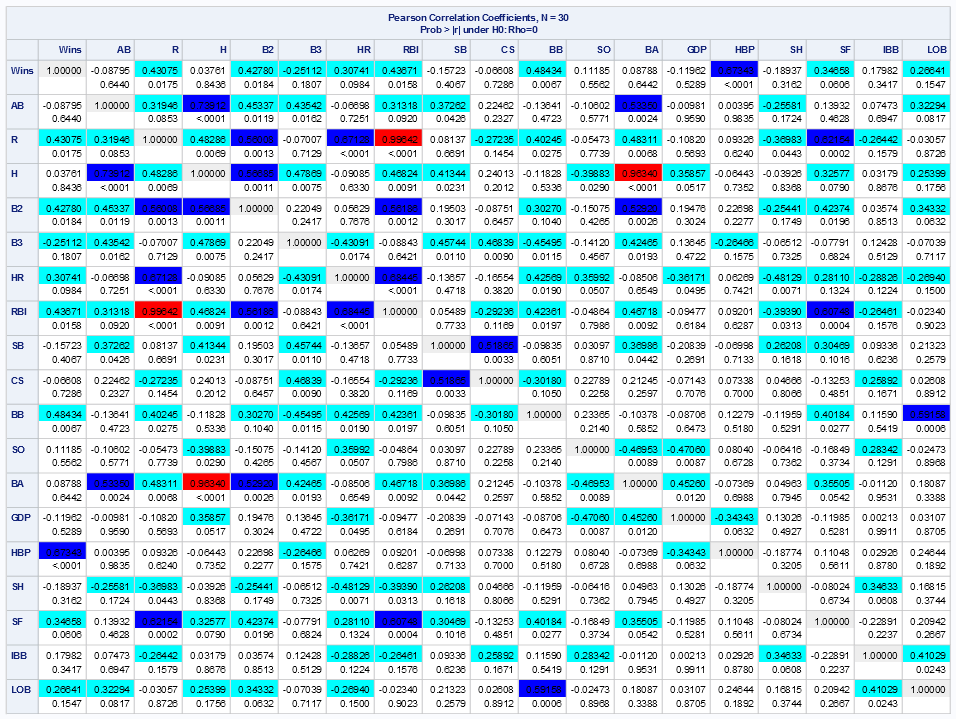


Figure 5 Correlation Matrix

**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.

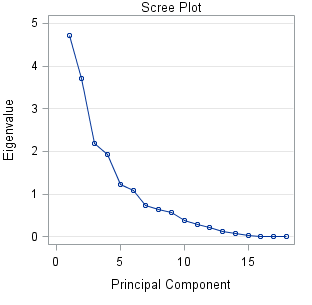
****

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.

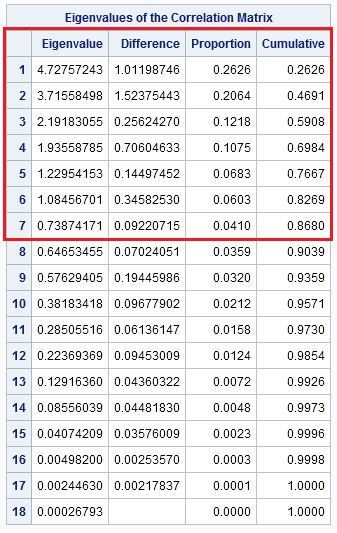
****

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.

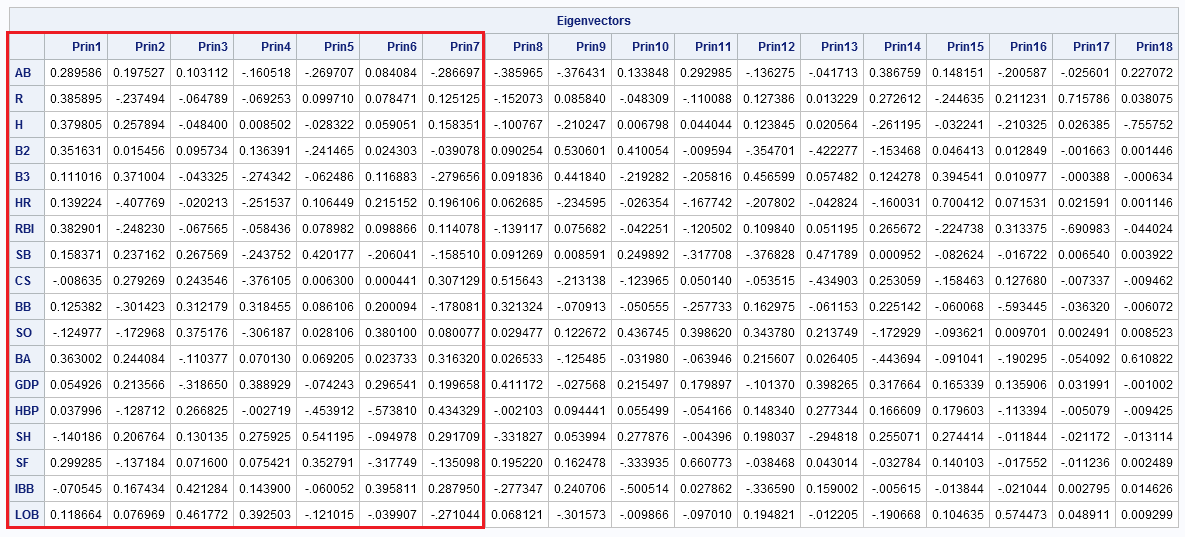
****

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=ß0 + ß1P1 + ß2P2 +ß3P3 + ß4P4 + ß5P5 + ß6P6 + ß7P7

**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% (R2=0.6843) of the variation in wins. The initial model fit statistics are presented in Figure 9.

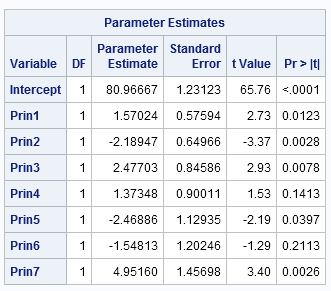
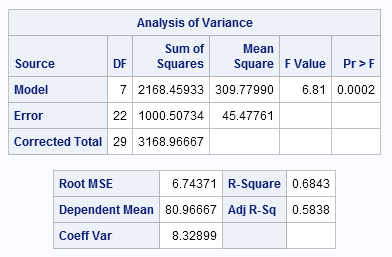


Figure 9 Initial Model Fit

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=ß0 + ß1P1 + ß2P2 +ß3P3 + ß5P5 + ß7P7

The refit model is presented in Figure 10. The overall model is significant and each parameter is now significant. The R2 slightly reduced and now explains about 63% (R2=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.

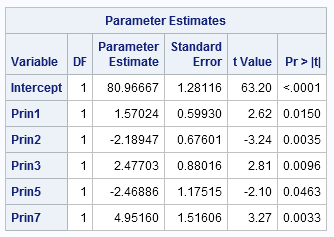
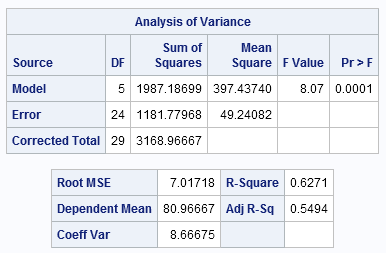


Figure 10 Final Model Fit

**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.

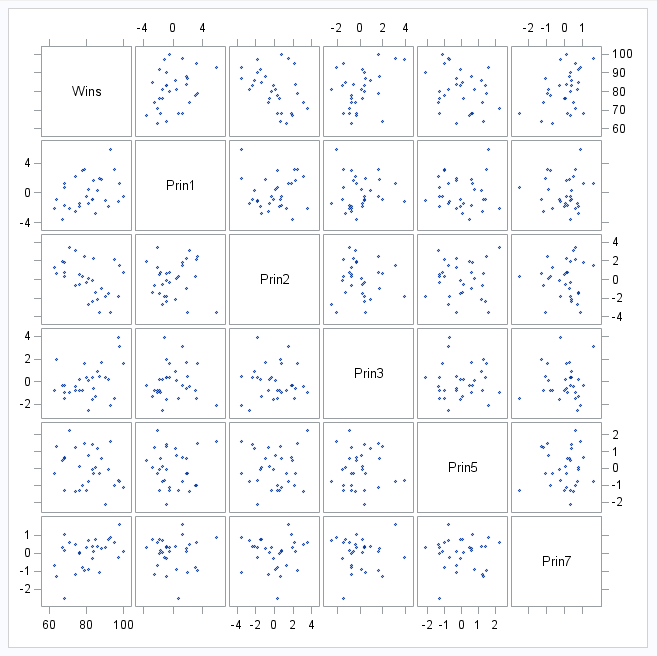


Figure 11 Scatterplot Matrix

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.

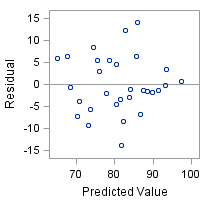
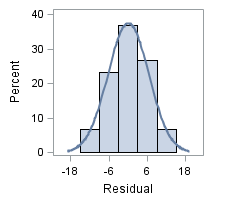
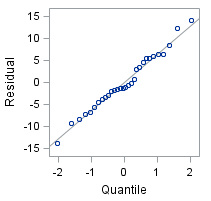


Figure 12 QQ Plot, Histogram, and Residual Plot

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.

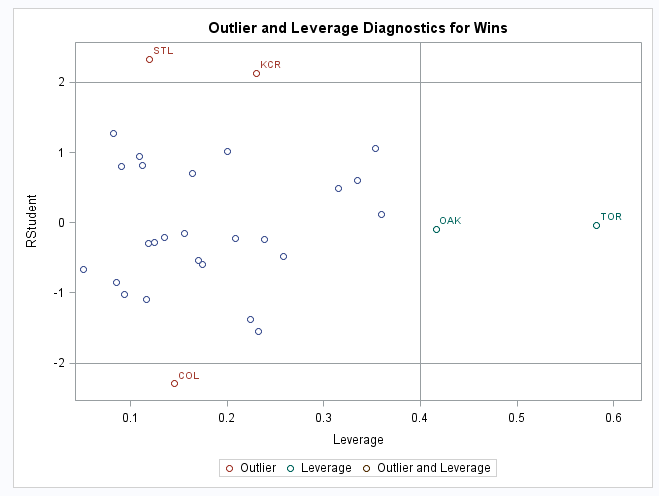
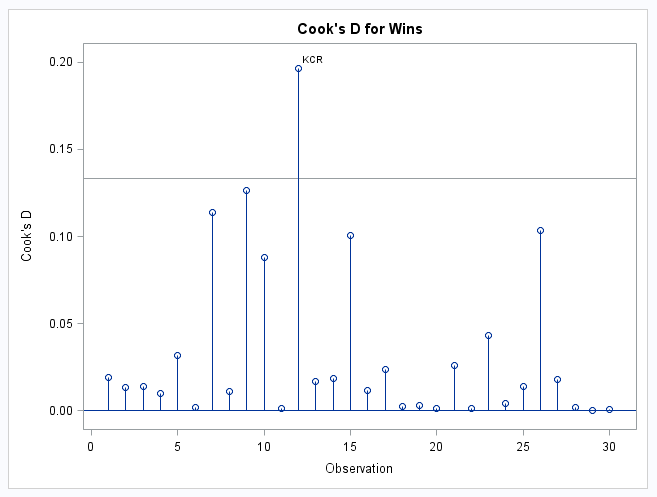
 

Figure 13 Studentized Residual and Cook’s D Plots

**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.

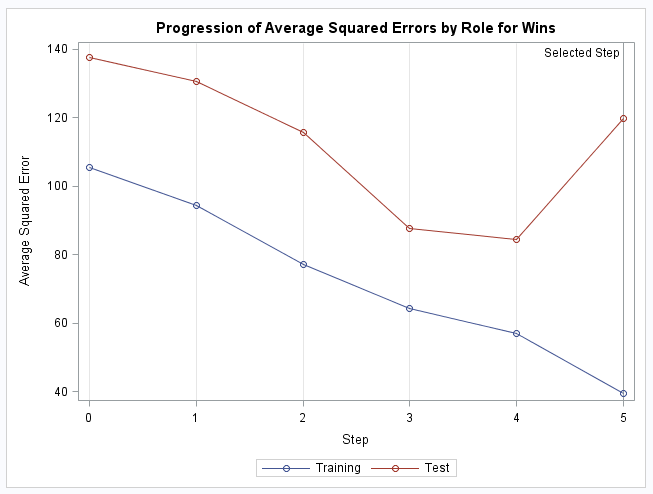


Figure 14 Model Validation - ASE

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% (R2=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=**12**in height=**12**in;

**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**;