Investigation of MLB Data with Multivariate Statistics

Vincent Milano
Statistics Department
Cal Poly State University, San Luis Obispo, CA

October 2009

Table of Contents

I. Introduction

The game of baseball is one of the most popular sports in the United States, and is even known as 'America's Pastime.' Major League Baseball has been going on for over 100 years with a strong fan base throughout. Although football appears to be the dominant sport in the United States, there was a time when baseball was considered in the highest regard and there are many fans toady that have a great affection for the game.

For as long as I can remember I have been a big sports fan, and baseball was always my favorite. Everyone in my family is a big fan, and baseball season has always been a time when I am feeling just a little happier. I remember being young, rooting for the Oakland A's and seeing greats like Ricky Henderson and Mark McGuire grace the field. Ever since those days I knew I wanted to be as close to the sport as I could be.

I used to play baseball when I was young, but my lack of ability eventually halted that dream. Despite this, I still wanted to be involved with something to do with the sport. Studying statistics in college made me wonder how the techniques I have learned can be implemented on the various statistics recorded in Major League Baseball.

Being that there are a lot of recorded statistics in baseball, it occurred to me that the use of multivariate statistical analysis would apply well to data taken from Major League Baseball. This is how I decided on what I wanted to do for my senior project. My initial goals were trying to predict player and team performance, but my project evolved into more of an exploratory analysis of baseball data using multivariate techniques.

Throughout working, I had three main goals that I wanted to accomplish. The first was to investigate team success. I wanted to find out which variables were most related to team success as well as how much they influence team success. The second main goal was to explore the differences (if any) between the American and National Leagues in Major League Baseball. The third goal was to develop a model to predict player performance for the next upcoming season. The following sections will further discuss the data and the analyses I performed.

II. Background

There were three main data sets used for my exploration. The primary data set is season totals of various player statistics for every player in the league for the 2008 Major League Baseball Season. Another of the main data sets is the season totals for every player in the 2007 Major League Baseball Season. The last data set is season total across the same variables for every team in the league for the 2008 season. The following are all the variables considered in my project and a brief explanation of the variable. All the variables are measured in totals except for AVG, OBP, SLG, and OPS, which are all averages. Also, the League variable is quantitative binary.

Variable	Explanation
Games	Amount of games played (where there are 162 games in a season)
At-Bats (AB)	Amount of times a player came up to the plate to bat
Runs	Amount of runs (points) scored
Hits	Amount of hits
Singles (1B)	Amount of hits that result in the player reaching first base
Doubles (2B)	Amount of hits that result in the player reaching second base
Triples (3B)	Amount of hits that result in the player reaching third base
Homeruns (HR)	Amount of homeruns
Extra Base Hits (Xtra)	All hits that result in extra bases, i.e. all hit that aren't singles
Runs Batted In (RBI)	Amount of runs that the player hits in
Total Bases (TB)	Total amount of bases touched by the player
Walks	Amount of walks player receives (batter is thrown four balls)
Strikeouts (SO)	Amount of strikeouts (batter is thrown three strikes)
Stolen Bases (SB)	Amount of bases player steals off of the defense

Caught Stealing (CS)	Amount of times a player is tagged out while trying to steal a base
On-Base-Percentage (OBP)	Percentage of time the player reached base when up to bat
Slugging Percentage (SLG)	TB divided by AB. Measure of power.
On-Base Plus Slugging Percentage (OPS)	OBP and SLG added together
Batting Average (AVG)	Percentage of time player gets a hit when up to bat versus recording an out
League	League the player belongs to. Player is in either the American League or National League depending on what team they are on.

For the data sets measured on the individual players, only players with over 100 at-bats were considered. This was done so that only players who had some significant playing time and contributions to the team were in the data. The cutoff of 100 at-bats was chosen somewhat arbitrarily, but in baseball many times a player's stats aren't considered 'official' unless they've had a minimum of 100 at-bats for the season. The main 2008 data set has 440 observations.

The data was found at the website MLB.com. Unfortunately, there was no digital file containing all the data, so it was necessary to copy and paste over fifty pages of data onto a spreadsheet. This led to some problems in that many lines and columns in the data were being skipped after the data was transferred to the spreadsheet. In order to get the data in a workable form, it was necessary for me to use the SAS and R statistical software packages and write some code. The coding was a challenging and time consuming process, but I was eventually able to create the right combination of R and SAS code to fix the entire data set.

Since there are so many variables measured, it allowed me the opportunity to analyze the data within many different combinations of variables. For this report, we will explore the data for three different combinations of the variables. For the sake of ease, I will refer to these combinations of variables as Groups A, B, and C. Group A will contain the variables of Games, 2B, 3B, HR, Walks, OBP, and SLG. Group B will have Games, RBI, 1B, Xtra, SO, OPS. Lastly, Group C will have the variables Games, Hits, OPS, AVG, Runs, SB, and CS.

III. Analysis of Team Success

One of the areas I wanted to investigate with this data is trying to find out how the variables are related to team success. In order to do this we will be looking at the data set that contains the season totals of the variables for each team. We can perform a Regression Analysis using team win percentage as the response variable. The Regression will result in a model that give us information on which variables contribute most to team success as well as a way to predict win percentage given certain team statistics. There are thirty teams in the league and so thirty observations in the data. Since there are so few observations, any model I come up with shouldn't have more than three or four terms in order to save degrees of freedom as well as not have too many effects in the model for a limited data set. Below is the best three variable model from the regression analysis.

```
Regression Analysis: WinPct versus AB, Runs, SB

The regression equation is
WinPct = 2.12 - 0.000401 AB + 0.000701 Runs + 0.000921 SB


                                                    Standardized
Predictor         Coef    SE Coef      T      P  Coefficients
Constant        2.1154     0.6671   3.17  0.004
AB          -0.0004013  0.0001276  -3.15  0.004   -0.420798
Runs         0.0007010  0.0001314   5.33  0.000    0.719146
SB           0.0009206  0.0002724   3.38  0.002    0.394807


S = 0.0403048   R-Sq = 67.3%   R-Sq(adj) = 63.5%
```

Calculations for the standardized coefficients (Appendix: Regression1A pg 20).

Before we can use the results of the regression analysis, there are assumptions that must be satisfied in order for the model to be valid. The first main assumption is that the observations are independent of each other. The second assumption is that the residuals are normally distributed. The next assumption is that there is equal variance across the data. These assumptions are satisfied (Appendix: Regression1B pg 20) and so we can consider the inference results based on the model to be valid.

The three variable model shows that the variables AB, Runs, and SB are most responsible for predicting team success. The p-values show that the three variables are significant predictors of win percentage. The null hypothesis in each case is that the coefficient is equal to zero. The alternative is that the coefficient is not zero. At the five percent level, these three variables were found to be significant. Possible interactions were investigated, but none were found to be significant in the models.

Reported are the regression model coefficients as well as the standardized coefficients. The standardized coefficients tell us which variables contribute most to the model for win percentage. Variables with larger standardized coefficients in magnitude contribute more to the model. For this model, Runs contributes most followed by AB and then SB. Note that the coefficient for AB is negative, implying that team at-bats has a negative association with win percentage. This means that a lower number of AB (at-bats) is associated with a higher win percentage versus a larger number of AB which is associated with a smaller win percentage.

The R^2 term is known as the Coefficient of Determination and is a measure of how well the model fits the data. The adjusted R^2 (R-Sq(adj) in output) is used to compare models that do not have the same amount of terms. Values closer to 100% imply the model fits the data very well. For this model, the adjusted R^2 value is 63.5%. This value is not as high as we would like to see, but it still shows the model is somewhat sufficient. A value higher than 75 or 80% would be considered a 'good' fit, but at 63.5% the model is an ok fit to the data meaning that the explanatory variables do a reasonable job of predicting win percentage.

This next analysis shows the best four variable model for win percentage. The only difference with this model is the addition of the variable AVG. The AVG variable has a low p-value, but still not low enough to be considered a significant predictor. Despite this, the addition of the variable raises the adjusted R^2 value of the model, and so I feel it is worth including.

Regression Analysis: WinPct versus AB, SB, Runs, AVG

```
The regression equation is
WinPct = 2.61 - 0.000565 AB + 0.000834 SB + 0.000593 Runs + 1.91 AVG


                                                 Standardized
Predictor         Coef    SE Coef       T      P   Coefficients
Constant        2.6102     0.7274    3.59  0.001
AB          -0.0005649  0.0001645   -3.43  0.002     -0.592347
SB           0.0008341  0.0002718    3.07  0.005      0.357711
Runs         0.0005931  0.0001465    4.05  0.000      0.608453
AVG              1.911      1.255    1.52  0.140      0.300205


S = 0.0393209   R-Sq = 70.0%   R-Sq(adj) = 65.3%
```

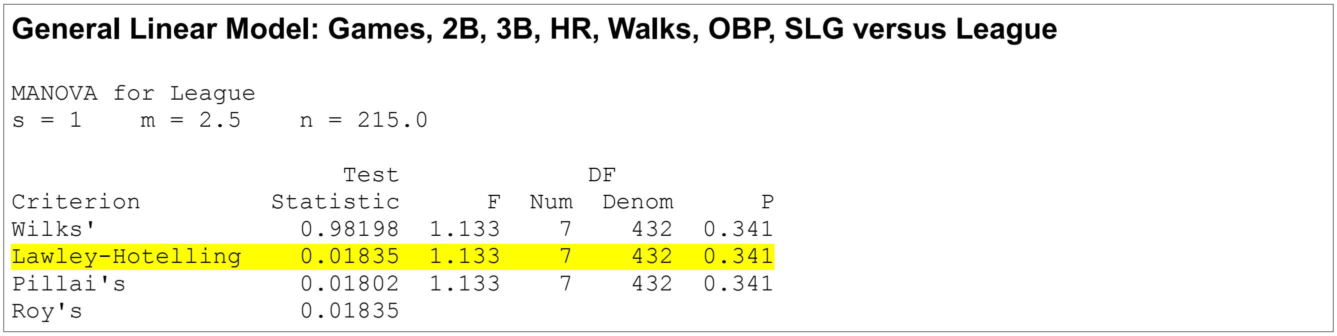
Calculations for the standardized coefficients in appendix (Appendix: Regression2A pg 21).

The main regression assumptions are satisfied for this model (Appendix: Regression2B pg 21) and thus we can further consider the model.

Though the AVG term has a higher p-value than what is considered acceptable, the p-value is still relatively low for a smaller data set. Also, including the AVG variable results in a model with an adjusted R^2 value of 65.3%, which is a bit higher than the value for the three variable model. This implies that the four term model may be a better fit for the data than the three term model. Again, this value Looking at the standardized coefficients we see that AB and Runs contribute most to the model over SB and AVG. In this model, the AB variable has a negatively affects win percentage just as in the three variable model.

IV. Analysis of League Differences

Here I aimed to explore the differences, if any, between the players in the American and National Leagues. To be more specific, I wanted too see if there were differences between the leagues with respect to the offensive characteristics as well as find out where those differences lie. First, we can use Hotelling's T^2 test in order to see whether there is a difference between the leagues across the offensive characteristics. The Hotelling's T^2 is the multivariate form of the student t-test. With the t-test we can compare the average of one variable across two levels, whereas with Hotelling's T^2 we can compare the averages of multiple variables across two levels. So to investigate if there are differences between the American and National Leagues, I applied the Hotelling's T^2 test to each group of variables. The test compares the multivariate mean for each League, which is a vector of means for each of the variables. The null hypothesis is that the multivariate mean for the American League is equal to the multivariate mean of the corresponding variables in the National league. The alternative hypothesis is that the multivariate means are not equal for the two leagues. If a difference is detected, then I will use Linear Discriminant Analysis to further investigate why there is a difference. Below is the Minitab software output of the Hotelling's T^2 test applied to Group A.

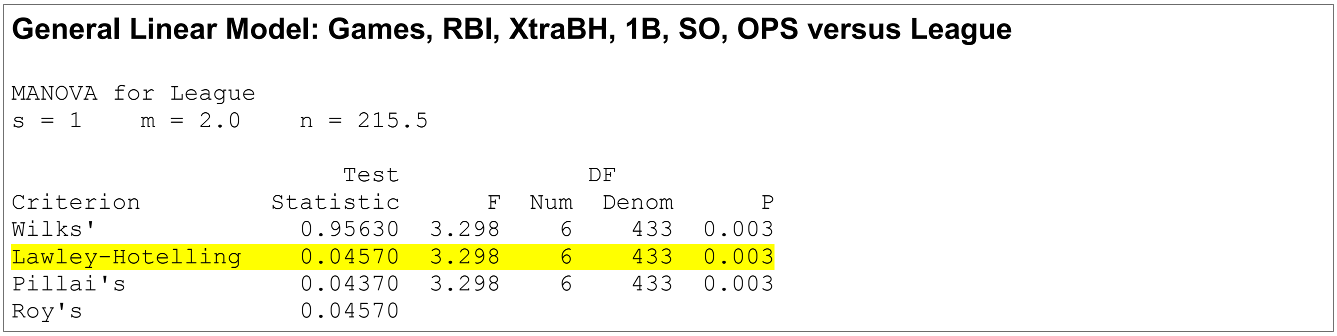


General Linear Model: Games, 2B, 3B, HR, Walks, OBP, SLG versus League

```
MANOVA for League
s = 1    m = 2.5    n = 215.0

                        Test              DF
Criterion          Statistic       F  Num  Denom       P
Wilks'               0.98198  1.133    7    432  0.341
Lawley-Hotelling     0.01835  1.133    7    432  0.341
Pillai's             0.01802  1.133    7    432  0.341
Roy's                0.01835
```

As with other significance tests, a low p-value demonstrates a significant difference. If the value is low, or below 0.05, then we reject the null hypothesis and say that a significant different is present. For Group A, the Hotelling's T^2 test statistic is 0.01835 and the corresponding p-value is 0.341. This p-value is not low, so we do not reject the null hypothesis and thus we cannot conclude that differences exist between the two leagues for Group A. For this reason there is no need to further analyze Group A with respect to American/National League differences.

The following Minitab output shows the results of the Hotelling's T^2 test applied on the variables in Group B.



General Linear Model: Games, RBI, XtraBH, 1B, SO, OPS versus League

```
MANOVA for League
s = 1    m = 2.0    n = 215.5

                       Test             DF
Criterion         Statistic       F  Num  Denom      P
Wilks'              0.95630  3.298    6    433  0.003
Lawley-Hotelling    0.04570  3.298    6    433  0.003
Pillai's            0.04370  3.298    6    433  0.003
Roy's               0.04570
```

For Group B, the Hotelling's T^2 test statistic is 0.04570 and the p-value is 0.003. This p-value is very low and we can reject the null hypothesis. So there is a significant difference between the American and National Leagues for the variables in Group B.

To further investigate this difference found between the leagues, we can use Linear Discriminant Analysis. The discriminant analysis yields a linear combination of the variables which best separates the two leagues. This linear combination is called the discriminant function and gives information on the group differences. Using the discriminant function we can determine which variables contribute to the group separation as well as which variables contribute most to the group separation. This analysis was done using the R software and the code can be found in the appendix (Appendix: Discriminant pg 22). Table 1 on the next page shows the results of the discriminant analysis.

Table 1 Discriminant Analysis on Group B

Variable	Function Coefficients	Standardized Coefficients	F-Statistic	p-value
Games	0.089668	6.940590	12.894951	0.0004
RBI	-0.008305	-5.030114	5.078477	0.0247
SO	-0.001933	-1.472584	1.245750	0.2650
OPS	0.999887	2.332757	4.183092	0.0414
Xtra	0.004782	2.011772	0.601684	0.4384
1B	-0.007001	-4.826380	10.224344	0.0015

The regular function coefficients do not give too much information, but the standardized coefficients tell us how much each variable contributes to the American and National League difference. The standardized coefficients are not standardized to any scale, like a z-score, but instead they are compared with respect to the other standardized coefficients. The coefficients larger in magnitude contribute more to the group difference. A partial-F test was performed to determine which variables are significant in explaining the league separation and the corresponding test statistics and p-values are reported in the table. The variables Games, RBI, OPS, and 1B had p-values less than 0.05 and thus were found to be significant in explaining the league difference. Looking at the standardized coefficients, we see that the order of importance of the variables is Games, RBI, 1B, then OPS.

The discriminant function can also be used to classify new observations into either American or National League based on their characteristics, if the league is unknown. Though this is not the case, we can still apply the discriminant function to the players in the data set and check how accurately the function classifies the players into each league. This can be done with the R software and the results are on the next page

Confusion Matrix

Predicted Actual	AL	NL
AL	99	108
NL	84	149

Correct Classification Rate: 0.5636
Error Rate: 0.4364

The 'Confusion Matrix' displays how many players were correctly classified and wrongly classified into the two leagues. Only 99 American League and 149 National League players were classified correctly whereas 108 American League and 84 National League players were classified incorrectly. The Correct Classification Rate measures the percentage of players that were classified correctly. For Group B, the Correct Classification Rate is 0.5636, meaning about 56 percent of the players were correctly classified and about 44 percent of the players were incorrectly classified. These results show that the discriminant function for Group B is not very effective at classifying the players into the leagues. If the discriminant function had no power to classify the players, we would expect the Correct Classification Rate and Error Rate to both be around 0.5, or a pure 50/50 chance. A rate of 0.5636 is not very high in comparison and unfortunately this suggests that the discriminant function is not very reliable and gives little information about Group B.

The following Minitab output shows the results of the Hotelling's T^2 test applied on the variables in Group C.

General Linear Model: Games, Hits, OPS, AVG, Runs, SB, CS versus League

```
MANOVA for League
s = 1    m = 2.5    n = 215.0

                        Test             DF
Criterion          Statistic      F  Num  Denom      P
Wilks'               0.95988  2.579    7    432  0.013
Lawley-Hotelling     0.04180  2.579    7    432  0.013
Pillai's             0.04012  2.579    7    432  0.013
Roy's             0.04180
```

For Group C, the Hotelling's T^2 test statistic is 0.04180 and the p-value is 0.013. This p-value is also very low and we can reject the null hypothesis. So there is a significant difference between the American and National Leagues for the variables in Group C.

Since a difference is found, once again we can use Linear Discriminant Analysis to further investigate the league differences. Table 2 below shows the results of the discriminant analysis using the R software.

Table 2 Discriminant Analysis on Group C

Variable	Function Coefficients	Standardized Coefficients	F-Statistic	p-value
Games	0.005862	4.398569	11.2536	0.0086
Runs	-0.006024	-3.575777	3.6965	0.0551
Hits	-0.002523	-2.614048	1.4551	0.2283
SB	0.004285	0.838213	1.0975	0.2954
CS	0.003534	0.197482	0.0656	0.7980
AVG	-0.244581	-0.177531	0.0306	0.8613
OPS	0.969574	2.262036	5.1184	0.0241

As Table 2 shows, only Games and OPS are significant in the discriminant function, and thus are they only two variables that significantly explain the league separation. Looking at the standardized coefficients, we see that Games is the more important variable in explaining the differences in the leagues.

Looking at the Confusion Matrix and classification rates for Group C we can see how well the discriminant function classifies the data.

Confusion Matrix

Predicted Actual	AL	NL
AL	99	108
NL	85	148

Correct Classification Rate: 0.5614
Error Rate: 0.4386

For Group C, the discriminant function correctly classified only 99 American League and 149 National League players whereas 108 American League and 84 National League players were classified incorrectly. In this case, the Correct Classification Rate is 0.5614, meaning about 56 percent of the players were correctly classified and about 44 percent of the players were incorrectly classified. Similar to the results for Group B, these numbers show that the discriminant function for Group C is not very effective at classifying the players into the leagues. The Correct Classification and Error Rates are both close to 0.5, unfortunately meaning that the function has little ability to classify the players.

V. Principal Component Analysis

In this section, I performed a Principal Component Analysis to learn about the linear combinations of the variables that explain variation in the data. This analysis yields linear combinations of the variables that are known as principal components. The first principal component, or first linear combination, explains the most variability in the data. The second principal component explains the second most variability in the data, and so on. Usually only the first two or three principal components are considered, in this case we will explore the first three. The principal components will show the combinations of the offensive statistics that are most responsible for the variability in the data. Observe the results of the principal component analysis for Group A.

Principal Component Analysis for Group A

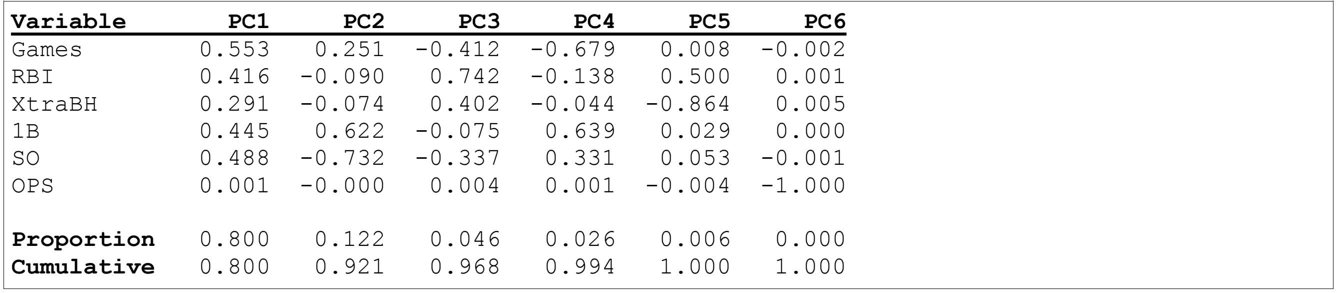
```
Variable        PC1     PC2     PC3     PC4     PC5     PC6     PC7
Games         0.850  -0.483  -0.183   0.098  -0.019   0.001  -0.000
2B            0.224  -0.006   0.704  -0.671  -0.057  -0.003   0.001     NOTE:  All Principal
3B            0.022  -0.015  -0.008  -0.086   0.996  -0.004   0.002     Components were
HR            0.162   0.186   0.643   0.722   0.066  -0.005   0.004     extracted from the
Walks         0.447   0.855  -0.238  -0.107  -0.008  -0.001  -0.001     covariance matrix
OBP           0.000   0.001   0.000  -0.001   0.001   0.506   0.862     of the data.
SLG           0.001   0.002   0.007   0.002   0.005   0.862  -0.506

Proportion    0.857   0.103   0.024   0.013   0.002   0.000   0.000
Cumulative    0.857   0.961   0.985   0.998   1.000   1.000   1.000
```

The row labeled 'Proportion' shows the proportion of the variability in the data that each principal component explains with the cumulative totals in the row underneath. The first principal component is responsible for 85.7% of the variability, the second is responsible for 10.3%, and the third responsible for 2.4%. Together the first three principal components explain 98.5% of the data. The coefficients of the principal components are measured on a scale of -1 to 1, and the coefficients that are larger in magnitude contribute most to the principal component. For example, looking at the first principal component we can see that the variables Games and Walks have larger coefficients than the rest of the variables. So this principal component emphasizes Games and Walks. Since each principal component is a linear combination, we can consider each principal component as a single variable itself. By looking at what is emphasized in each principal component we can determine what the variable is measuring. The first principal component has a positive emphasis on Games and Walks, so this variable can be considered as 'Walk Production.' The second principal component is a contrast between Games and Walks, where the coefficient of Games is negative. This principal component emphasizes more walks to less games. This variable will be considered as 'Walk Efficiency.' The third component is a positive emphasis on HR and 2B. These two variables are very common measurement of power, so this variable will be 'Power.'

The principal component analysis on Group A showed that the variables 'Walk Production,' 'Walk Efficiency,' and 'Power' are responsible for the most variability in the data. Later on, I explore how the principal components can be considered as predictors for regression.

Below are the results of the principal component analysis for Group B.

Principal Component Analysis for Group B



Variable	PC1	PC2	PC3	PC4	PC5	PC6
Games	0.553	0.251	-0.412	-0.679	0.008	-0.002
RBI	0.416	-0.090	0.742	-0.138	0.500	0.001
XtraBH	0.291	-0.074	0.402	-0.044	-0.864	0.005
1B	0.445	0.622	-0.075	0.639	0.029	0.000
SO	0.488	-0.732	-0.337	0.331	0.053	-0.001
OPS	0.001	-0.000	0.004	0.001	-0.004	-1.000
Proportion	0.800	0.122	0.046	0.026	0.006	0.000
Cumulative	0.800	0.921	0.968	0.994	1.000	1.000

As shown in the table, the first principal component accounts for 80% of the variability in the data, the second accounts for 12.2%, and the third accounts for 4.6%. Cumulatively, the first three principal components are responsible for 96.8% of the variability in the data.

Looking at the first principle component we can see that most of the variable coefficients are close except for OPS which doesn't contribute much at all. So this principal component can be considered as a weighted average of offensive statistics. Since this principal component is a weighted average, it is reasonable to say that it is a measure of overall offensive performance. The first principal component will be the variable 'Overall Performance.' The second principal component appears to have a strong contrast between 1B (base hits) and SO (strikeouts). It is measuring hitting versus striking out, so we can consider it a measurement of ability to get a hit. This variable will be 'Ability to Hit.' The third principal component has a strong contrast of RBI and XtraBH (extra base hits) versus Games and SO.

The variables RBI and XtraBH emphasize scoring runs, while Games with SO emphasizes recording outs. Since this principal component seems to be measuring score versus recording outs, this variable will be 'Ability to Score.'

For Group B, the principal component analysis showed that the variables 'Overall Performance,' 'Ability to Hit,' and 'Ability to Score' account for the most variability in the data.

Here are the analysis results for Group C.

```
Variable        PC1     PC2     PC3     PC4     PC5     PC6     PC7
Games         0.529   0.842  -0.107  -0.022   0.007   0.002  -0.000
Hits          0.741  -0.412   0.468  -0.250   0.013  -0.001   0.001
OPS           0.001  -0.003   0.000   0.005   0.003   0.975   0.223
AVG           0.000  -0.001   0.001  -0.000   0.000   0.223  -0.975
Runs          0.409  -0.319  -0.607   0.601  -0.019  -0.004  -0.001
SB            0.063  -0.139  -0.623  -0.740   0.200   0.002   0.000
CS            0.019  -0.022  -0.110  -0.166  -0.980   0.004   0.001

Proportion    0.925   0.043   0.018   0.012   0.001   0.000   0.000
Cumulative    0.925   0.969   0.987   0.999   1.000   1.000   1.000
```

These results show the first principal component accounts for 92.5% of the variability, the second accounts for 4.3%, and the third accounts for 1.8%. Together the first three principal components for Group C account for 98.7% of the variability in the data.

Looking at the principal components individually, we see that the first emphasizes Games, Hits, and Runs. Since it emphasizes Runs Hits, and Games, it seems that the variable is measuring offensive production over the season. So this variable is called 'Offensive Productivity.' The second principal component has a contrast of Games versus Hits and Runs. There is a contrast between games played and offensive ability, so this variable might be a measure of how efficient one's offensive is over the season. This variable can be 'Offensive Efficiency.' The third principal component contrasts Hits with Runs and SB (stolen bases). The variables Runs and SB are conducive to scoring and they are contrasted with games played. It seems this variable measures one's ability to score runs without getting hits. This variable will be 'Scoring without Hitting.'

The principal component analysis for Group C showed us that the majority of the variability in the data can be explained by the variables 'Offensive Production,' 'Offensive Efficiency,' and 'Scoring without Hitting.'

VI. Performance Prediction

In this section my intent is to find a model for predicting future performance. To be more specific, find a model that predicts next season's performance for a certain offensive variable. To do this, I used linear regression with two different data sets. The response variable comes from the normal 2008 season data but all the predictor variables will come from 2007 season data. This 2007 data set contains 331 observations all of which played in both the 2007 and 2008 MLB seasons. In this case, RBI will be the response variable.

For each variable group (A, B, and C) I will show the best model using the regular variables as well as the best model using the first three principal components as new variables. The principal components used in this section come from the 2007 data set rather than the 2008 data set (Appendix: Principal07 pg 23). Despite this, the principal components from both data sets are very similar and result in the same 'new variables' for the 2007 data set.

Below is regression analysis that yielded the best model for Group A.

```
Regression Analysis: 2008RBI versus Games, 2B, HR, OBP

The regression equation is
2008RBI = 5.4 - 0.130 Games + 0.531 2B + 1.70 HR + 85.0 OBP



Predictor       Coef  SE Coef      T      P
Constant        5.43    12.46   0.44  0.663
Games       -0.13009  0.06312  -2.06  0.040
2B            0.5313   0.1958   2.71  0.007
HR            1.6991   0.1672  10.16  0.000
OBP            84.99    35.88   2.37  0.018

S = 21.9491   R-Sq = 46.3%   R-Sq(adj) = 45.6%
```

Note: The main regression assumptions are satisfied for this model (Appendix: RegressionA1 pg 24).

The regression analysis for Group A found the variables Games, 2B, HR, and OBP to be significant. Interaction terms were also investigated as predictors but none were found to be significant. The adjusted R^2 value here is 45.6%, which is a low value. Generally values under 50%

are considered poor. Unfortunately this suggests that the model is not very strong at predicting 2008 RBI.

Here is the regression analysis using the principal components as variables.

Regression Analysis: 2008RBI versus Walk Production, Walk Efficiency, Power

```
The regression equation is
2008RBI = 35.9 + 0.348 Walk Production + 0.383 Walk Efficiency + 1.69 Power



Predictor             Coef  SE Coef      T      P
Constant            35.890    4.662   7.70  0.000
Walk Production    0.34801  0.02961  11.75  0.000
Walk Efficiency    0.38265  0.08029   4.77  0.000
Power               1.6937   0.1676  10.11  0.000


S = 22.2553   R-Sq = 44.6%   R-Sq(adj) = 44.1%
```

The main regression assumptions are satisfied for this model (Appendix: RegressionA2 pg 25).

All of the firs three principal components were kept in the model. The variables are Walk Production, Walk Efficiency, and Power. But, this model also has a low adjusted R^2 value of 44.1% and thus would not be very reliable in predicting 200 RBI. Although neither of these models were great, they are the best models using the regular variables and principal components to predict the 2008 RBIs for Group A.

Next is the result of the regression analysis for Group B.

Regression Analysis: 2008RBI versus Games, RBI, SO, XtraBH

```
The regression equation is
2008RBI = 36.0 - 0.333 Games + 0.543 RBI + 0.107 SO + 0.471 XtraBH

Predictor       Coef  SE Coef      T      P
Constant      36.023    4.677   7.70  0.000
Games       -0.33278  0.06319  -5.27  0.000
RBI           0.5427   0.1047   5.18  0.000
SO           0.10724  0.05319   2.02  0.045
XtraBH        0.4714   0.1712   2.75  0.006

S = 21.9766   R-Sq = 46.1%   R-Sq(adj) = 45.5%
```

The main regression assumptions are satisfied for this model (Appendix: RegressionB1 pg 25).

For Group B, the regression analysis found the variables Games, RBI, SO, and XtraBH to be significant. Once again the adjusted R^2 value is low at 45.5%. It seems that it is also the case with Group B that the regression model is not strong at predicting the next season's RBIs.

Below is the regression analysis performed using the principal components for Group B.

Regression Analysis: 2008RBI versus Overall Performance, Ability to Hit, Ability to Score

```
The regression equation is
2008RBI = 27.4 + 0.259 Overall Performance + 0.223 Ability to Hit
          - 0.632 Ability to Score


Predictor                Coef  SE Coef      T      P
Constant               27.426    3.955   6.93  0.000
Overall Performance   0.25946  0.01960  13.24  0.000
Ability to Hit        0.22294  0.04715   4.73  0.000
Ability to Score     -0.63211  0.07778  -8.13  0.000


S = 22.2427   R-Sq = 44.6%   R-Sq(adj) = 44.1%
```

The main regression assumptions are satisfied for this model (Appendix: RegressionB2 pg 26).

The three variables of Overall Performance, Ability to Hit, and Ability to Score were all kept in the model. The adjusted R^2 value is 44.1%, another low value. Like with Group A, the best models using the regular variables and principal components from Group B are not very strong at predicting 2008 RBI.

The regression models for Group C also had low R^2 values and wouldn't be considered as strong predictors of 2008 RBI. The analysis can be seen in the appendix (Appendix: RegressionC pg 27).

VII. Conclusion

In analyzing team success I found that the variables of Runs, At-bats, Stolen Bases, and Batting Average are most associated a team's winning percentage. For Groups B and C, there was a significant league difference and we looked at the discriminant function to further see which variables contributed most to the difference. Though, in both cases the discriminant function was shown to be only slightly reliable. In trying to predict performance for next season, I found models to predict the next season's RBI (Runs Batted In) using both the normal variables and new variables created with the principal components. Sadly all the models were weak at predicting RBI and should not be considered reliable.

Unfortunately, the lack of significant and interesting results for Groups A, B, and C mirror all the results I saw throughout working on my project. Though this is true, it is no reason to think that this project would not be able to go any further. There are a number of areas that can still be investigated. For instance, time can be looked at more closely. Instead of including that statistics for just one or two seasons, the last ten, fifty, or even one hundred seasons can be looked at to create better models and make more comparisons. Also, I only looked at the offensive statistics for my project without considering pitching. Investigating the pitching statistics can be a whole project in its own and we could combine the pitching and offensive statistics to more accurately understand team success and league differences.

These are just some of the ideas I have for possibly continuing this project, though I am sure there are many other ideas that I and other people can come up with. There are a lot of directions that this project can go in. I am glad that I was able to work with this data and I plan to continue my investigation as time goes on.

VII. Appendix

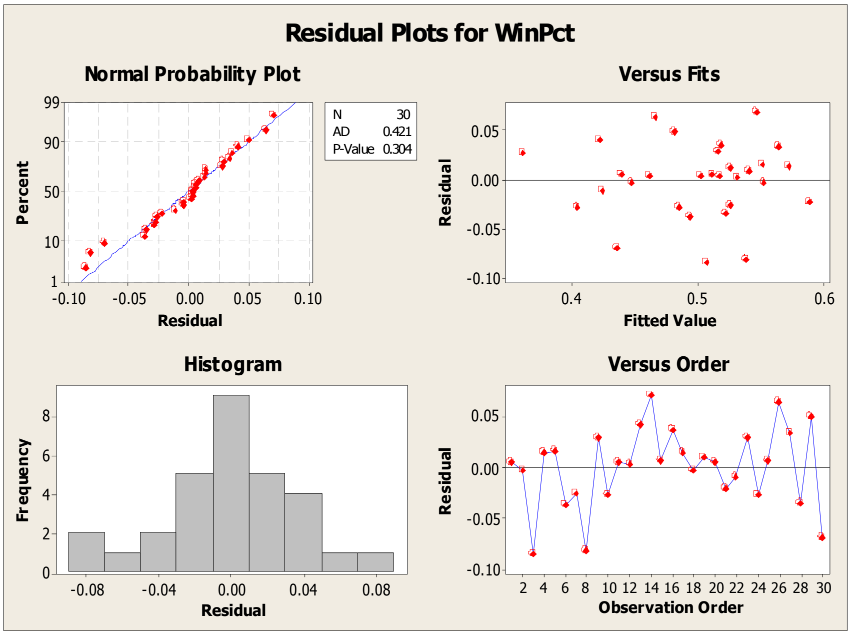
Regression1A: Calculation of Standardized Coefficients and 95% Confidence Intervals

The calculation for the standardized coefficient is:

$std\,coef_x = \frac{sd_x}{sd_y} \square coef_x \square$ where x denotes the independent variable and y denotes the dependent variable.

AB:	std coef = (69.9486/0.0667075)(-0.0004013) = -0.420798
Runs:	std coef = (68.4343/0.0667075)(0.0007010) = 0.719146
SB:	std coef = (28.6081/0.0667075)(0.0009206) = 0.394807

Regression1B: Regression Diagnostics for First Model



Based on the nature of the data, independence of the data is not a problem, thus the independence assumption is always satisfied.

The Normal Probability Plot gives us information on the normality assumption. More importantly, we can look at the Anderson-Darling test results next to the plot. In this test, the null hypothesis is that the data is normally distributed, thus we want to fail to reject the null and thus we want to see a high p-value. The p-value for the test is 0.304, thus we fail to reject the null hypothesis. This means that there is not enough evidence to suggest that the normality assumption is violated.

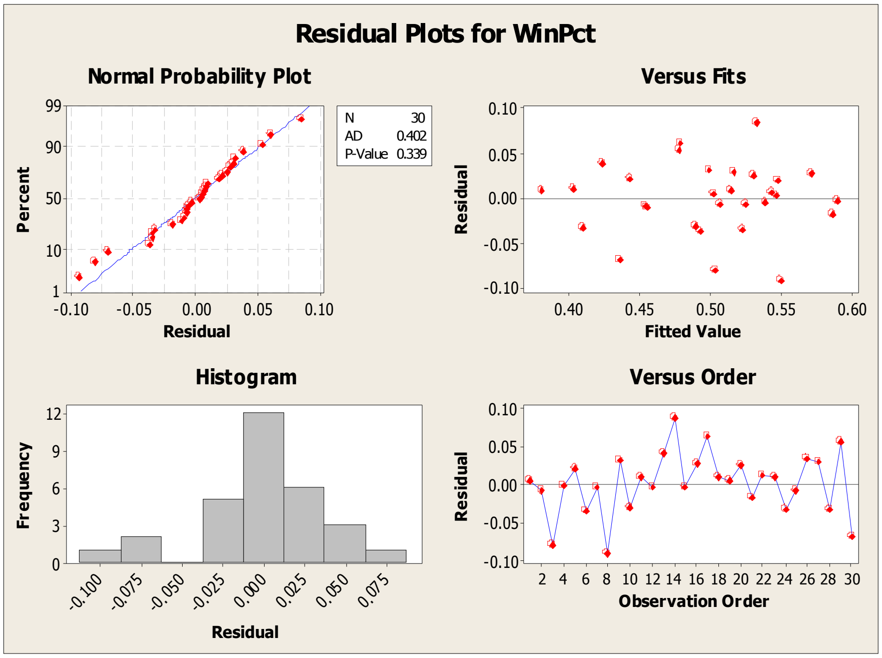
We look at the Residual Versus fits plot to get information on the assumption of equal variance. If there are any specific patterns in the data then we need to further investigate. The plot appears to be randomly scattered and so we can assume equal variance.

Regression2A: Calculation of Standardized Coefficients and 95% Confidence Intervals

Standardized coefficients:

AB: std coef = (69.9486/0.0667075)(-0.0005649) = -0.592347
Runs: std coef = (68.4343/0.0667075)(0.0005931) = 0.608453
SB: std coef = (28.6081/0.0667075)(0.0008341) = 0.357711
AVG: std coef = (0.0104793/0.0667075)(1.91100) = 0.300205

Regression2B: Regression Diagnostics for Second Model



The p-value for the Anderson-Darling test is 0.339, so we fail to reject the null hypothesis. Thus there is not enough evidence to suggest the data are not normally distributed.
The Residual Versus fits Plot doesn't have any patterns or trends, so we can assume equal variance.

Discriminant: R Software Functions for Discriminant Analysis

Function for obtaining discriminant functions:

```
#Obtain Discriminant Functions
#Note: 'Y' denotes data matrix and 'group' is groupong variable
discrim <- function(Y, group){
        Y <- data.matrix(Y)
        group <- as.factor(group)
        m1 <- manova(Y ~ group)
        nu.h <- summary(m1)$stats[1]
        p <- ncol(Y)
        SS <- summary(m1)$SS
        E.inv.H <- solve(SS$Residuals) %*% SS$group
        eig <- eigen(E.inv.H)
        s <- min(nu.h, p)
        lambda <- Re(eig$values[1:s])
        a <- Re(eig$vectors[,1:s])
        a.star <- (sqrt(diag(SS$Residuals)) * a)
        return(list("a"=a, "a.stand"=a.star))
 }
```

Function for partial f test:

```
partial.F <- function(Y, group){
        Y <- data.matrix(Y)
        group <- as.factor(group)
        p <- ncol(Y)
        m1 <- manova(Y ~ group)
        nu.e <- m1$df
        nu.h <- m1$rank-1
        Lambda.p <- summary(m1,test="Wilks")$stats[3]
        Lambda.p1 <- numeric(p)
        for(i in 1:p){
                dat <- data.matrix(Y[,-i])
                m2 <- manova(dat ~ group)
                Lambda.p1[i] <- summary(m2,test="Wilks")$stats[3]
        }
        Lambda <- Lambda.p / Lambda.p1
        F.stat <- ((1 - Lambda) / Lambda) * ((nu.e - p + 1)/nu.h)
        p.val <- 1 - pf(F.stat, nu.h, nu.e - p + 1)
        out <- cbind(Lambda, F.stat, p.value = p.val)
        dimnames(out)[[1]] <- dimnames(Y)[[2]]
        ord <- rev(order(out[,2]))
        return(out[ord,])
 }
```

Function for Rates and Confusion Matrix:

```
rates <- function(data,group,method="l") {
        library(MASS)
        data <- as.matrix(data)
        group <- as.matrix(group)
        da.obj <- lda(data,group)
        if (method=="q") {
        da.obj <- qda(data,group)
        method <- "QDA"
        }
    tab <- table(original=group,predicted=predict(da.obj)$class)
        if (method=="l") method <- "LDA"
        cor.rate <- sum(predict(da.obj)$class==group)/nrow(data)
        er.rate <- 1-cor.rate
        return(list("Correct Class Rate"=cor.rate,"Error Rate"=er.rate,
           "Method"=method,"Confusion Matrix"=tab))
}
```

Principal07: Principal Component Analysis for Groups A, B, and C for 2007 data

Principal Component Analysis for Group A

Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Games	0.817	-0.514	-0.202	0.164	-0.027	0.001	-0.000
2B	0.238	-0.094	0.581	-0.772	-0.036	-0.003	0.000
3B	0.019	-0.028	-0.041	-0.068	0.996	-0.004	0.001
HR	0.173	0.166	0.759	0.601	0.074	-0.005	0.003
Walks	0.496	0.836	-0.210	-0.108	-0.002	-0.001	-0.001
OBP	0.000	0.001	0.001	-0.001	0.001	0.530	0.848
SLG	0.001	0.001	0.007	0.001	0.004	0.848	-0.530
Proportion	0.840	0.114	0.026	0.016	0.003	0.000	0.000
Cumulative	0.840	0.955	0.981	0.997	1.000	1.000	1.000

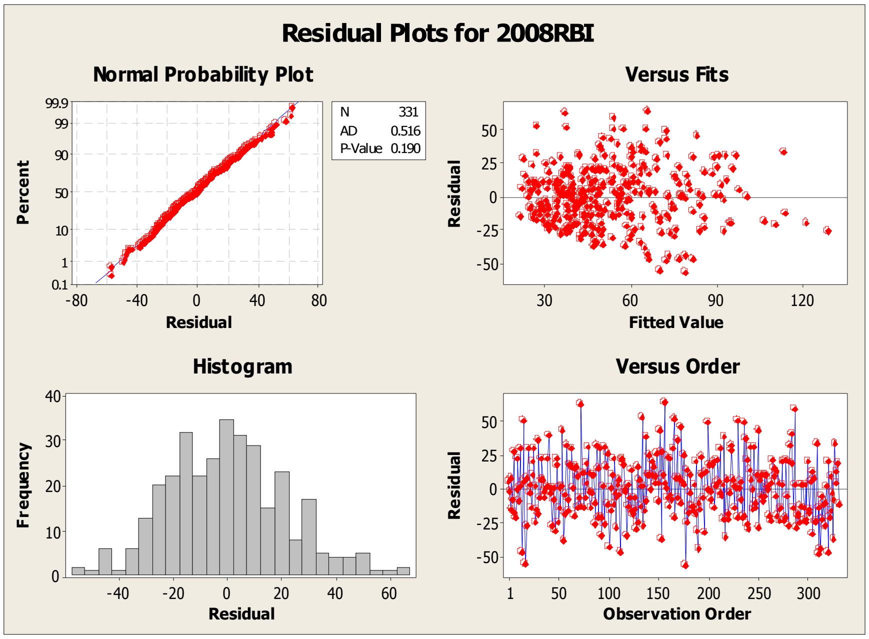
Principal Component Analysis for Group B

Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Games	0.526	-0.203	0.291	-0.770	-0.065	0.002
RBI	0.445	0.081	-0.767	0.031	-0.454	-0.001
SO	0.480	0.721	0.401	0.294	-0.060	0.000
OPS	0.001	0.000	-0.003	0.002	0.003	1.000
1B	0.453	-0.654	0.218	0.565	-0.003	-0.000
XtraBH	0.300	0.073	-0.344	-0.019	0.887	-0.004
Proportion	0.781	0.135	0.050	0.027	0.008	0.000
Cumulative	0.781	0.916	0.965	0.992	1.000	1.000

Principal Component Analysis for Group C:

```
Variable        PC1     PC2     PC3     PC4     PC5     PC6     PC7
Games         0.497   0.801   0.334   0.026  -0.004   0.002   0.001
Runs          0.420  -0.419   0.434  -0.677   0.025  -0.004   0.001
Hits          0.756  -0.260  -0.523   0.295  -0.016  -0.001  -0.001
SB            0.070  -0.334   0.644   0.656  -0.197   0.003  -0.000
CS            0.018  -0.057   0.112   0.154   0.980   0.004  -0.001
AVG           0.000  -0.001  -0.001   0.000  -0.000   0.224   0.974
OPS           0.001  -0.002  -0.002  -0.005  -0.003   0.974  -0.224


Proportion    0.918   0.043   0.025   0.014   0.001   0.000   0.000
Cumulative    0.918   0.961   0.985   0.999   1.000   1.000   1.000
```

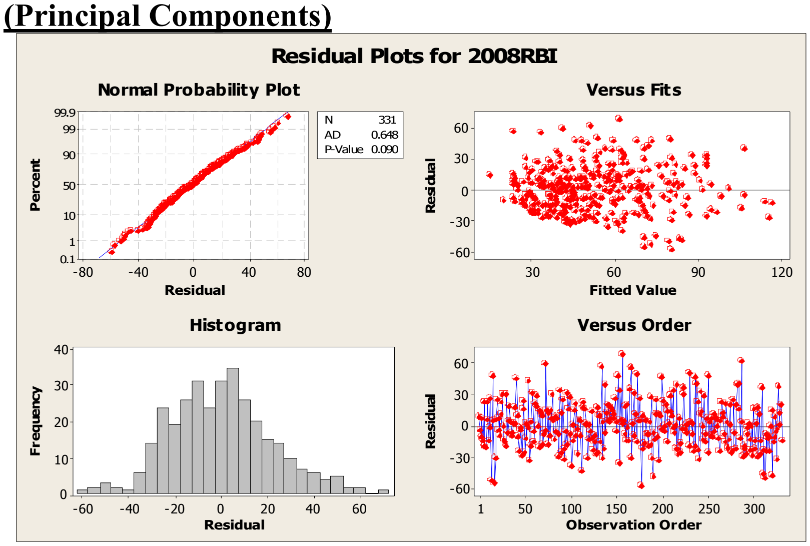
RegressionA1: Regression Diagnostics and Confidence Interval Calculations for Group A



The Anderson-Daring p-value is 0.190, so we fail to reject the null, and thus can not conclude that the normality assumption is violated.

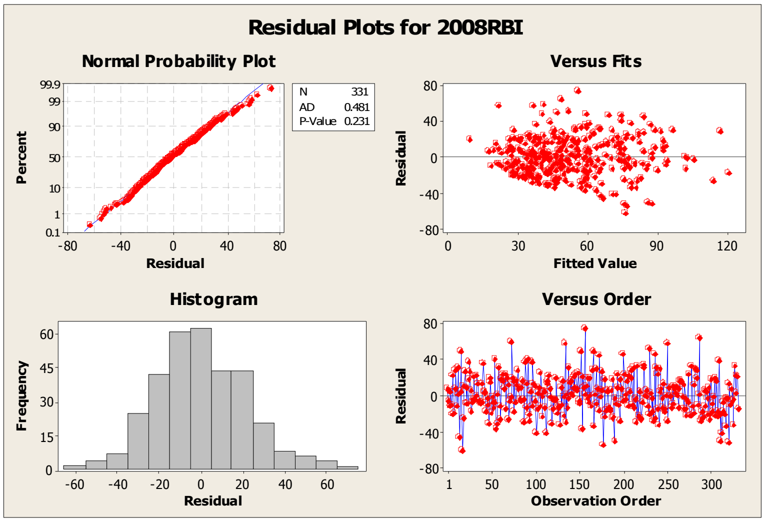
The Residual vs Fits plot looks fairly scattered and random and so there is no evidence to suggest that the equal variance assumption is violated.

RegressionA2: Regression Diagnostics and Confidence Interval Calculations for Group A (Principal Components)



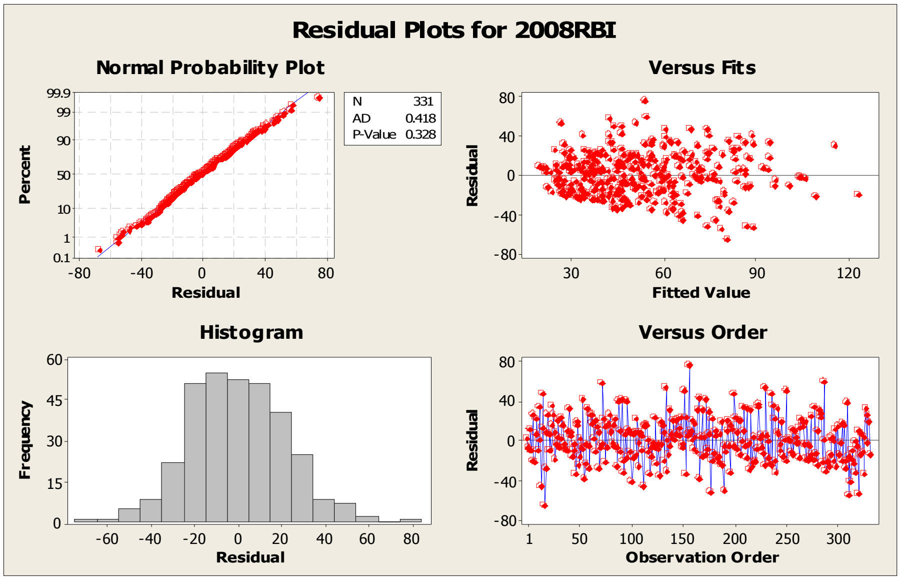
Here the Anderson-Darling p-value is 0.09. This p-value is pretty low and might suggest a possible normality violation. But at the 5% level, we still fail to reject the null hypothesis meaning there is not enough evidence to conclude a violation of normality for this model. The Residual vs Fits plot does not have any major patterns and looks scattered, so there is not evidence to suggest that the equal variance assumption is violated.

RegressionB1: Regression Diagnostics and Confidence Interval Calculations for Group B



The Anderson-Darling test yields a p-value of 0.231. For this model we fail to reject the null hypothesis that the data are normal. The Residual vs Fits plot looks fairly random with no patterns, so there is no evidence to suggest the equal variance assumption is violated.

RegressionB2: Regression Diagnostics and Confidence Interval Calculations for Group B (Principal Components)



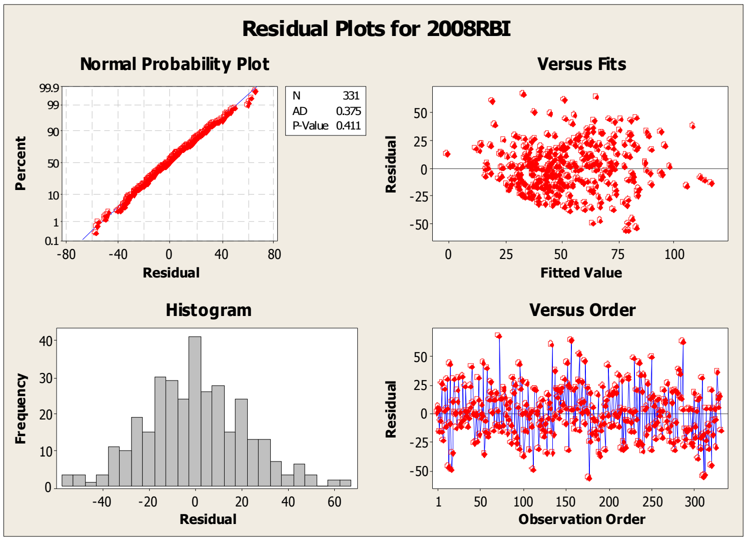
For this model, the Anderson-Darling p-value is 0.328, which means we fail to reject the null hypothesis. Thus we can assume normality for this model. The Residual vs Fits plot looks without major pattern so we can assume equal variance for this model.

RegressionC: Regression Analysis for Group C

Regression results and diagnostics for Group C:

Regression Analysis: 2008RBI versus Games, Hits, SB, OPS*Games, AVG*Hits

```
The regression equation is
2008RBI = 37.3 - 1.37 Games + 1.37 Hits - 0.335 SB + 1.45 OPS*Games
              -  3.42 AVG*Hits
              -


Predictor      Coef  SE Coef      T      P
Constant     37.267    5.083   7.33  0.000
Games       -1.3739   0.1573  -8.73  0.000
Hits         1.3749   0.2298   5.98  0.000
SB          -0.3346   0.1259  -2.66  0.008
OPS*Games    1.4544   0.1440  10.10  0.000
AVG*Hits    -3.4219   0.5931  -5.77  0.000

S = 22.1887   R-Sq = 45.2%   R-Sq(adj) = 44.4%
```



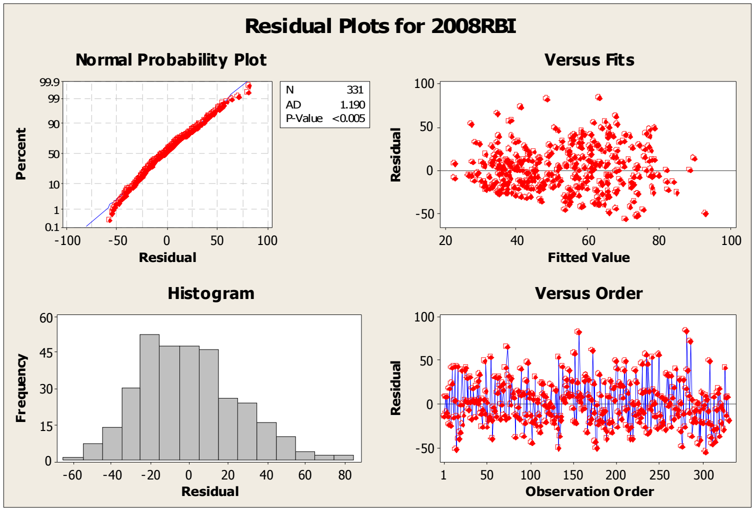
The Anderson-Darling p-value is 0.411 so we fail to reject the null hypothesis and there is not enough evidence to conclude that the data are not normal. The Residual vs Fits plot seems to have a slight curve trend, but still looks scattered enough to assume equal variance.

Regression Analysis for Group C (Principal Components)

Regression Analysis: 2008RBI versus Offensive Production, Offensive Efficiency, Score w/o Hitting

```
The regression equation is
2008RBI = 27.5 + 0.219 Offensive Production - 0.182 Offensive Efficiency
          - 0.341 Score w/o Hitting


Predictor                 Coef  SE Coef      T      P
Constant                27.513    5.618   4.90  0.000
Offensive Production   0.21930  0.02158  10.16  0.000
Offensive Efficiency  -0.18246  0.09963  -1.83  0.068
Score w/o Hitting      -0.3413   0.1319  -2.59  0.010


S = 25.7633   R-Sq = 25.7%   R-Sq(adj) = 25.0%
```

Before this model is considered valid, let’s look at the diagnostics:



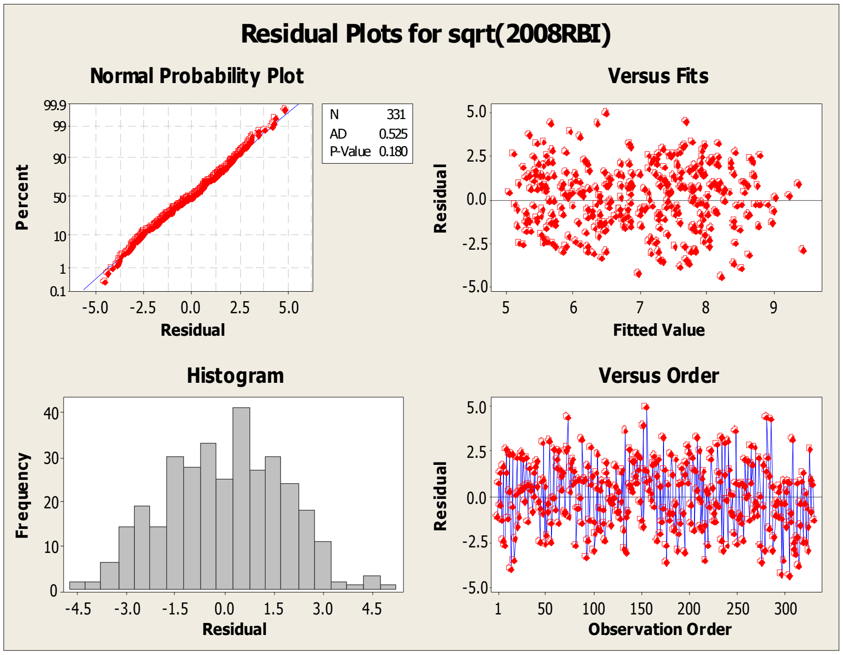
The plot shows that the Anderson-Darling p-value for this model is less than 0.005, and thus we reject the null hypothesis that the data are normal. Therefore we can not assume normality for this model. To correct this, I applied a transformation on the response variable. The transformation used was the square root.

Observe the results of the regression analysis on the transformed variable on the next page. This is the best model using the principal components as predictors of the square root of 2008 RBI.

Regression Analysis: sqrt(2008RBI) versus Offensive Production, Score w/o Hitting

```
The regression equation is
sqrt(2008RBI) = 4.67 + 0.0156 Offensive Production - 0.0244 Score w/o Hitting



Predictor                 Coef   SE Coef      T      P
Constant                4.6653    0.2938  15.88  0.000
Offensive Production  0.015611  0.001525  10.24  0.000
Score w/o Hitting    -0.024371  0.009318  -2.62  0.009


S = 1.82034   R-Sq = 25.4%   R-Sq(adj) = 24.9%
```

Here are the diagnostics for the model with the transformed response.



With the new transformation, the Anderson-Darling p-value is now 0.180, and this time we fail to reject the null hypothesis. It would appear that the transformation fixed the normality violation and so we assume normality. The residual vs Fits plot looks randomly scattered giving now evidence of an equal variance violation.