Multiple Regression Modeling of
Batting Statistics and Team Wins for Chicago Cubs

Samantha Morello

Introduction

The Chicago Cubs are an American professional baseball team established in 1876. Being one of the oldest franchises in Major League Baseball (MLB), the franchise has an abundance of historical data available, making it possible to conduct a detailed analysis of the team's batting performance over the last 25 years. In order to maximize the team's efficiency, it is crucial to identify the key batting factors that contribute to wins for the Chicago Cubs through the development of predictive regression models. This study aims to identify and develop a robust multiple regression model that can accurately predict the number of wins based on various batting statistics for the Chicago Cubs.

Abstract Executive Summary

The goal of this analysis is to determine the best two-variable model and the best three-variable model for predicting team wins for the Chicago Cubs. To do this, it is crucial to first identify the key batting metrics that most significantly contribute to wins for the Chicago Cubs. For this analysis, several batting variables will be correlated alongside wins in order to determine the best two and three-variable model. The various

batting factors to be analyzed include runs, hits, doubles, and more. In total, this analysis will consider eleven different batting statistics that may impact team wins.

The full list of variables analyzed are below:

- R (Runs)
- H (Hits)
- DB (Doubles)
- TP (Triples)
- HR (Home Runs)
- RBI (Runs Batted In)
- SB (Stolen Bases)
- BB (Walks)
- SO (Strikeouts)
- BA (Batting Average)
- BatAge (Age of Batters)

These batting statistics are used in order to identify the most significant association with wins. To conduct the analysis, Multiple Regression will be utilized with the goal of understanding influencing factors that contribute to wins for the Chicago Cubs.

The data collected focuses on batting statistics for the Chicago Cubs from the years of 1999 to 2024, excluding the year 2020. This dataset includes 25 observations, one for each year, and 27 batting related variables. From the 27 batting related variables, the analysis will only utilize four, RBI (Runs Batted In) , R (Runs) , BB (Walks), DB (Doubles). These are the four variables that were found to be most strongly associated with team wins and will be further correlated for our multiple regression models. Determining the best two-variable and the best three-variable multiple regression model involves a multi-step process of testing all combinations of variables with wins, the dependent variable. This requires six two-variable models and two three-variable models. Following the creation of these models, their output is analyzed

and compared based on several specific statistical criteria and the best model is revealed. To further enhance the analysis, additional models incorporating interaction terms and second-order terms are generated based on the best initial model. The purpose of this is to assess whether these additional variables have an impact on the model's overall efficiency and predictive accuracy, or if the original model is still the best.

After diving into each of the multiple regression, interaction, and second order models, the findings indicate that the three-variable Model (Model 2: Runs Batted In & Doubles & Walks) is the most significant model in predicting wins for the Chicago Cubs. Additionally our analysis revealed that the interaction and second order model did not add improvements in the model for predicting wins for the Chicago Cubs. In order to better understand these results, further explanation regarding the process of this analysis will be displayed through the exploratory analysis, correlation analysis, methodology and Multiple Regression techniques used.

Research Question

The analysis of this study focuses on the following research question:
What is the best two-variable model and the best three-variable model for predicting team wins for the Chicago Cubs based on the four batting variables that are most associated with team wins? (RBI (Runs Batted In) , R (Runs) , BB (Walks) , DB (Doubles))

Data

The dataset used within this study was retrieved from Baseball Reference, an official website for current and historical baseball players, teams, scores and leader statistics. It contains only data gathered from the years of 1999 to 2024, excluding the year of 2020. Therefore, the dataset focuses on 25 years of batting statistics for the Chicago Cubs. As stated above, this dataset includes 25 observations, one observation for each year analyzed, and 27 batting related variables. For the purpose of this analysis, four specific batting variables are examined:

RBI (Runs Batted In): The number of runs a player contributes by hitting the ball, allowing teammates on base to score.

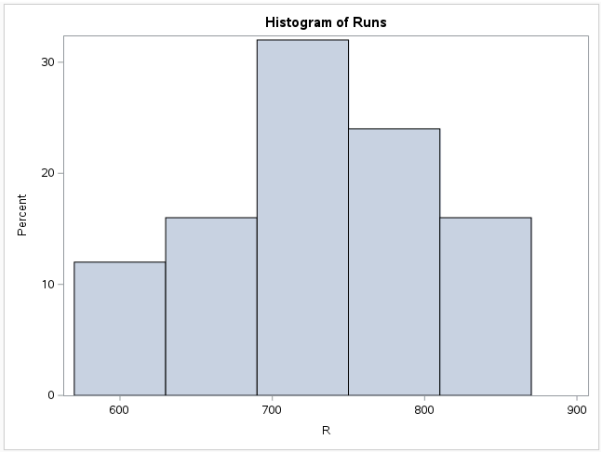
R (Runs): The total number of times players successfully reach home plate.

BB (Walks): The number of times a batter reaches first base due to pitches being called as balls.

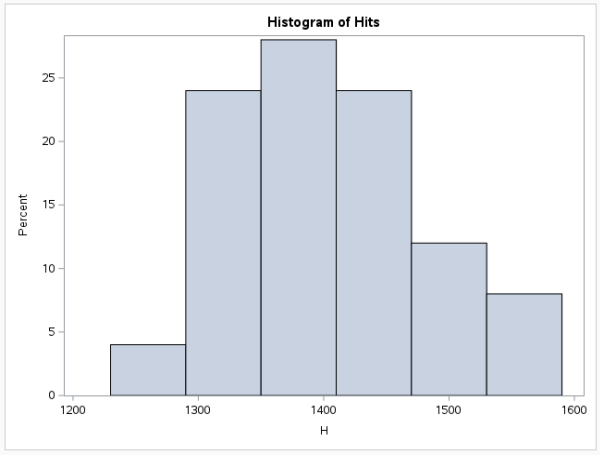
DB (Doubles): Hits where the batter safely reaches second base.

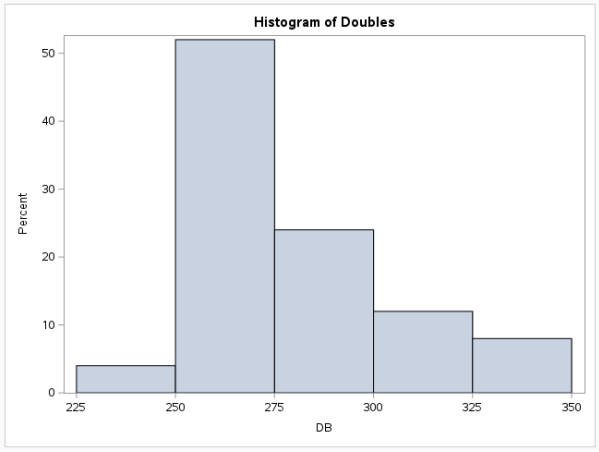
These variables were chosen because they were found to represent the four variables most strongly associated with team wins for the Chicago Cubs. The data gathered from these four batting variables in correlation to wins will help to identify which multiple regression model has the most significant impact on team wins for the Chicago Cubs.

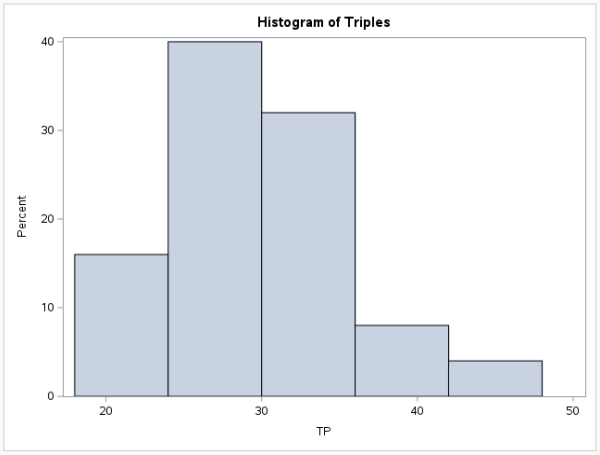
Exploratory Data Analysis



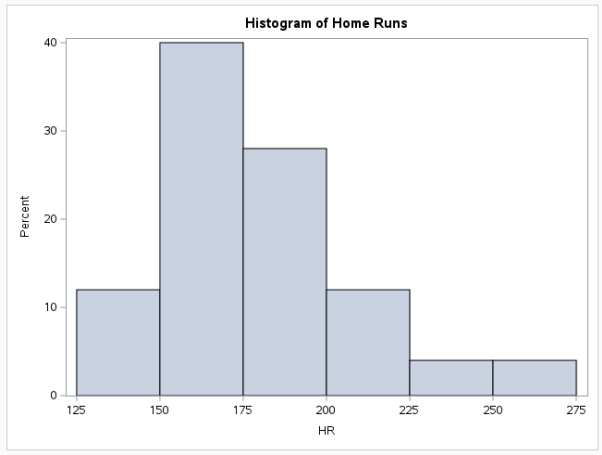
The very first step of the analysis involved conducting Exploratory Data Analysis (EDA), which is used to analyze and investigate datasets and summarize main characteristics. To better understand the distribution of the variables, each component is visualized and displayed with a histogram. The eleven important batting statistic variables are to be displayed in histogram form (R (Runs), H (Hits), DB (Doubles), TP (Triples), HR (Home Runs), RBI (Runs Batted In), SB (Stolen Bases), BB (Walks), SO (Strikeouts), BA (Batting Average), BatAge (Age of Batters)). Each histogram displays a unique distribution of data, although several appear to be slightly right skewed. To begin, the histogram for runs displays most observations falling between just under 700-800 runs. As for the histogram for hits, its distribution seems to be symmetrical and it is clear that the majority of hits lie around 1300-1450. Around 1500 hits, there appears to be a drop in observations. The

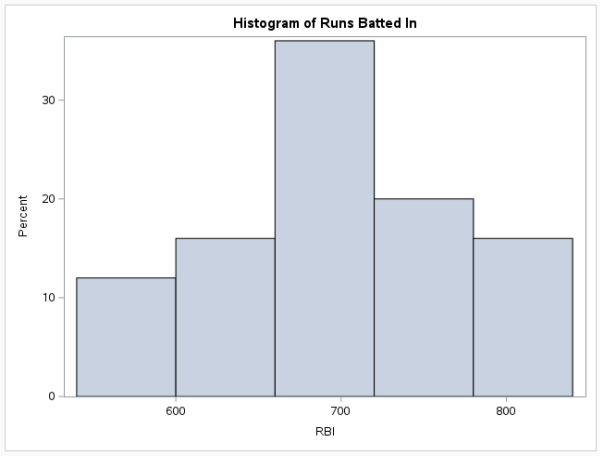


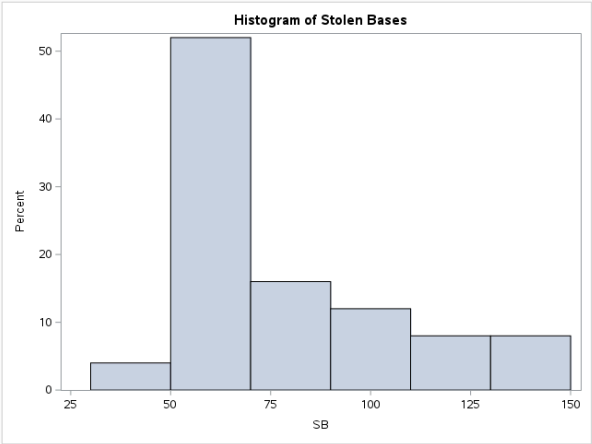




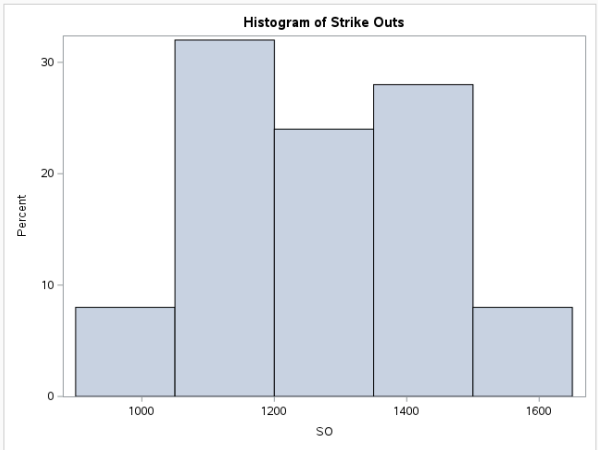
histogram for doubles shows observations concentrated between 250-300 with a right skew in the data, like previously mentioned. Although the peak of observations is between 250-275, indicating a very small range of performance for doubles. The next histogram, for triples, also displays a right skew and most observations are seen to be between 25-35. The graph shows that there are more observations towards the lower end. The histogram for Home Runs displays most observations having between 150-200 home runs. Additionally, it displays a right skew toward lower home run totals. Although observations are very low just before the peak between 125-150. The histogram for runs batted in has a high concentration of observations lying between 650-750. The data seems to be fairly symmetrical and centered without a clear skew, this indicates a normal spread of data. The next histogram, stolen bases, is another example of a right skewed graph. Most of the observations display that fewer than 75 bases were stolen and a very





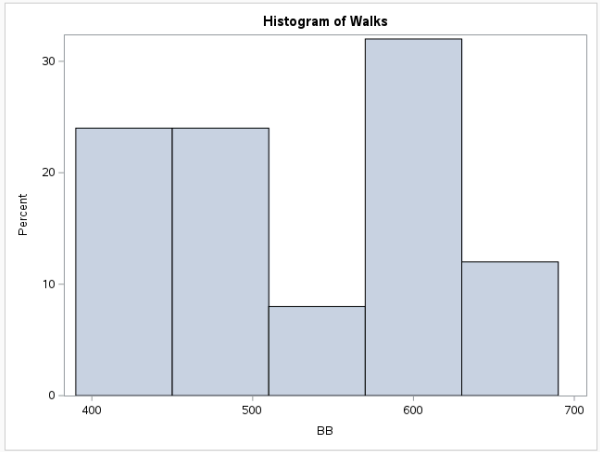


drop off between 500-550. The next histogram, displaying strike outs, is another example of a relatively symmetrical distribution. Again this indicates a normal spread of data. Although it could be said that there might be a very slight right skew, with the

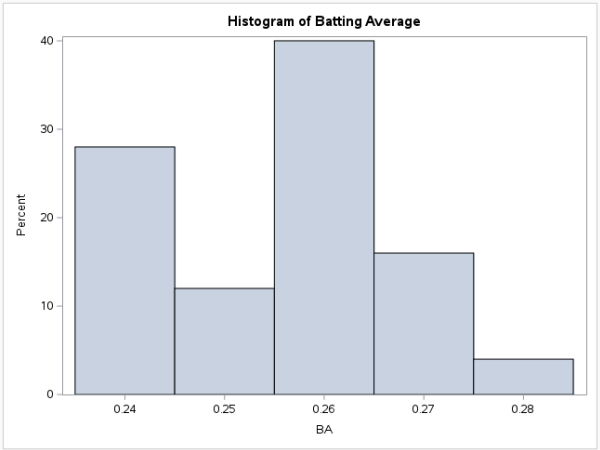


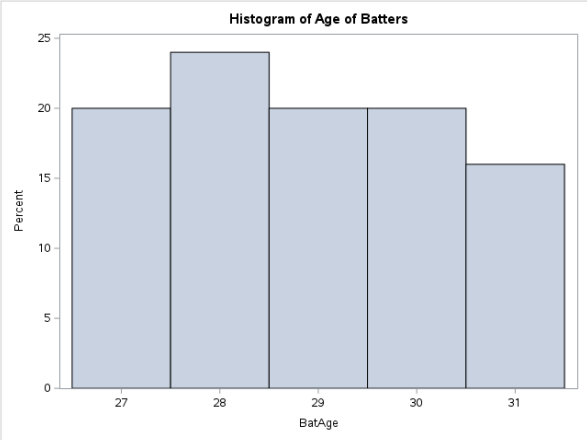
above 0.25 and a bit below 0.27. The graph also displays that there are relatively high batting averages on the

little number of observations above that. The peak appears to be between 50-75, which is the reason for the long right tail. Next up is the histogram displaying walks, in this graph the highest number of observations lies between around 550-650. Although, observations also seem to be a bit high from 400-500, with a

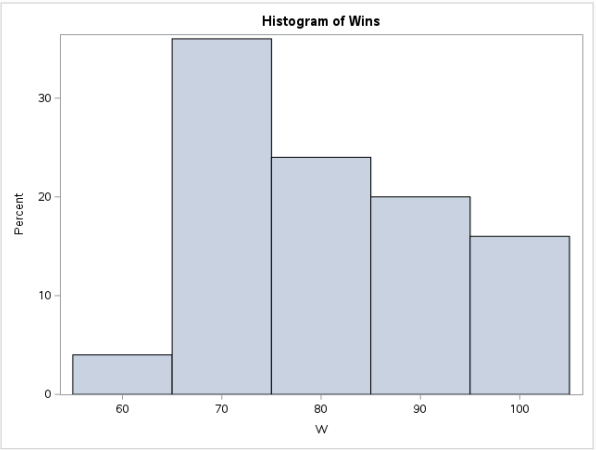


majority of observations lying just above 1000 to around 1500. As for the histogram for batting averages, there is a higher concentration of observations that sit just





lower end, around 0.24, indicating a right skew in the data. The last of our eleven batting statistic variables, displays the age of batters. This histogram appears to be relatively symmetrical, with a spread of ages around 26 up to around 32 years old. There does not appear to be a skew within this graph, just a tiny peak around the year 28. Another important histogram to visualize is the dependent variable of the analysis, wins. In this histogram it is clear to see a right skew, with a peak in observations around 70-80. This indicates that the Chicago Cubs typically have seasons with around 70-80 wins. This can also suggest that high winning seasons are less common for the Chicago Cubs during this time frame.



The visualization of each batting metric variable displayed that many of the variables are right skewed, including doubles, triples, home runs, etc. Additionally, it revealed several other variables displaying a more symmetrical distribution, including hits, strike outs, and age of batters. The exploration and visualization of these key components provide a foundational understanding of the dataset as a whole and aid in the prediction of wins for the Chicago Cubs.

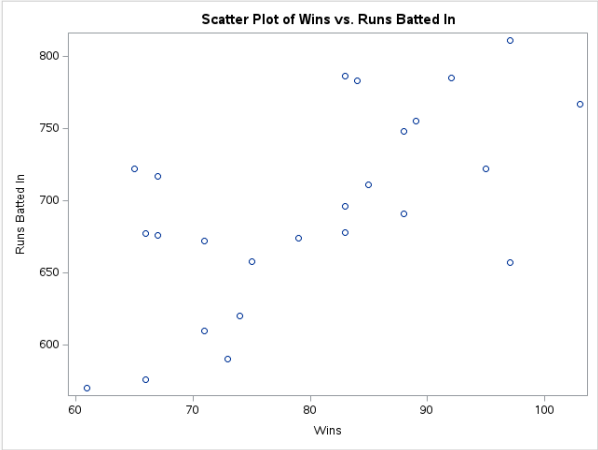
Correlation Analysis

In order to determine the four batting statistic variables that are most strongly associated with team wins, a correlation analysis was implemented. This is done by analyzing the output of a 'proc corr' procedure in SAS, which calculates the correlation coefficients between the specific batting variables (all 11 variables previously mentioned) and wins. The correlation coefficient measures the strength and direction of the linear relationship between two variables. This means that the four variables with the highest correlation coefficients represent the strongest associations. The four variables that were found to have the highest correlation coefficient and therefore have the strongest relationship to wins are:

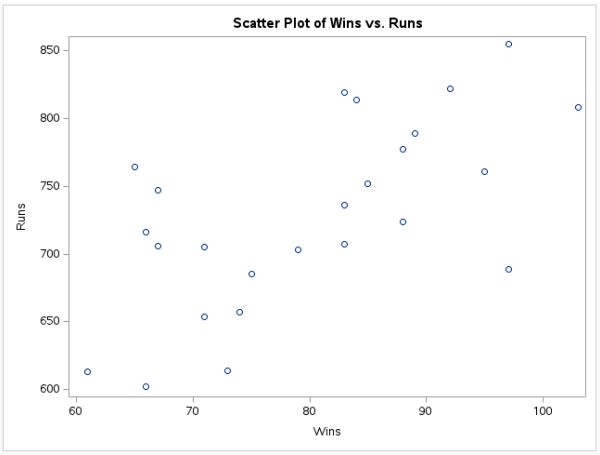
RBI (Runs Batted In): 0.64058 BB (Walks): 0.53792

R (Runs): 0.63919 DB (Doubles): 0.39551

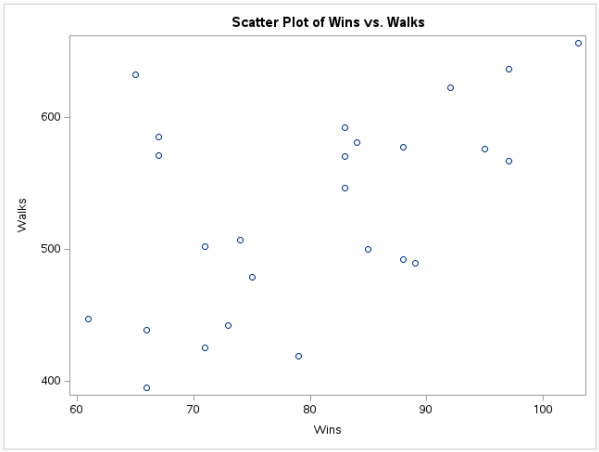
By identifying the four variables with the strongest correlation to wins, we can now proceed to the analysis and implementation of our multiple regression models. But before that, visualizations for each of these variables are displayed to effectively assess these relationships between each of the four selected variables and wins. This step is done to spot any outliers or trends in the data.

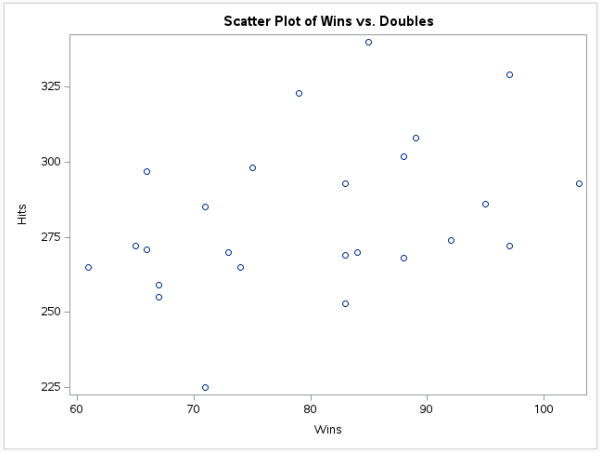


Since the correlation between runs batted in and wins is the strongest, let's start there. This scatterplot somewhat displays a positive direction, although it is



not clear. This plot also somewhat displays a linear trend within the data. The distribution of points indicates a moderate relationship. Lastly, it is hard to tell but may include one questionable outlier. As for the scatterplot of runs and wins, a slight positive direction can also be seen. As well as a slight linear trend. With the inconsistent, but somewhat linear distribution of points, it can be considered to have a weak to moderate trend. Lastly, it does not necessarily appear to have any outliers as all points are a bit wide spread.The remaining two plots have similar qualities and trends displayed. It appears that both plots seem to have no clear trend or pattern. For the scatter plot displaying walks and wins, there is no clear direction or form. The strength of the plot is very weak, with no evident pattern. There may be outliers, but again it is hard to tell how they impact the model. Overall, there does not appear to be a very strong correlation between the two variables. The same interpretation can be said for the scatter plot displaying doubles and wins. These





scatterplots do not display a correlation between walks and doubles and team wins. Although these scatter plots do not display extremely clear trends, they help to provide visuals of the relationships before diving into the more in depth regression analysis.

Multiple Regression

Jumping into the Multiple Regression analysis, using wins as the dependent variable and the four key batting metrics as the independent variables, the relationship between batting performance and team success was conducted. This multi-step process incorporates all combinations of the four variables within several multiple regression models with the goal of identifying which combinations of variables most effectively predict the number of wins for the Chicago Cubs. This requires six two-variable models, two three-variable models and two more complex models.

Each model was analyzed and evaluated based on key statistical criteria within the outputs of each model. These metrics included the Adjusted R-Square, F Test P-Value, Standard Deviation (Root MSE), and Coefficient Variation. It is crucial to consider these metrics when examining multiple regression models because they provide a comprehensive view of the model's overall performance and accuracy.

Why are each of these statistical metrics important? The importance of the adjusted r square lies in its ability to display how well the independent variables explain the variability in the dependent variable. A higher adjusted r square indicates a better fit, therefore the model that displays the highest adjusted r square has potential to be the best. Next, it is important to look at the F Test P-Value because it reveals the overall significance of the model. Having p-values less than 0.05, or alpha, indicates that the

variables are statistically significant. The standard deviation, or root mse, indicates the predictive accuracy of a model. Having a lower root mse means that the model's predictions are closer to the actual values, which indicates better predictive accuracy. Lastly it is important to look at the coefficient of variation because it provides insights to how much variability there is relative to the mean of the dependent variable. Through this analysis, all of these key metrics are considered and used in determining the best overall multiple regression model that effectively predicts the number of wins for the Chicago Cubs.

Results

Beginning with the six two-variable multiple regression models, they were generated, pairing each combination of the four variables, with the objective being to examine how each pair of batting metrics correlated with wins. Each model was analyzed based on the specific key metrics stated above. In order to better visualize each model and its key statistics, the following table was generated.

2-variable Model Key Statistic Table

Key metrics from all six models considered when determining the best two-variable model.

First Order Two Variable Models	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model 1: *Runs Batted In & Runs* Wins = 2.02743 + 0.08919(RBI) + 0.02215(R)	*0.3569*	*0.0030*	*9.46317*	*11.81714*
Model 2: *Runs Batted In & Walks* Wins = 3.07676 + 0.09110(RBI) + 0.02600(BB)	*0.3717*	*0.0023*	*9.35360*	*11.68031*
Model 3: *Runs Batted In & Doubles* Wins = -28.61247 + 0.10274(RBI) + 0.13272(DB)	***0.4460***	***0.0006***	***8.78262***	***10.96731***

Model 4: *Runs & Walks* Wins = 1.56177 + 0.08865(R) + 0.02626(BB)	*0.3700*	*0.0024*	*9.36584*	*11.69561*
Model 5: *Runs & Doubles* Wins = -29.71086 + 0.10001(R) + 0.13102(DB)	*0.4416*	*0.0006*	*8.81774*	*11.01116*
Model 6: *Walks & Doubles* Wins = -25.22385 + 0.09022(BB) + 0.20412(DB)	*0.4412*	*0.0006*	*8.82128*	*11.01558*

This table displays the key statistics gathered from each two-variable model, these statistics help to determine the best model. Again each model is evaluated based on the following key statistical criteria: Adjusted R-Square, F Test P-Value, Standard Deviation (Root MSE), and Coefficient Variation. From the table, it can be seen that Model 3: Runs Batted In (RBI) & Doubles (DB) is the best performing two-variable model. Out of all six models, it has the highest r square adjusted of 0.4460. This suggests that runs batted in and doubles were the most effective pair of predictors. Additionally the model presented a p value of 0.0006, which is <0.05, indicating that the model is statistically significant. This model also had the lowest standard deviation, or root mse, of 8.78262. This means that this model's predictions were more accurate on average compared to the other models. Lastly, model 3 appears to have the lowest Coefficient Variation coming in at 10.96731, while all other models appear to be over 11.0. This means the model has less variability relative to the mean than the other models, which suggests more consistency in its predictions. Overall, this model presented the highest adjusted r square, a p value <0.05, the lowest standard deviation, and lowest coefficient of variation, making it the most reliable model. Given these

factors, Model 3: Runs Batted In (RBI) & Doubles (DB) was revealed to be the best performing two-variable model for predicting team wins for the Chicago Cubs.

Building upon the two-variable models, two three-variable models were created by adding a third variable. Similarly, each possible combination of the four variables is implemented into two multiple regression models with the objective being to examine how each of the three batting metrics is correlated with wins. Again, both models were analyzed based on the specific key metrics stated above. In order to better visualize each model and its key statistics, the following table was generated.

3-variable Model Key Statistic Table

Key metrics from both models considered when determining the best three-variable model.

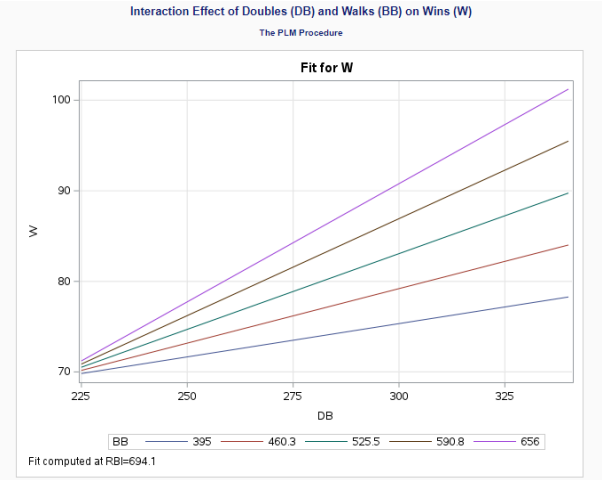
First Order Three Variable Models	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model 1: *Runs Batted In & Doubles & Runs* Wins = -28.08574 + 0.13493(RBI) + 0.13340(DB) - 0.03165(R)	*0.4199*	*0.0022*	*8.98711*	*11.22267*
Model 2: *Runs Batted In & Doubles & Walks* **Wins = -35.26724 + 0.06011(RBI) + 0.16583(DB) + 0.05079(BB)**	***0.4736***	***0.0008***	***8.56147***	***10.69114***

This table displays the key statistics gathered from both three-variable models, these statistics help to determine the best model. Again each model is evaluated based on the following key statistical criteria: Adjusted R-Square, F Test P-Value, Standard Deviation (Root MSE), and Coefficient Variation. From the table, it can be seen that Model 2: Runs Batted In & Doubles & Walks is the best performing three-variable model. Between the two models, it has the highest r square adjusted of 0.4736, even higher than the adjusted r square of the best two-variable model. This suggests that

adding walks as a third variable with runs batted in and doubles increased the model's predictive ability. This also suggests that these three variables are the most effective predictors so far. Additionally, this three-variable model displays a p value of 0.0008, which is <0.05, indicating that the model is statistically significant. This model also had the lowest standard deviation, or root mse, of 8.56147, again even lower than the best two-variable model. This means that this model's predictions were more accurate on average compared to the other models, including the two-variable models. Lastly, model 2 appears to have the lowest coefficient variation coming in at 10.69114, while again the other model is over 11.0. The coefficient variation of this three-variable model is also lower than the best two-variable model. This means the model has less variability relative to the mean than all of the other models, including the two-variable models. This suggests that this three-variable model is more consistent in its predictions. Overall, this model presented the highest adjusted r square, a p value <0.05, the lowest standard deviation, and lowest coefficient of variation, across all previously evaluated models, making it the most reliable model. This model's results suggest that when combined, runs batted in, doubles, and walks are the most impactful predictors of team wins. Given these factors, Model 2: Runs Batted In & Doubles & Walks was revealed to be the best performing three-variable model for predicting team wins for the Chicago Cubs.

To further enhance the analysis, an interaction term model was developed based on the best overall model, the three-variable model. This is done to ensure that all possible relationships were explored in the exploration of finding the best overall model for predicting wins for the Chicago Cubs. The first step of this process was the creation of an interaction plot to determine whether interaction and/or second order terms might

be needed. The plot generated displayed the slopes between variables tested, if the slopes appear to be different, an interaction term could be useful in improving the model, if slopes are parallel, an interaction term is not necessary. Because the plot reveals different slopes for the relationship between variables, it suggests that an interaction term could enhance the model. Based on the output of the plot, interaction terms were added and tested in the regression analysis with the objective being to examine how each of the three batting metrics is correlated with wins.The model was analyzed based on the same specific key metrics stated above. In order to better visualize each model and its key statistics, the following table was generated.



Interaction Model Key Statistic Table

Interaction Model	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model: *Wins = -101.91701 + 0.55681(RBI)* *- 0.08823(DB) - 0.21558(BB)* *- 0.00103(RBI_DB)* *- 0.00039707(RBI_BB)* *+ 0.00191(DB_BB)*	*0.4215*	*0.0112*	*8.97525*	*11.20786*

This table displays the key statistics gathered from the interaction model, these statistics help to determine the best model. Again each model is evaluated based on the following key statistical criteria: Adjusted R-Square, F Test P-Value, Standard Deviation (Root MSE), and Coefficient Variation. From the table, it can be seen that the interaction

model does not outperform the original three-variable model, Model 2: Runs Batted In & Doubles & Walks. The interaction model has an adjusted r square of 0.4215, which is lower than previous results (adjusted r square of previous model: 0.4736). The model has a p value of 0.0112, <0.05, which indicates that model is statistically significant. The standard deviation lies at 8.97525, which does not compare to the results of the previous model (standard deviation of previous model: 8.56147). Lastly, this model appears to have a coefficient variation of 11.20786, which again does not compare (coefficient variation of previous model: 10.69114). This suggests that the three-variable model is more consistent in its predictions without interaction terms. Overall, this model did not perform as well as the three-variable model with a lower adjusted r square, a higher standard deviation, and higher coefficient of variation. This model's results suggest that adding more terms does not always result in a better model. Despite including interaction terms, the three-variable model, Model 2: Runs Batted In & Doubles & Walks, remains the best performing three-variable model for predicting team wins for the Chicago Cubs.

In addition to the interaction terms, a second-order term model was also generated based on the best overall model, the three-variable model, to explore the possibility of non-linear relationships between batting metrics and wins. Second order terms were added and tested in the regression analysis with the objective being to examine how each of the three batting metrics is correlated with wins.The model was analyzed based on the same specific key metrics stated above. In order to better visualize each model and its key statistics, the following table was generated.

Second Order Model Key Statistic Table

Second Order Model	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model 2: *Runs Batted In & Doubles & Walks* Wins = -264.86198 + 0.08995(RBI) + 1.01794(DB) + 0.42211(BB) - 0.00001633(RBI_2) - (0.00147)DB_2 - 0.00035868(BB_2)	*0.4225*	*0.0110*	*8.96708*	*11.19766*

This table displays the key statistics gathered from the second order model, these statistics help to determine the best model. Again each model is evaluated based on the following key statistical criteria: Adjusted R-Square, F Test P-Value, Standard Deviation (Root MSE), and Coefficient Variation. From the table, it can be seen that similar to the interaction model, the second order model does not outperform the original three-variable model, Model 2: Runs Batted In & Doubles & Walks. The second order model has an adjusted r square of 0.4225, which again is lower than previous results (adjusted r square of previous model: 0.4736). The second order model has a p value of 0.0110, <0.05, which still indicates that model is statistically significant. The standard deviation of this model is 8.96708, which again does not compare to the results of the previous model (standard deviation of previous model: 8.56147). Lastly, this model has a coefficient variation of 11.19766, which again does not compare (coefficient variation of previous model: 10.69114). Just like the interaction model, this suggests that the three-variable model is more consistent in its predictions without second order terms. Overall, this model did not perform as well as the three-variable model with a lower adjusted r square, a higher standard deviation, and higher coefficient of variation. The results from this model reiterate that adding more terms does not

always result in a better model. Despite including second order terms, the three-variable model, Model 2: Runs Batted In & Doubles & Walks, remains the best performing three-variable model for predicting team wins for the Chicago Cubs.

Through this multi-step analysis, the most effective model for predicting wins based on various combinations of batting metrics was successfully identified. The best overall model for predicting team wins for the Chicago Cubs is the three-variable model, Model 2: Runs Batted In & Doubles & Walks. While the best two-variable model was first revealed to be Model 3: Runs Batted In & Doubles, adding an additional term (BB walks) for the purpose of creating a three-variable model resulted in a better overall model. These results suggest that when combined, runs batted in, doubles, and walks are the most impactful predictors of team wins. Despite the exploration of an interaction model and second-order model, the three-variable model, Model 2: Runs Batted In & Doubles & Walks, remained on top. This model was chosen as the best overall model due to its strong statistical performance and interpretability. In the end, the analysis identified Model 2: Runs Batted In & Doubles & Walks as the most effective model for predicting wins based on key batting statistics for the Chicago Cubs.

Methodology

For the methodology of this analysis, several statistical techniques were implemented within SAS. SAS, also known as the Statistical Analysis System, is a software suite that allows users to perform data management and analysis, advanced analytics, predictive analytics and much more. To begin, a ‘proc freq’ statement was used in SAS to gather important basic information regarding the dataset including the

number of observations. For the purpose of the exploratory data analysis portion of this study, histograms were implemented within SAS. In SAS, histograms are generated with the ‘proc sgplot’ statement. Each variable was displayed with this technique in order to gather a better understanding before jumping into the bulk of the analysis. Histograms help to visualize distribution of variables in order to identify potential skewness, outliers, and variability. The purpose of using histograms in regards to this study is to better visualize each batting variable. Additionally, in SAS, a correlation analysis was done to find the four most relevant variables in regards to the dependent variable. This is done with a ‘proc corr’ statement and includes all eleven variables that were correlated with wins. After discovering the four most relevant variables in regards to wins, the ‘sgplot’ was utilized once more for the creation of scatterplots. Scatter plots are an impactful tool that help provide visual insights of the relationships between variables. Each of the four variables was displayed with this technique alongside the dependent variable in order to gather a better understanding of the distributions of data before jumping into the bulk of the analysis. Following this, our multiple regression models were built with the ‘proc reg’ statement. This was utilized in the testing of all combinations of variables to find the best model. The ‘proc reg’ statement was also utilized in the creation of the interaction term model and the second order term model. Although before that step, a ‘proc glm’ and ‘proc plm’ statement was used to generate an interaction plot for the justification of adding an interaction term. These techniques create a plot or a graph in which the slopes of the variables can be seen. This is helpful in determining if an interaction term is needed. Additionally, for the creation of both the interaction and second order models, two new data sets were created in order to introduce the

interaction terms and second order terms. Once generated, the new data sets including these new terms were utilized in each corresponding model. In summary, several statistical techniques were used within SAS for the purpose of this analysis in order to achieve the overall goal of identifying the best model for predicting team wins.

Conclusion

In conclusion, this analysis revealed that Model 2: Runs Batted In & Doubles & Walks (three-variable model) is the best overall model for predicting wins based on key batting statistics for the Chicago Cubs based on the years of 1999-2024. While one of the two-variable models, Model 3: Runs Batted In (RBI) & Doubles (DB) additionally revealed to perform very well in predicting wins for the Chicago Cubs, it was not enough to outshine the three-variable model. Model 2: Runs Batted In & Doubles & Walks, presented the highest overall adjusted r square, a p value <0.05, the lowest overall standard deviation, and lowest overall coefficient of variation, across all evaluated models, making it the most reliable and effective model. Despite testing interaction and second order models the three-variable model presented the best results. This suggests that when combined, runs batted in, doubles, and walks are the most impactful predictors of team wins for the Chicago Cubs. Given these outcomes, Model 2: Runs Batted In & Doubles & Walks is the best performing model for predicting team wins for the Chicago Cubs for the years of 1999-2024.

If I were to complete this analysis again, I would consider incorporating more batting variables. Likewise, I would be interested to expand the metrics to include things like pitching or defensive statistics. Investigating more potential key statistics could

ultimately enhance the models overall efficiency and further improve the team's success on the field. Additionally, I would consider expanding the dataset to include additional years, or even conducting the same analysis of the team's early years. This would offer valuable insight into how the Chicago Cubs performance has evolved over time. Additionally I would consider exploring the use of different software applications for better visualizations and analysis. Software I would be interested in using is R or Python as they offer powerful libraries such as ggplot2 (R) and Matplotlib (Python). These libraries are extremely useful for creating detailed, interactive plots that could provide better visualizations of the relationships between variables. I would keep these ideas in mind if I were to conduct a similar analysis, with the goal of refining the model by integrating more variables and testing different combinations of batting factors and incorporating the use of different softwares. With this I could improve predictions and provide deeper insights into team winning patterns over time.

I gained a substantial amount of knowledge through my analysis of building a multiple regression model that can accurately predict the number of wins based on various batting statistics for the Chicago Cubs. Through this, I was able to successfully determine the best performing model and therefore I am happy with the results. I am looking forward to implementing what I have learned from this analysis on future projects involving predictive modeling and more advanced statistical techniques.

IMPORTANT TABLES

2-variable Model Key Statistic Table

First Order Two Variable Models	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model 1: *Runs Batted In & Runs* Wins = 2.02743 + 0.08919(RBI) + 0.02215(R)	*0.3569*	*0.0030*	*9.46317*	*11.81714*
Model 2: *Runs Batted In & Walks* Wins = 3.07676 + 0.09110(RBI) + 0.02600(BB)	*0.3717*	*0.0023*	*9.35360*	*11.68031*
Model 3: *Runs Batted In & Doubles* **Wins = -28.61247 + 0.10274(RBI) + 0.13272(DB)**	***0.4460***	***0.0006***	***8.78262***	***10.96731***
Model 4: *Runs & Walks* Wins = 1.56177 + 0.08865(R) + 0.02626(BB)	*0.3700*	*0.0024*	*9.36584*	*11.69561*
Model 5: *Runs & Doubles* Wins = -29.71086 + 0.10001(R) + 0.13102(DB)	*0.4416*	*0.0006*	*8.81774*	*11.01116*
Model 6: *Walks & Doubles* Wins = -25.22385 + 0.09022(BB) + 0.20412(DB)	*0.4412*	*0.0006*	*8.82128*	*11.01558*

3-variable Model Key Statistic Table

First Order Three Variable Models	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model 1: *Runs Batted In & Doubles & Runs* Wins = -28.08574 + 0.13493(RBI) + 0.13340(DB) - 0.03165(R)	*0.4199*	*0.0022*	*8.98711*	*11.22267*
Model 2: *Runs Batted In & Doubles & Walks* **Wins = -35.26724 + 0.06011(RBI) + 0.16583(DB)** **+ 0.05079(BB)**	***0.4736***	***0.0008***	***8.56147***	***10.69114***

Interaction Model Key Statistic Table

Interaction Model	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model: *Wins = -101.91701 + 0.55681(RBI)* *- 0.08823(DB) - 0.21558(BB)* *- 0.00103(RBI_DB)* *- 0.00039707(RBI_BB)* *+ 0.00191(DB_BB)*	*0.4215*	*0.0112*	*8.97525*	*11.20786*

Second Order Model Key Statistic Table

Second Order Model	*R Squared Adjusted*	*F Test P-Value*	*Standard Deviation*	*Coefficient Variation*
Model 2: *Runs Batted In & Doubles & Walks* Wins = -264.86198 + 0.08995(RBI) + 1.01794(DB) + 0.42211(BB) - 0.00001633(RBI_2) - (0.00147)DB_2 - 0.00035868(BB_2)	*0.4225*	*0.0110*	*8.96708*	*11.19766*

SAS CODE + OUTPUT

```
/* Generated Code (IMPORT) */
/* Source File: sportsref_download 3(Worksheet).csv */
/* Source Path: /home/u63987812/sasuser.v94/Lab 2 */
/* Code generated on: 10/22/24, 6:23 PM */

%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/home/u63987812/sasuser.v94/Lab 2/sportsref_download 3(Worksheet).csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=WORK.cubs;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.cubs; RUN;


%web_open_table(WORK.IMPORT);

proc print data=cubs; run;
```

Obs	Year	Lg	W	L	Finish	R/G	G	PA	AB	R	H	DB	TP	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	E	DP	Fld%	BatAge
1	2024	NL Central	83	79	2	4.54	162	6116	5441	736	1318	253	29	170	696	143	30	546	1362	0.242	0.317	0.393	0.71	76	91	0.987	27.8
2	2023	NL Central	83	79	2	5.06	162	6220	5504	819	1399	269	30	196	786	140	34	570	1391	0.254	0.33	0.421	0.751	92	137	0.984	28.4
3	2022	NL Central	74	88	3	4.06	162	6072	5425	657	1293	265	31	159	620	111	37	507	1448	0.238	0.311	0.387	0.698	96	139	0.984	27.9
4	2021	NL Central	71	91	4	4.35	162	5972	5306	705	1255	225	26	210	672	86	37	502	1596	0.237	0.312	0.407	0.719	87	149	0.985	29.1
5	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	2019	NL Central	84	78	3	5.02	162	6195	5461	814	1378	270	26	256	783	45	24	581	1460	0.252	0.331	0.452	0.783	118	141	0.981	27.7
7	2018	NL Central	95	68	2	4.67	163	6389	5624	761	1453	286	34	167	722	66	38	576	1388	0.258	0.333	0.41	0.744	104	155	0.983	27.2
8	2017	NL Central	92	70	1	5.07	162	6283	5496	822	1402	274	29	223	785	62	31	622	1401	0.255	0.338	0.437	0.775	95	139	0.984	27.1
9	2016	NL Central	103	58	1	4.99	162	6335	5503	808	1409	293	30	199	767	66	34	656	1339	0.256	0.343	0.429	0.772	101	116	0.983	27.4
10	2015	NL Central	97	65	3	4.25	162	6200	5491	689	1341	272	30	171	657	95	37	567	1518	0.244	0.321	0.398	0.719	111	120	0.982	26.9
11	2014	NL Central	73	89	5	3.79	162	6102	5508	614	1315	270	31	157	590	65	40	442	1477	0.239	0.3	0.385	0.684	103	137	0.983	26.8
12	2013	NL Central	66	96	5	3.72	162	6079	5498	602	1307	297	18	172	576	63	32	439	1230	0.238	0.3	0.392	0.693	100	129	0.983	27.9
13	2012	NL Central	61	101	5	3.78	162	5967	5411	613	1297	265	36	137	570	94	45	447	1235	0.24	0.302	0.378	0.68	105	148	0.982	27.8
14	2011	NL Central	71	91	5	4.04	162	6130	5549	654	1423	285	36	148	610	69	23	425	1202	0.256	0.314	0.401	0.715	134	128	0.978	29.3
15	2010	NL Central	75	87	5	4.23	162	6140	5512	685	1414	298	27	149	658	55	31	479	1236	0.257	0.32	0.401	0.721	126	137	0.979	29.4
16	2009	NL Central	83	78	2	4.39	161	6244	5486	707	1398	293	29	161	678	56	34	592	1185	0.255	0.332	0.407	0.738	105	144	0.983	29.9
17	2008	NL Central	97	64	1	5.31	161	6384	5588	855	1552	329	21	184	811	87	34	636	1186	0.278	0.354	0.443	0.797	99	118	0.983	30.1
18	2007	NL Central	85	77	1	4.64	162	6268	5643	752	1530	340	28	151	711	86	33	500	1054	0.271	0.333	0.422	0.754	94	134	0.984	29.3
19	2006	NL Central	66	96	6	4.42	162	6147	5587	716	1496	271	46	166	677	121	49	395	928	0.268	0.319	0.422	0.741	106	122	0.982	28.6
20	2005	NL Central	79	83	4	4.34	162	6161	5584	703	1506	323	23	194	674	65	39	419	920	0.27	0.324	0.44	0.764	101	136	0.983	29.7
21	2004	NL Central	89	73	3	4.87	162	6281	5628	789	1508	308	29	235	755	66	28	489	1080	0.268	0.328	0.458	0.786	86	126	0.986	30.1
22	2003	NL Central	88	74	1	4.47	162	6187	5519	724	1431	302	24	172	691	73	31	492	1158	0.259	0.323	0.416	0.739	106	157	0.983	31.3
23	2002	NL Central	67	95	5	4.36	162	6242	5496	706	1351	259	29	200	676	63	21	585	1269	0.246	0.321	0.413	0.734	114	144	0.981	30.2
24	2001	NL Central	88	74	3	4.8	162	6219	5406	777	1409	268	32	194	748	67	36	577	1077	0.261	0.336	0.43	0.766	109	113	0.982	30.7
25	2000	NL Central	65	97	6	4.72	162	6397	5577	764	1426	272	23	183	722	93	37	632	1120	0.256	0.335	0.411	0.746	100	139	0.983	30.8
26	1999	NL Central	67	95	6	4.61	162	6201	5482	747	1411	255	35	189	717	60	44	571	1170	0.257	0.329	0.42	0.749	139	135	0.977	31.2

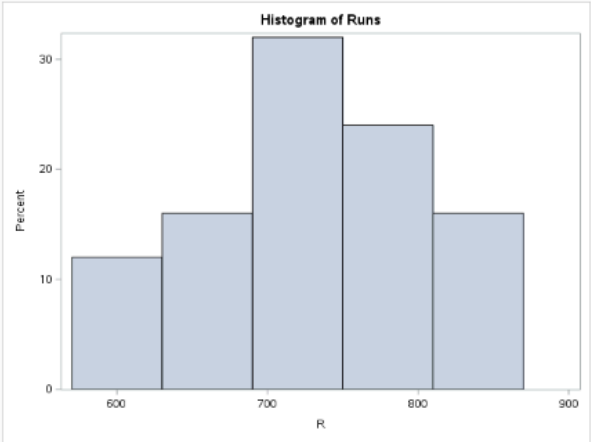
```
/* EDA */
/* total # observations */
proc freq data=cubs;
    run;
```

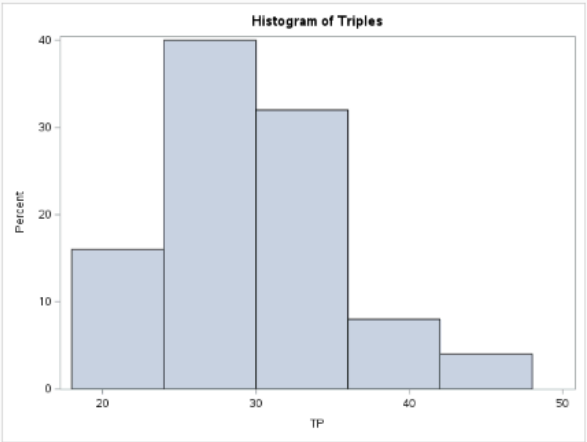
The FREQ Procedure

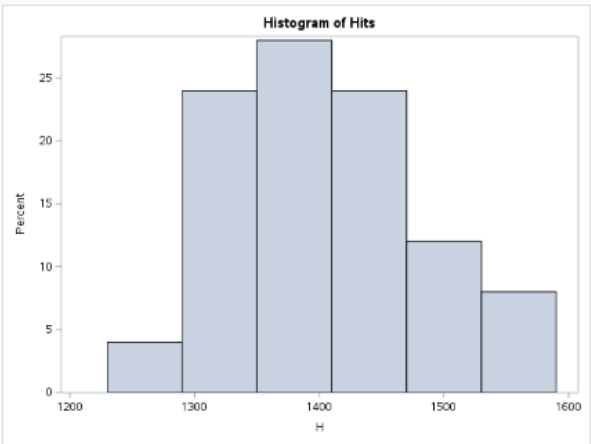
Year	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1999	1	4.00	1	4.00
2000	1	4.00	2	8.00
2001	1	4.00	3	12.00
2002	1	4.00	4	16.00
2003	1	4.00	5	20.00
2004	1	4.00	6	24.00
2005	1	4.00	7	28.00
2006	1	4.00	8	32.00
2007	1	4.00	9	36.00
2008	1	4.00	10	40.00
2009	1	4.00	11	44.00
2010	1	4.00	12	48.00
2011	1	4.00	13	52.00
2012	1	4.00	14	56.00
2013	1	4.00	15	60.00
2014	1	4.00	16	64.00
2015	1	4.00	17	68.00
2016	1	4.00	18	72.00
2017	1	4.00	19	76.00
2018	1	4.00	20	80.00
2019	1	4.00	21	84.00
2021	1	4.00	22	88.00
2022	1	4.00	23	92.00
2023	1	4.00	24	96.00
2024	1	4.00	25	100.00
Frequency Missing = 124				

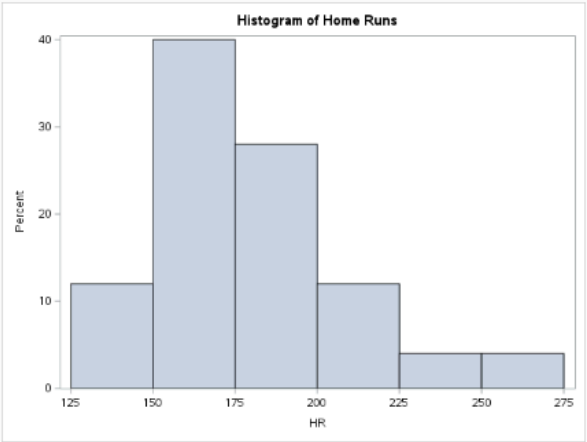
Lg	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NL Central	25	100.00	25	100.00
Frequency Missing = 124				

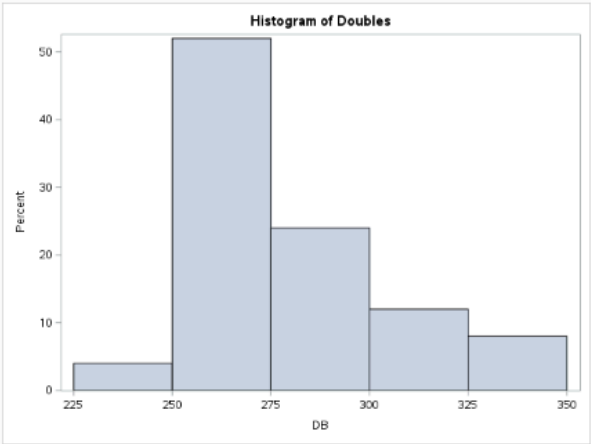
```
/* HISTOGRAMS */
/* 11 variables */
/* R (Runs) , H (Hits) , 2B (Doubles) , 3B (Triples) , HR (Home Runs) , RBI (Runs Batted In) ,
SB (Stolen Bases) , BB (Walks) , SO (Strikeouts) , BA (Batting Average) , BatAge (Age of Batters) */

/* histogram runs */
proc sgplot data=cubs;
    histogram R;
    TITLE "Histogram of Runs";
    run;

/* histogram hits */
proc sgplot data=cubs;
    histogram H;
    TITLE "Histogram of Hits";
    run;

/* histogram doubles */
proc sgplot data=cubs;
    histogram DB;
    TITLE "Histogram of Doubles";
    run;

/* histogram triples */
proc sgplot data=cubs;
    histogram TP;
    TITLE "Histogram of Triples";
    run;

/* histogram home runs */
proc sgplot data=cubs;
    histogram HR;
    TITLE "Histogram of Home Runs";
    run;

/* histogram runs batted in */
proc sgplot data=cubs;
    histogram RBI;
    TITLE "Histogram of Runs Batted In";
    run;

/* histogram stolen bases */
proc sgplot data=cubs;
    histogram SB;
    TITLE "Histogram of Stolen Bases";
    run;

/* histogram walks */
proc sgplot data=cubs;
    histogram BB;
    TITLE "Histogram of Walks";
    run;

/* histogram strike outs */
proc sgplot data=cubs;
    histogram SO;
    TITLE "Histogram of Strike Outs";
    run;

/* histogram batting average */
proc sgplot data=cubs;
    histogram BA;
    TITLE "Histogram of Walks";
    run;

/* histogram Age of Batters */
proc sgplot data=cubs;
    histogram BatAge;
    TITLE "Histogram of Age of Batters";
    run;

/* histogram wins */
proc sgplot data=cubs;
    histogram W;
    TITLE "Histogram of Wins";
    run;
```

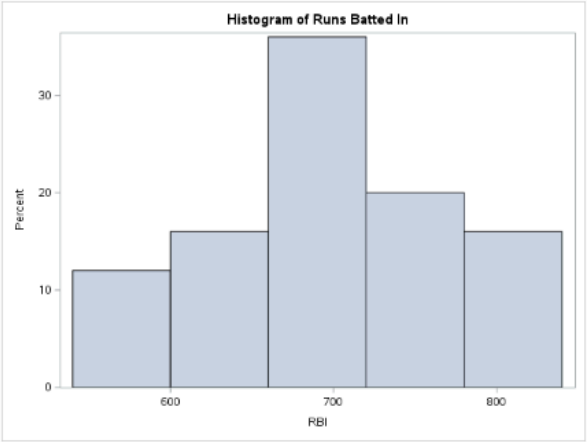

Histogram of Runs
Percent
0
10
20
30
600
700
800
900
R

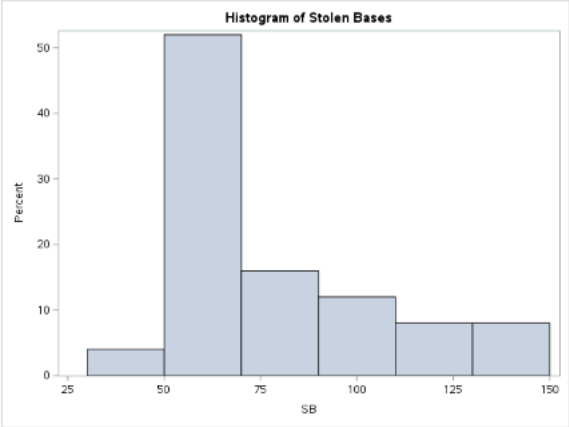

Histogram of Triples
Percent
0
10
20
30
40
20
30
40
50
TP

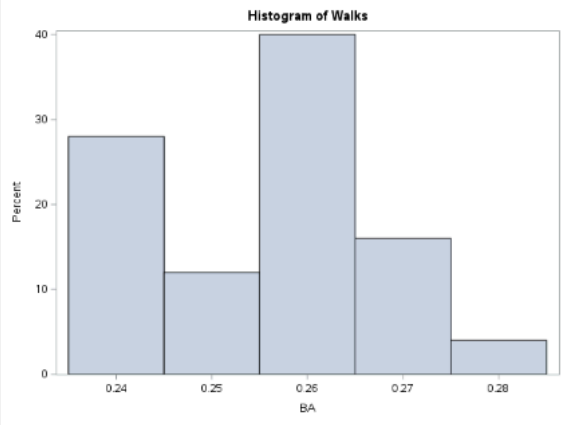

Histogram of Hits
Percent
0
5
10
15
20
25
1200
1300
1400
1500
1600
H

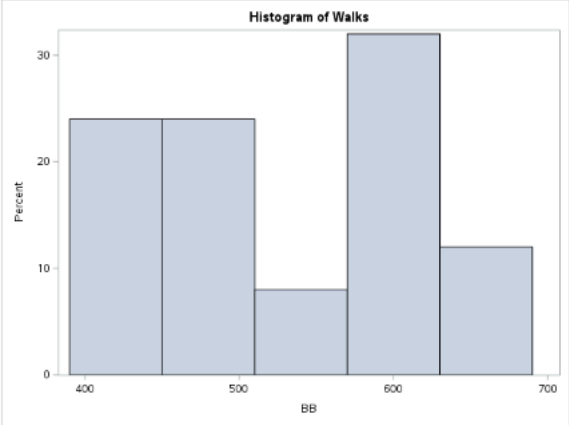

Histogram of Home Runs
Percent
0
10
20
30
40
125
150
175
200
225
250
275
HR

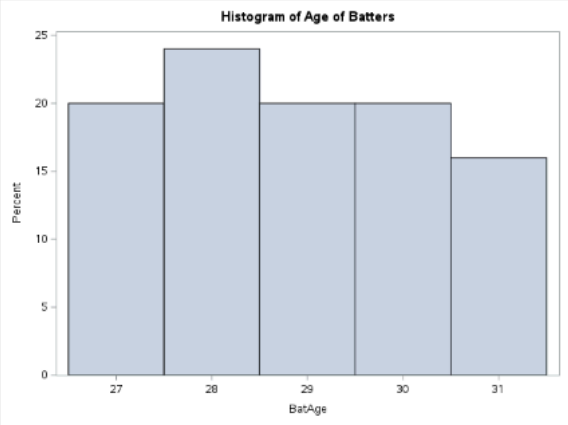

Histogram of Doubles
Percent
0
10
20
30
40
50
225
250
275
300
325
350
DB

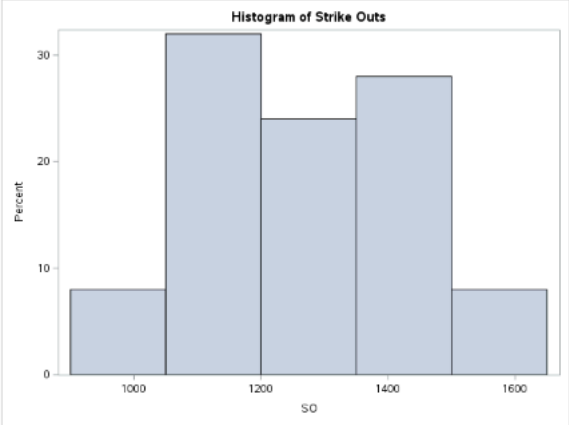

Histogram of Runs Batted In
Percent
0
10
20
30
600
700
800
RBI

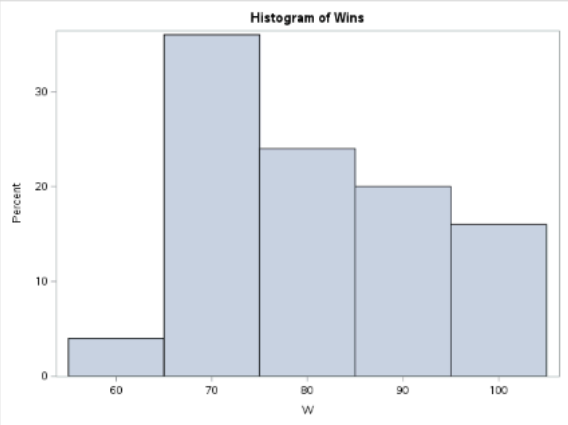

Histogram of Stolen Bases
Percent
0
10
20
30
40
50
25
50
75
100
125
150
SB


Histogram of Walks
Percent
0
10
20
30
40
0.24
0.25
0.26
0.27
0.28
BA


Histogram of Walks
Percent
0
10
20
30
400
500
600
700
BB


Histogram of Age of Batters
Percent
0
5
10
15
20
25
27
28
29
30
31
BatAge


Histogram of Strike Outs
Percent
0
10
20
30
1000
1200
1400
1600
SO


Histogram of Wins
Percent
0
10
20
30
60
70
80
90
100
W

```
/* four variables most strongly associated w team wins (W) based on correlation coefficients:
RBI (Runs Batted In): 0.64058
R (Runs): 0.63919
BB (Walks): 0.53792
DB (Doubles): 0.39551*/
proc corr data = cubs;
var R  H DB TP  HR  RBI  SB  BB  SO  BA  BatAge;
with W;
run;
```

The CORR Procedure

1 With Variables:	W
11 Variables:	R H DB TP HR RBI SB BB SO BA BatAge

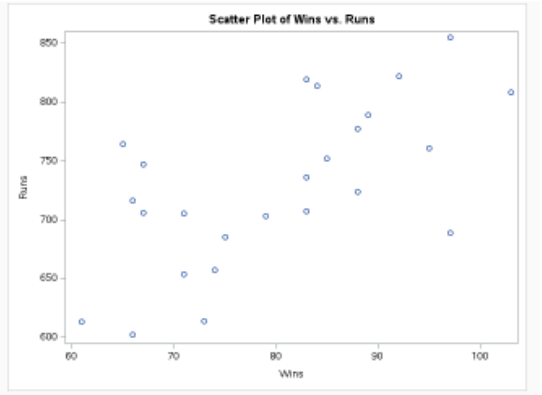
Simple Statistics

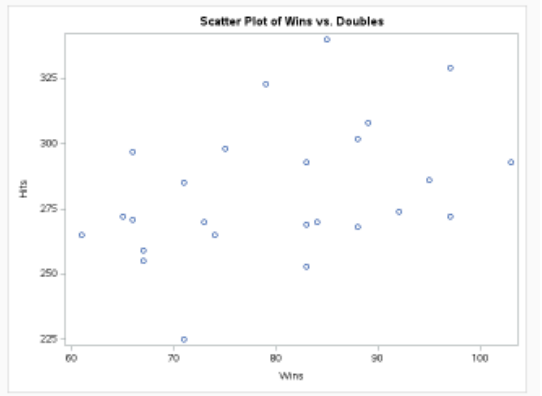
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
W	25	80.08000	11.80014	2002	61.00000	103.00000
R	25	728.76000	69.14916	18219	602.00000	855.00000
H	25	1401	78.95427	35022	1255	1552
DB	25	281.68000	25.85169	7042	225.00000	340.00000
TP	25	29.28000	5.66804	732.00000	18.00000	46.00000
HR	25	181.72000	28.62912	4543	137.00000	256.00000
RBI	25	694.08000	67.63438	17352	570.00000	811.00000
SB	25	79.88000	25.84138	1997	45.00000	143.00000
BB	25	529.88000	75.58236	13247	395.00000	656.00000
SO	25	1257	178.02949	31430	920.00000	1596
BA	25	0.25420	0.01149	6.35500	0.23700	0.27800
BatAge	25	28.90400	1.40993	722.60000	26.80000	31.30000

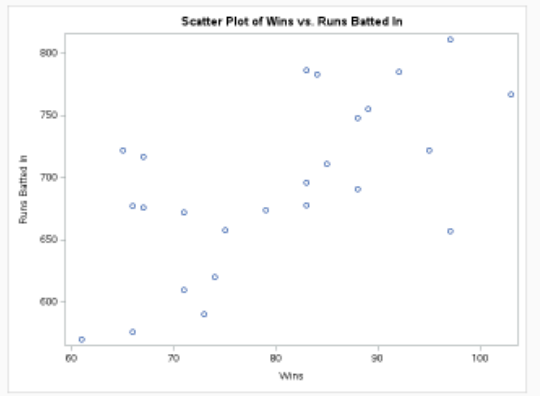
Pearson Correlation Coefficients, N = 25
Prob > |r| under H0: Rho=0

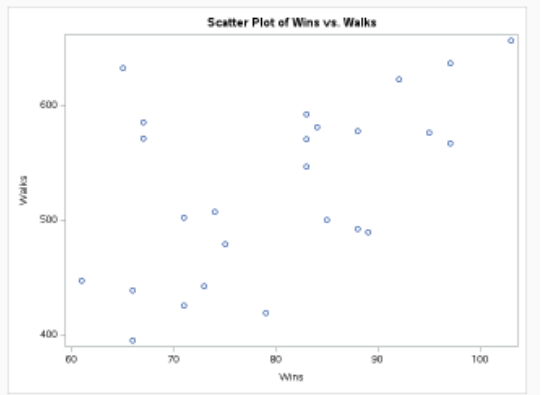
	R	H	DB	TP	HR	RBI	SB	BB	SO	BA	BatAge
W	0.63919 0.0006	0.37487 0.0648	0.39551 0.0504	-0.18849 0.3669	0.30607 0.1368	0.64058 0.0006	-0.09357 0.6564	0.53792 0.0055	0.17679 0.3979	0.38589 0.0568	-0.19937 0.3393

```
/* PROC CORR */
/* scatterplots */
/* wins vs runs scatterplot */
proc sgplot data=cubs;
    scatter x=W y=R;
    title "Scatter Plot of Wins vs. Runs";
    xaxis label="Wins";
    yaxis label="Runs";
    run;

/* wins vs doubles scatterplot */
proc sgplot data=cubs;
    scatter x=W y=DB;
    title "Scatter Plot of Wins vs. Doubles";
    xaxis label="Wins";
    yaxis label="Hits";
    run;

/* wins vs runs batted in scatterplot */
proc sgplot data=cubs;
    scatter x=W y=RBI;
    title "Scatter Plot of Wins vs. Runs Batted In";
    xaxis label="Wins";
    yaxis label="Runs Batted In";
    run;

/* wins vs walks scatterplot */
proc sgplot data=cubs;
    scatter x=W y=BB;
    title "Scatter Plot of Wins vs. Walks";
    xaxis label="Wins";
    yaxis label="Walks";
    run;
```









```
/* PROC REGG */
/* 6 two variable model*/
/* Wins (W) = RBI (Runs Batted In) , R (Runs) , BB (Walks) , DB (Doubles) */

/* Model 1 */
/* Runs Batted In vs. Runs */
proc reg data=cubs;
    model W = RBI R;
    title "Multiple Regression: Runs Batted In vs. Runs";
    run;

/* Model 2 */
/* Runs Batted In vs. Walks */
proc reg data=cubs;
    model W = RBI BB;
    title "Multiple Regression: Runs Batted In vs. Walks";
    run;
```

Multiple Regression: Runs Batted In vs. Runs

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1371.70572	685.85286	7.66	0.0030
Error	22	1970.13428	89.55156		
Corrected Total	24	3341.84000			

Root MSE	9.46317	R-Square	0.4105
Dependent Mean	80.08000	Adj R-Sq	0.3569
Coeff Var	11.81714		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.02743	21.14775	0.10	0.9245
RBI	1	0.08919	0.33520	0.27	0.7926
R	1	0.02215	0.32785	0.07	0.9467

Multiple Regression: Runs Batted In vs. Walks

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1417.06554	708.53277	8.10	0.0023
Error	22	1924.77446	87.48975		
Corrected Total	24	3341.84000			

Root MSE	9.35360	R-Square	0.4240
Dependent Mean	80.08000	Adj R-Sq	0.3717
Coeff Var	11.68031		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.07676	19.69841	0.16	0.8773
RBI	1	0.09110	0.04016	2.27	0.0335
BB	1	0.02600	0.03594	0.72	0.4771

```
/* Model 3 */
/* Runs Batted In vs. Doubles */
proc reg data=cubs;
    model W = RBI DB;
    title "Multiple Regression: Runs Batted In vs. Doubles";
    run;

/* Model 4 */
/* Runs vs. Walks */
proc reg data=cubs;
    model W = R BB;
    title "Multiple Regression: Runs vs. Walks";
    run;
```

Multiple Regression: Runs Batted In vs. Doubles

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1644.88136	822.44068	10.66	0.0006
Error	22	1696.95864	77.13448		
Corrected Total	24	3341.84000			

Root MSE	8.78262	R-Square	0.4922
Dependent Mean	80.08000	Adj R-Sq	0.4460
Coeff Var	10.96731		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-28.61247	24.79153	-1.15	0.2608
RBI	1	0.10274	0.02694	3.81	0.0009
DB	1	0.13272	0.07047	1.88	0.0729

Multiple Regression: Runs vs. Walks

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1412.02165	706.01082	8.05	0.0024
Error	22	1929.81835	87.71902		
Corrected Total	24	3341.84000			

Root MSE	9.36584	R-Square	0.4225
Dependent Mean	80.08000	Adj R-Sq	0.3700
Coeff Var	11.69561		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.56177	20.27909	0.08	0.9393
R	1	0.08865	0.03936	2.25	0.0346
BB	1	0.02626	0.03601	0.73	0.4735

```
/* Model 5 */
/* Runs vs. Doubles */
proc reg data=cubs;
    model W = R DB;
    title "Multiple Regression: Runs vs. Doubles";
    run;

/* Model 6 */
/* Walks vs. Doubles */
proc reg data=cubs;
    model W = BB DB;
    title "Multiple Regression: Walks vs. Doubles";
    run;
```

Multiple Regression: Runs vs. Doubles

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1631.28518	815.64259	10.49	0.0006
Error	22	1710.55482	77.75249		
Corrected Total	24	3341.84000			

Root MSE	8.81774	R-Square	0.4881
Dependent Mean	80.08000	Adj R-Sq	0.4416
Coeff Var	11.01116		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-29.71086	25.12736	-1.18	0.2497
R	1	0.10001	0.02649	3.78	0.0010
DB	1	0.13102	0.07085	1.85	0.0779

Multiple Regression: Walks vs. Doubles

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1629.91176	814.95588	10.47	0.0006
Error	22	1711.92824	77.81492		
Corrected Total	24	3341.84000			

Root MSE	8.82128	R-Square	0.4877
Dependent Mean	80.08000	Adj R-Sq	0.4412
Coeff Var	11.01558		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-25.22385	24.42201	-1.03	0.3129
BB	1	0.09022	0.02392	3.77	0.0011
DB	1	0.20412	0.06993	2.92	0.0080

```
/* best 2 var model : Model 3 Runs Batted In vs. Doubles */
/*2 three variable models  */

/* Model 1 */
/* Runs Batted In vs. Doubles vs. Runs */
/* example models if RBI and R were best 2 var models */
proc reg data=cubs;
    model W = RBI DB R;
    title "Multiple Regression: Runs Batted In vs. Doubles vs. Runs";
    run;

/* Model 2 */
/* Runs Batted In vs. Doubles vs. Walks */
proc reg data=cubs;
    model W = RBI DB BB;
    title "Multiple Regression: Runs Batted In vs. Doubles vs. Walks";
    run;
```

Multiple Regression: Runs Batted In vs. Doubles vs. Runs

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1645.70860	548.56953	6.79	0.0022
Error	21	1696.13140	80.76816		
Corrected Total	24	3341.84000			

Root MSE	8.98711	R-Square	0.4925
Dependent Mean	80.08000	Adj R-Sq	0.4199
Coeff Var	11.22267		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-28.08574	25.89714	-1.08	0.2904
RBI	1	0.13493	0.31930	0.42	0.6769
DB	1	0.13340	0.07243	1.84	0.0797
R	1	-0.03165	0.31273	-0.10	0.9203

Multiple Regression: Runs Batted In vs. Doubles vs. Walks

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1802.56754	600.85585	8.20	0.0008
Error	21	1539.27246	73.29869		
Corrected Total	24	3341.84000			

Root MSE	8.56147	R-Square	0.5394
Dependent Mean	80.08000	Adj R-Sq	0.4736
Coeff Var	10.69114		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-35.26724	24.58946	-1.43	0.1662
RBI	1	0.06011	0.03917	1.53	0.1398
DB	1	0.16583	0.07231	2.29	0.0322
BB	1	0.05079	0.03463	1.47	0.1573

```
/* interaction term model */

proc glm data=cubs;
    model W = RBI DB BB DB*BB / solution;
    store GLMMODEL;
run;

proc plm restore=GLMMODEL noinfo;
    effectplot slicefit(x=DB sliceby=BB);
    title "Interaction Effect of Doubles (DB) and Walks (BB) on Wins (W)";
run;
```

The GLM Procedure

Number of Observations Read	149
Number of Observations Used	25

The GLM Procedure

Dependent Variable: W

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1835.341736	458.835434	6.09	0.0023
Error	20	1506.498264	75.324913		
Corrected Total	24	3341.840000			

R-Square	Coeff Var	Root MSE	W Mean
0.549201	10.83790	8.678993	80.08000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
RBI	1	1371.296823	1371.296823	18.21	0.0004
DB	1	273.584541	273.584541	3.63	0.0712
BB	1	157.686172	157.686172	2.09	0.1634
DB*BB	1	32.774200	32.774200	0.44	0.5170

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RBI	1	179.9154073	179.9154073	2.39	0.1379
DB	1	10.0754009	10.0754009	0.13	0.7184
BB	1	18.4377590	18.4377590	0.24	0.6262
DB*BB	1	32.7742001	32.7742001	0.44	0.5170

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	72.37071739	165.0735760	0.44	0.6658
RBI	0.06144195	0.0397558	1.55	0.1379
DB	-0.21022215	0.5747998	-0.37	0.7184
BB	-0.15627964	0.3158767	-0.49	0.6262
DB*BB	0.00071829	0.0010889	0.66	0.5170

```
data cubs_2;
    set cubs;
    RBI_DB=RBI*DB;
    RBI_BB=RBI*BB;
    DB_BB=DB*BB;
```

```
proc reg data=cubs_2;
    model W=RBI DB BB RBI_DB RBI_BB DB_BB;
    run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1891.84671	315.30779	3.91	0.0112
Error	18	1449.99329	80.55518		
Corrected Total	24	3341.84000			

Root MSE	8.97525	R-Square	0.5661
Dependent Mean	80.08000	Adj R-Sq	0.4215
Coeff Var	11.20786		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-101.91701	309.74016	-0.33	0.7459
RBI	1	0.55681	0.73235	0.76	0.4569
DB	1	-0.08823	0.97112	-0.09	0.9286
BB	1	-0.21558	0.52145	-0.41	0.6842
RBI_DB	1	-0.00103	0.00220	-0.47	0.6445
RBI_BB	1	-0.00039707	0.00050066	-0.79	0.4381
DB_BB	1	0.00191	0.00208	0.92	0.3712

```
/* second order term model */
data cubs_3;
    set cubs;
    RBI_2=RBI*RBI;
    DB_2=DB*DB;
    BB_2=BB*BB;


proc reg data=cubs_3;
    model W=RBI DB BB RBI_2 DB_2 BB_2;
    run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	149
Number of Observations Used	25
Number of Observations with Missing Values	124

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1894.48529	315.74755	3.93	0.0110
Error	18	1447.35471	80.40860		
Corrected Total	24	3341.84000			

Root MSE	8.96708	R-Square	0.5669
Dependent Mean	80.08000	Adj R-Sq	0.4225
Coeff Var	11.19766		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-264.86198	287.38405	-0.92	0.3689
RBI	1	0.08995	0.53859	0.17	0.8692
DB	1	1.01794	1.20551	0.84	0.4095
BB	1	0.42211	0.40696	1.04	0.3134
RBI_2	1	-0.00001633	0.00039145	-0.04	0.9672
DB_2	1	-0.00147	0.00210	-0.70	0.4942
BB_2	1	-0.00035868	0.00038844	-0.92	0.3680