STP 429 - Regression Analysis

Arizona State University

Takahiro Wada

Lab #2

Executive Summary

Winning the World Series is the ultimate goal for any MLB team. With each MLB team viciously competing with one another to get their hands on the Commissioner's trophy. Hence, it is important to understand what determines their wins. We will be focusing on the MLB team the New York Mets. Understanding what determines wins for the New York Mets allows the Mets to improve their roster by identifying key statistics from their batting statistics report.

Many factors determine wins for the New York Mets, so a study was conducted to try and predict major win factors for the Mets. Factors that are considered critical when predicting wins for the New York Mets are Runs, Hits, Doubles, Triples, Home Runs, Runs Batted In, Stolen Bases, Walks, Strikeouts, Batting Average, and Age of Batters. A statistical analysis was performed to determine which win factors are most important when building a model to predict wins for the New York Mets.

Using multiple regression techniques to help analyze each of their factors and relevance, I can conclude that Hits, Runs Batted In, and Walks can be used to predict wins for the New York Mets. The following report will include details of the analysis. The model that was developed will help the New York Mets manage how these win factors impact their wins.

Introduction

This study was generated in order to determine if there are factors that affect wins so we can build a prediction model for the New York Mets. We have 11 variables that may have a strong correlation to wins for the New York Mets such as Runs, Hits, Doubles, Triples, Home Runs, Runs Batted In, Stolen Bases, Walks, Strikeouts, Batting Average, and Age of Batters which are included for our prediction model in our study. The New York Mets rely on identifying win factors in order to maximize their chance of winning games.

Analysis

To determine the best model for predicting wins for the New York Mets, I first analyzed each variables' scatter plots. The scatter plots allow us to see the relationship between independent variables vs wins and determine whether the independent variable has any association to our dependent variable. The scatter plot alone does not tell us which independent variable is significant to our prediction model, but does give us an idea. So, I used a correlation matrix which allows us to determine which independent variables have a strong association to our dependent variable. The chart allows us to see the strength for each variable, also to see if we can identify any outliers within our data set. Once we decide which independent variables are acceptable to be used, I tested different combinations of models and ran a regression analysis to find our best prediction model for determining wins for the New York Mets with our independent variables. Next, I tested our best prediction model to see whether an interaction term or quadratic term will help better predict wins for the New York Mets. Finally, I used a stepwise method, so the analysis will stop when all independent variables have been confirmed and are acceptable to be used for the model.

Data Section

From Table 27 , the descriptive analysis for our 12 variables used for our model:

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
W	W	25	81.1600000	8.8866942	66.0000000	97.0000000
R	R	25	713.6400000	69.8032234	619.0000000	853.0000000
H	H	25	1404.20	81.9827218	1242.00	1553.00
HR	HR	25	163.4800000	36.5560301	95.0000000	242.0000000
_B	2B	25	277.1200000	21.4773524	228.0000000	323.0000000
_B0	3B	25	27.5200000	10.0917458	14.0000000	49.0000000
RBI	RBI	25	679.6000000	66.3525935	593.0000000	814.0000000
SB	SB	25	97.8800000	40.3074848	42.0000000	200.0000000
BB	BB	25	537.3200000	60.2416799	445.0000000	717.0000000
SO	SO	25	1147.92	148.8421759	928.0000000	1404.00
BA	BA	25	0.2549600	0.0122423	0.2340000	0.2790000
BatAge	BatAge	25	29.0320000	1.0490948	27.2000000	30.7000000

The Wins range from 66 to 97 games won. Runs range from 619 to 853 runs. Hits range from 1242 to 1553 hits. Home runs range from 95 to 242 home runs. Doubles range from 228 to 323 doubles. Triples range from 14 to 49 triples. Runs batted in range from 593 to 814. Stolen bases range from 42 to 200 bases stolen. Walks range from 445 to 717 walks. Strikeouts range from 928 to 1404 strikeouts. Batting average ranges from 0.234 to 0.279. Age of the batter ranges from 27 to 30.7 years old. We are given a total of 11 independent variables. At the end, I used Hits, Runs Batted In, and Walks for my 3 variables. So to identify which variables are viable for our model, I analyzed both the scatter plots and the correlation matrix of each variable to determine if there is a relationship between the independent variables and dependent variables. I first identified which variables displayed a clear relationship between the independent variables and the dependent variable from their respective scatter plots.. I determined that Runs, Runs Batted In and Walks has a moderate association with wins for the New York Mets. To ensure I have got all the viable variables for our model, I ran a correlation matrix to look at each p-value for each variable. I only accepted variables that had a p-value less than 0.05 which were Runs, Hits, Home Runs, Runs Batted In, and Walks. I rejected only the Batting Average which is explained in the results section.

Results

Looking at our scatter plots for each variable from Table 1 - Table 12, we see that only Runs, Runs Batted In, and Walks have a moderate, positive, and a linear relationship, with a little bit of variation towards the bottom left of the graph, with wins for the New York Mets while the other scatter plots show a weak or no relationship with wins for the New York Mets. Scatter plots alone are not viable for choosing our variables for our prediction model and we do not want to miss any important variables that will improve our model.

So I next looked at our correlation matrix in Table 13, we see that Runs, Hits, Home Runs, Runs Batted In, and Walks are best correlated with predicting wins. I chose not to accept the Batting Average as it is too close to our acceptable P-value = 0.05 and the Batting Average scatter plot in Table 11 is weak as

well. I also noticed that the most significantly correlated variable with wins for the New York Mets is Runs Batted In with a linear correlation coefficient of 0.69049. This is further supported by the scatter plot in Table 6, showing a moderate, positive and linear association as seen as the data points are moderately close together and follow a positive trend. The other variables such as Doubles, Triples, Stolen Bases, Strikeouts, Batting Average, and Age of Batters have little significance to our dependent variable which we can ignore for our prediction model.

With our 5 variables we accepted to be significant to our model, I tested a total of 10 different models to determine our best model for predicting wins for the New York Mets. Looking at Table 24, I tested and collected statistical data from each model, Table 14 -Table 23, and plugged it into our model table. Comparing each model to one another, I concluded that the model **Wins (y) = B0 + B1(Hits, x1) + B2(Runs Batted In, x2) + B3(Walks, x3)** best predicts wins for the New York Mets. The reason I chose this model is because it has the highest F-Value = 9.35 & lowest P-Value = 0.0004, highest R^2 Adjusted value = 0.5108, Lowest Coefficient of Variation = 7.65832, and the best T-Values & P-Values compared to all other model, Table 22.

Since we have decided our best model, I decided to test one interaction term and quadratic term to see if they help improve our model for predicting wins for the New York Mets. For the interaction term: Ho: B4 = 0; Ha: B4 does not equal to 0. When we look at Table 28, our t-value=1.34 and p-value=0.1999. We can see that the p-value is greater than the acceptable value of 0.05. It is safe to conclude that we fail to reject the null hypothesis and the interaction term has no significance in the model to predicting wins for the New York Mets. For the quadratic term: Ho: B7 = 0; Ha: B7 does not equal to 0. When we look at Table 28, our t-value = -1.53 and p-value = 0.1468. We can see that the p-value is greater than the acceptable value of 0.05. It is safe to conclude that we fail to reject the null hypothesis and the quadratic term has no significance in the model to predicting wins for the New York Mets.

The stepwise regression method resulted in a model which included the best 3 variables out of the 11 variables. This is because compared to all other variables and models, they have the most significant values which will best predict wins for the New York Mets. The final model is the following, Table 22:

Wins (y) = 31.41678 - 0.02798(Hits, x1) + 0.09406(Runs Batted In, x2) + 0.04673(Walks, x3)

The interpretation of each coefficient of the final model is as follows: B0 = 31.41678 represents the estimated mean wins for the New York Mets when all independent variables are equal to 0. B1 = -0.02798: We estimate the mean wins for the New York Mets to decrease by 0.02798 for every 1 unit increase in Hits, x1, when all other variables are held fixed. B2 = 0.09406: We estimate the mean wins for the New York Mets to increase by 0.09406 for every 1 unit increase in Runs Batted In, x2 when all other variables are held fixed. B3 = 0.04673: We estimate the mean wins for the New York Mets to increase by 0.04673 for every 1 unit increase in Walks, x3, when all other variables are held fixed.

The final model has an R-Square Adjusted = 0.5108 which means that 51.08% of the variation in wins for the New York Mets are explained by Hits, Runs Batted In, and Walks. 0.5108 is just above the acceptable value for R-Squared Adjusted of 50%. The F-Statistic = 9.35 and P-Value = 0.0004, Table 22, shows our model is significant.

The residual plot in Table 25 shows that there is a random disbursement of data points on the plot which is the appropriate graph outcome. The normal probability plot in Table 26 shows that points are generally close to the line. There is an outlier on the bottom left corner of the graph.

In conclusion, I am able to create a statistically significant model to predict wins for the New York Mets using Hits, Runs Batted In, and Walks.

Future Work

Given that I am not aware of the player's individual batting data, it is difficult to determine what affects the wins for the New York Mets and whether this study needs to be expanded to cover all the players batting data.

Appendix

Table 1 - Scatter Plot of Wins vs Runs

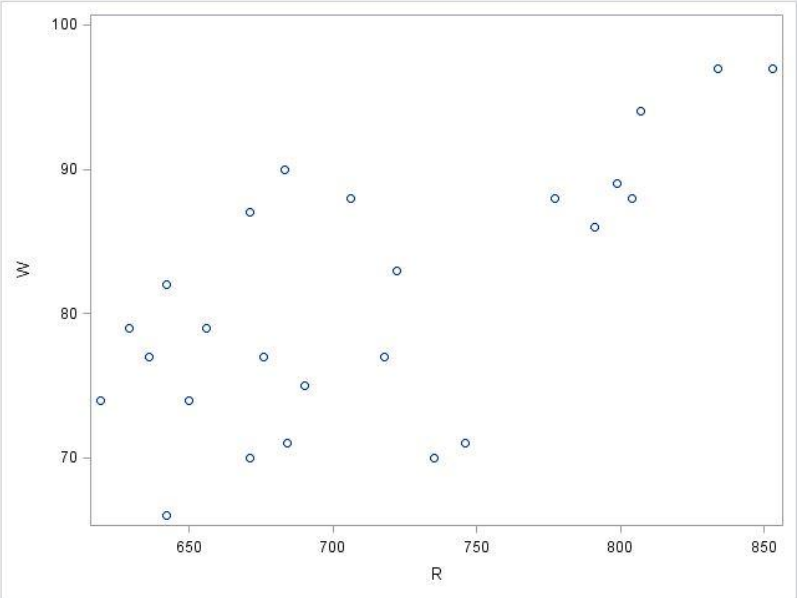

Table 2 - Scatter Plot of Wins vs Hits

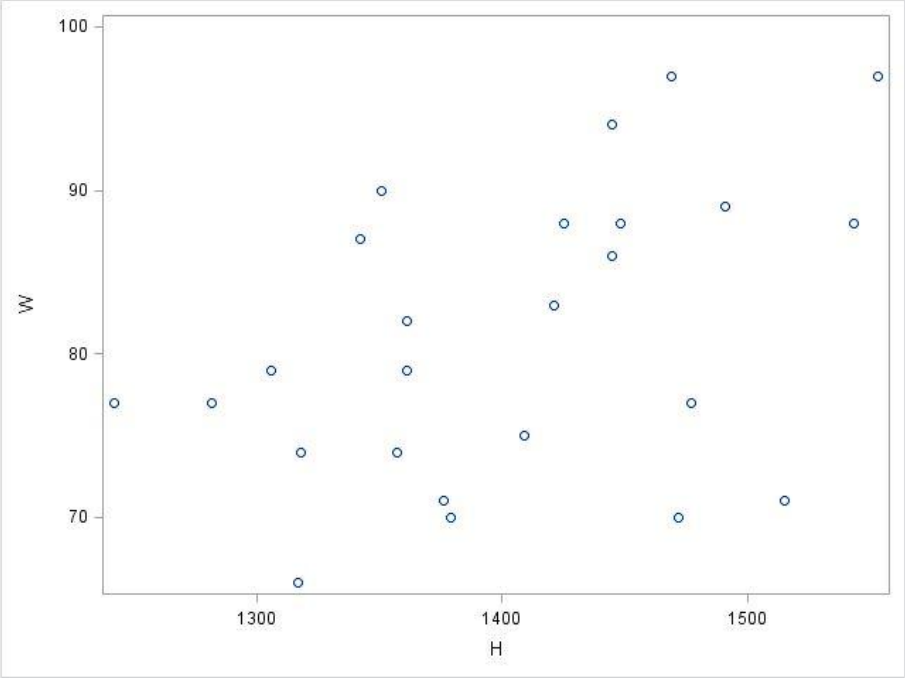

Table 3 - Scatter Plot of Wins vs Doubles

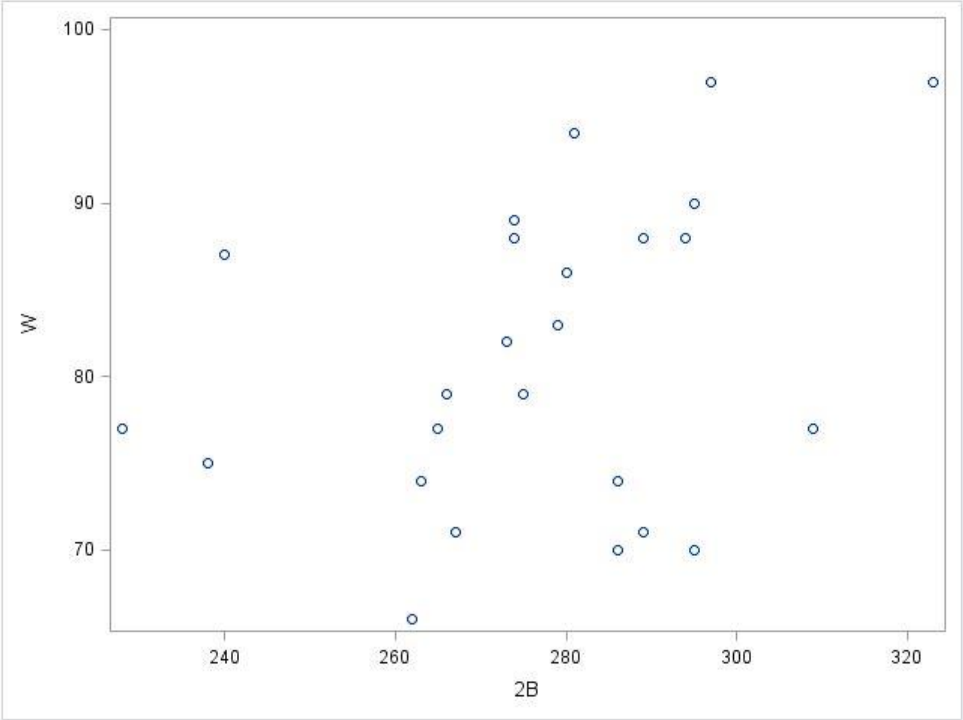

Table 4 - Scatter Plot of Wins vs Triples

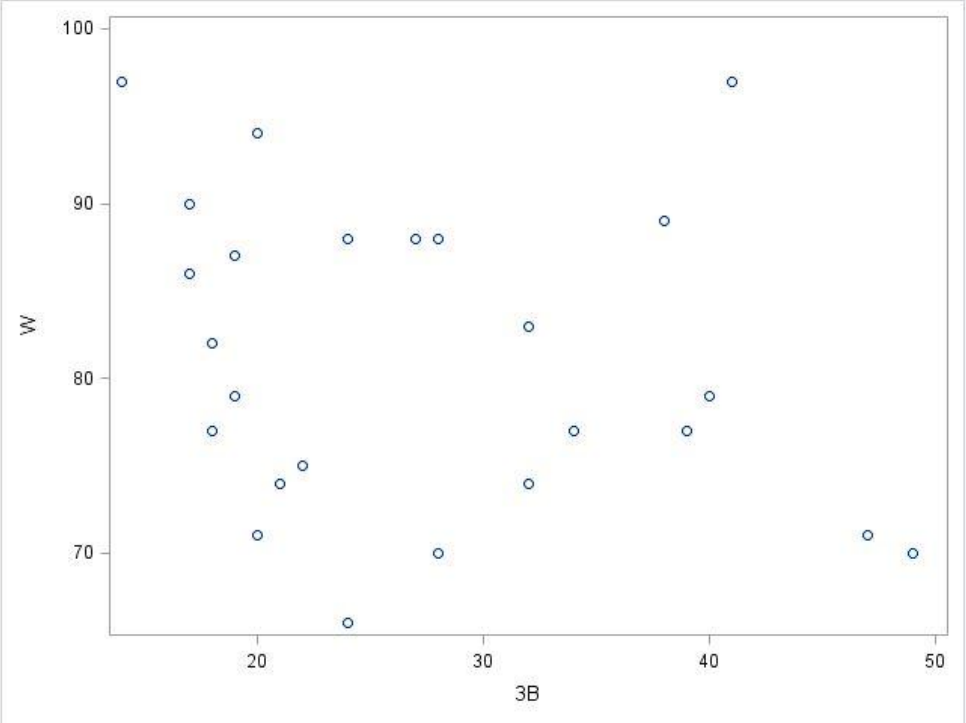

Table 5 -Scatter Plot of Wins vs Home Runs

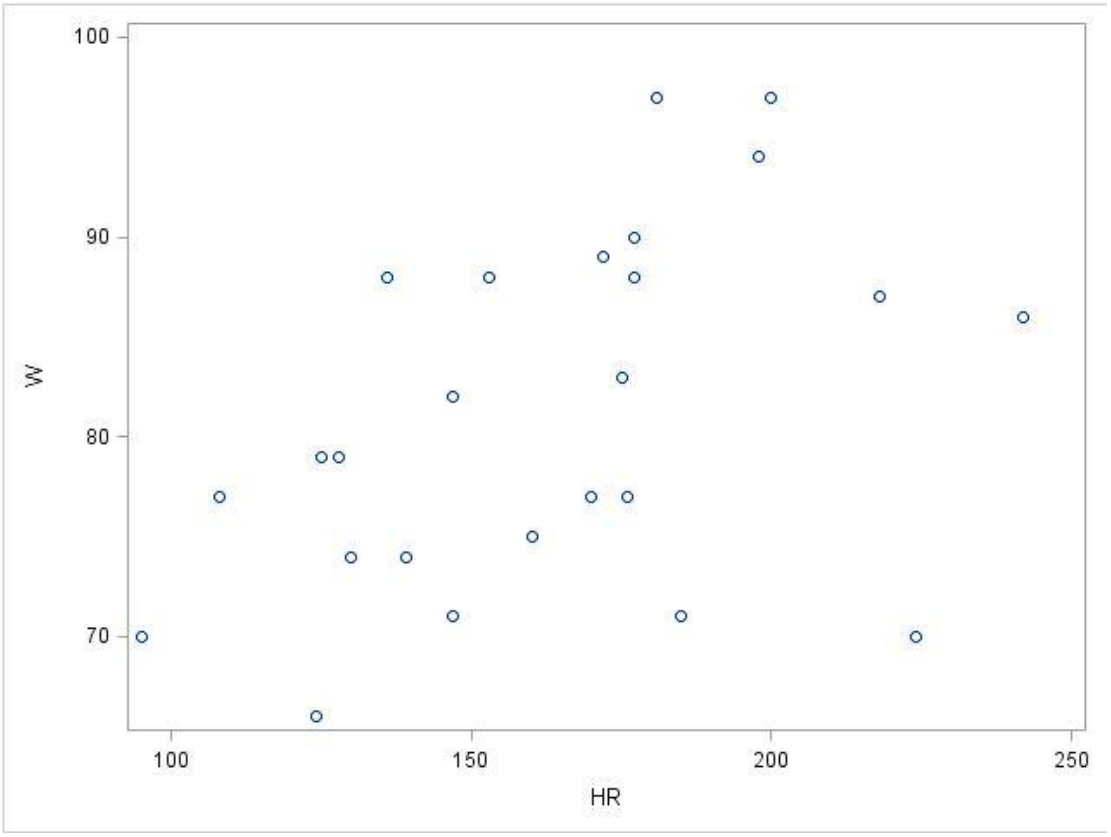

Table 6 - Scatter Plot of Wins vs Runs Batted In

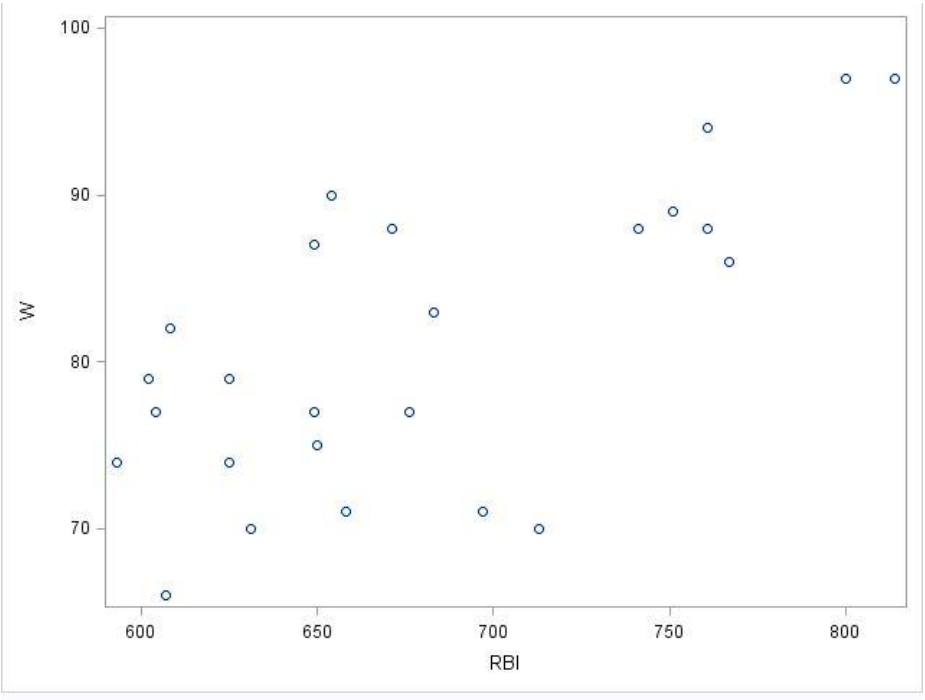

Table 7 - Scatter Plot of Wins vs Runs Batted In

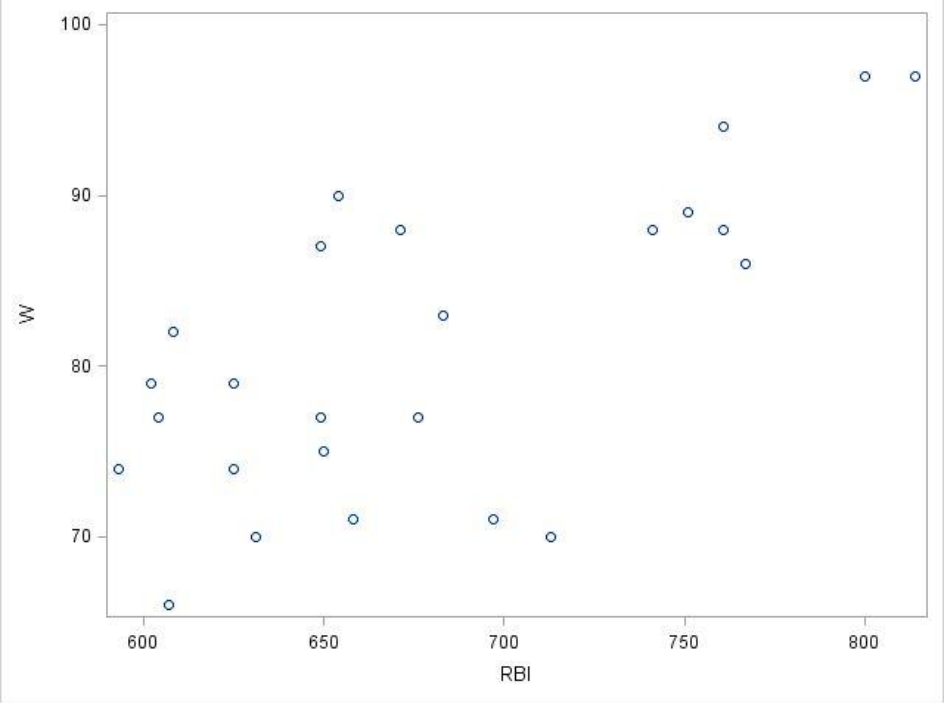

Table 8 - Scatter Plot of Wins vs Stolen Bases

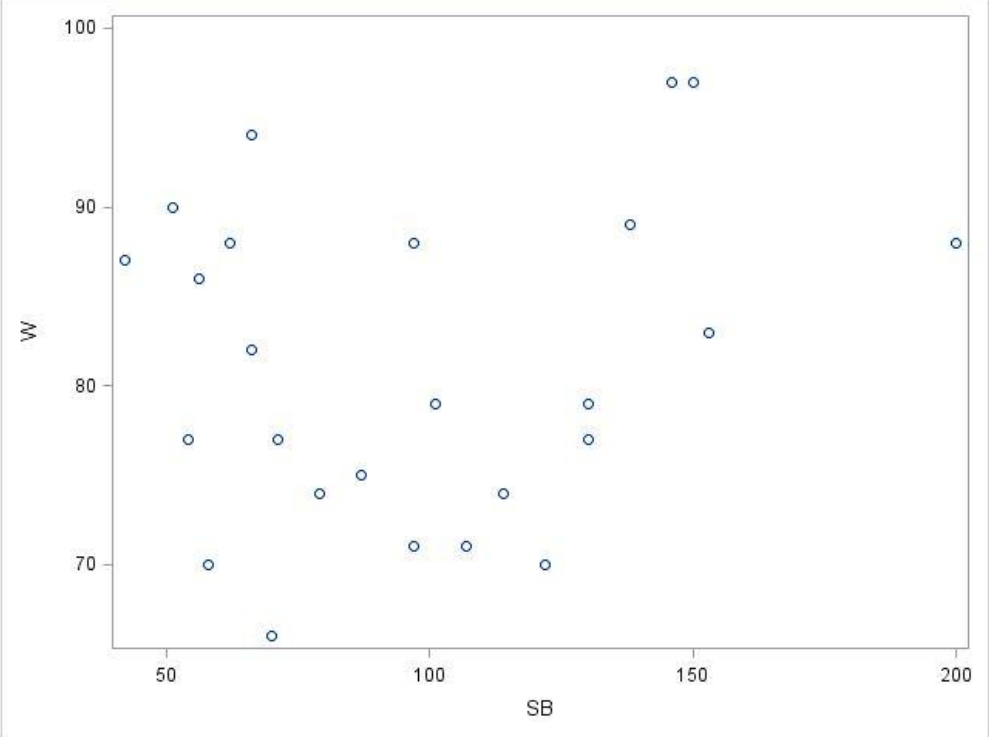

Table 9 - Scatter Plot of Wins vs Walks

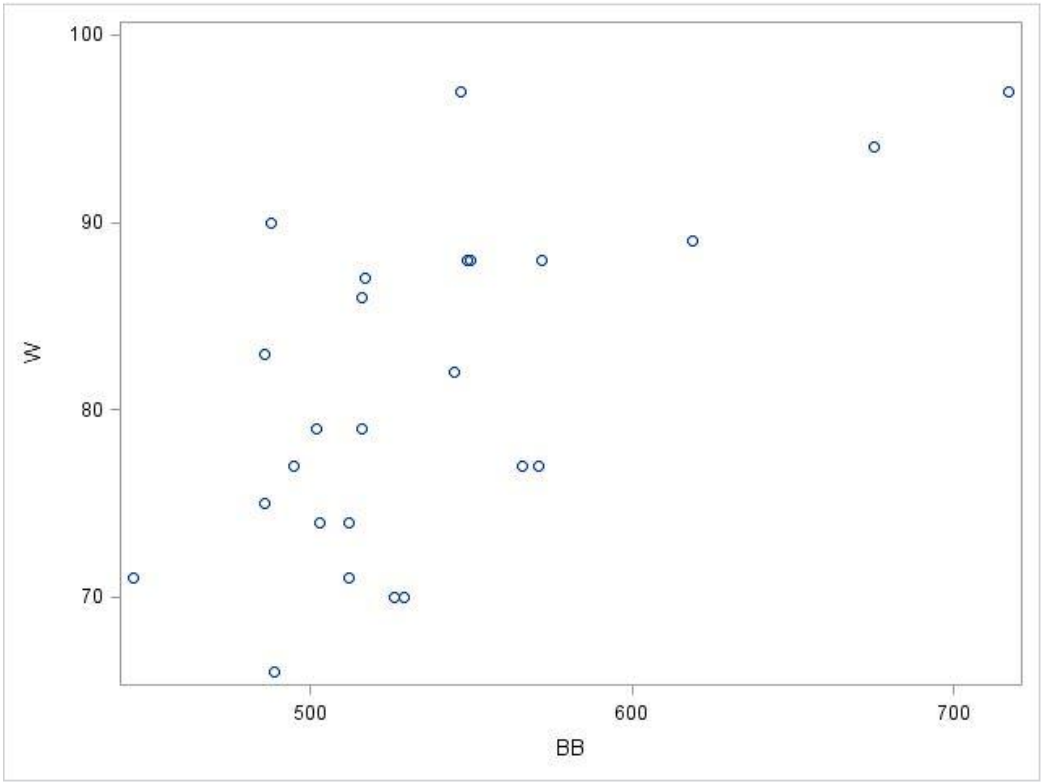

Table 10 - Scatter Plot of Wins vs Strikeouts

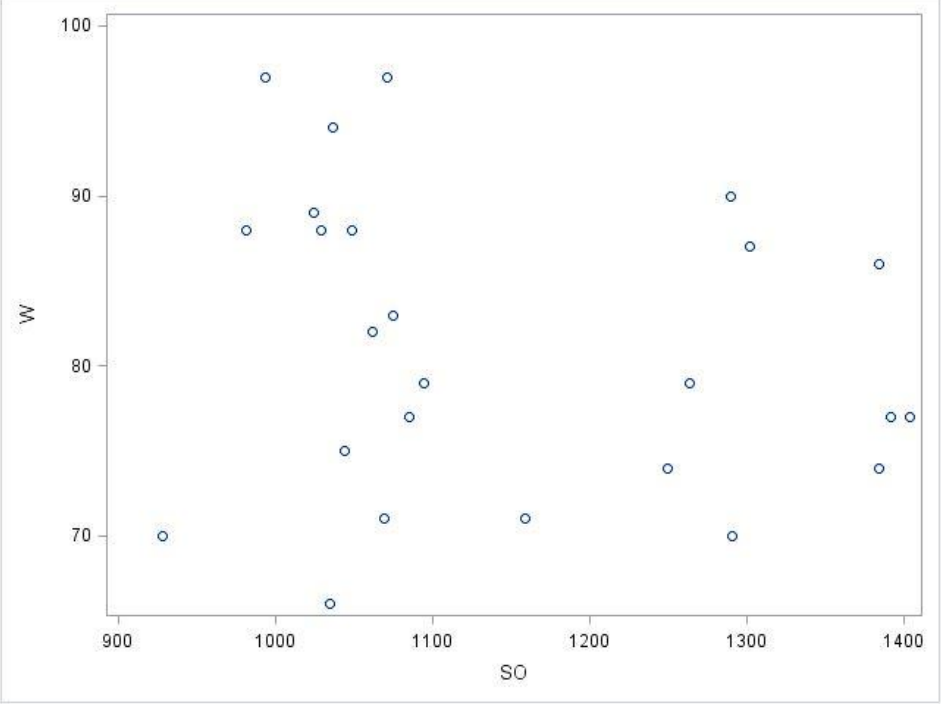

Table 11 - Scatter Plot of Wins vs Batting Average

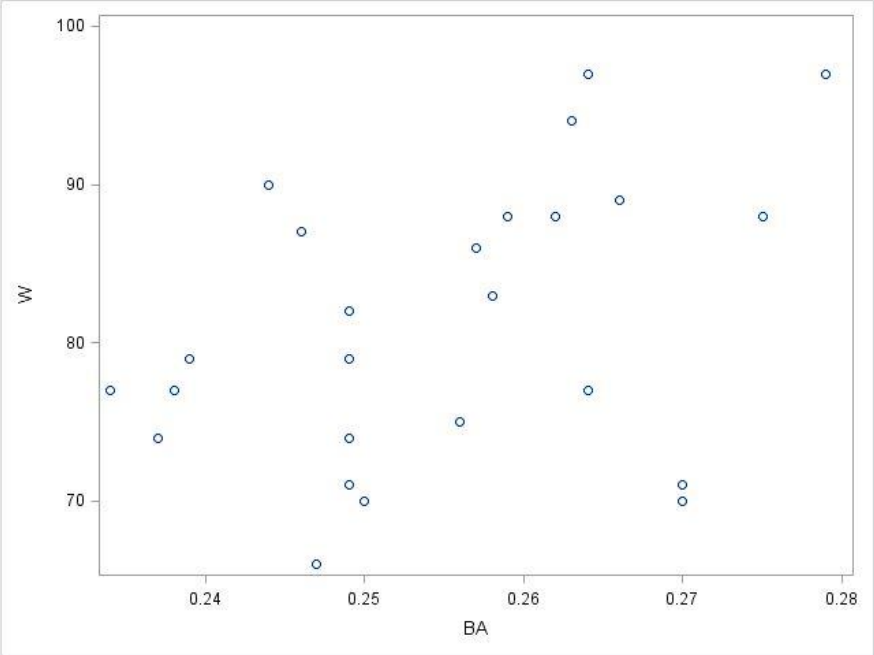

Table 12 - Scatter Plot of Wins vs Age of Batters

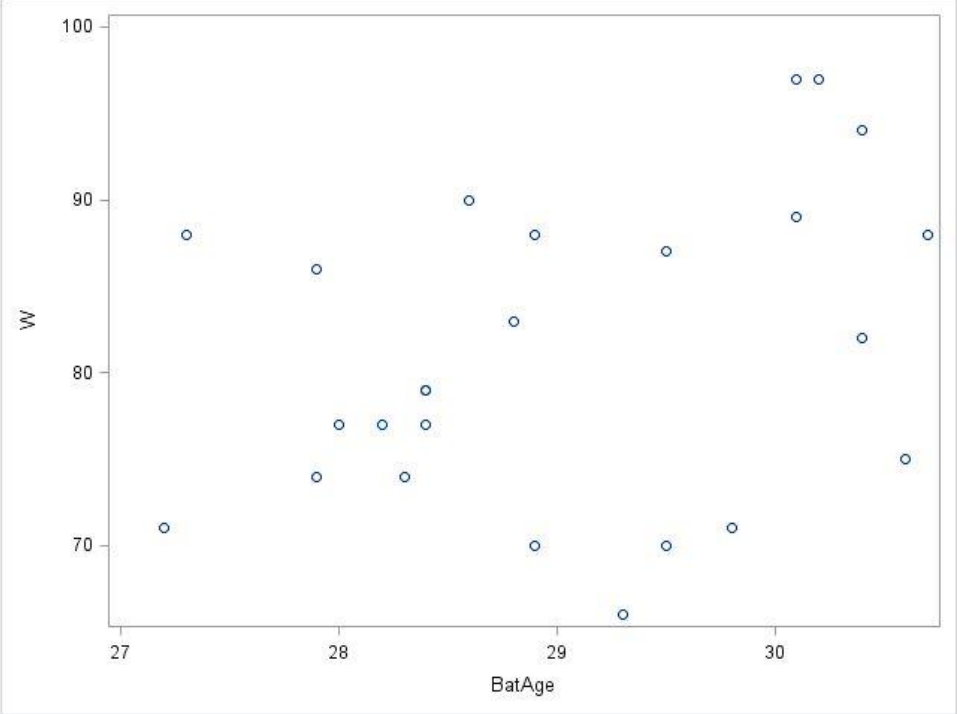

Table 13 - Correlation Matrix

Pearson Correlation Coefficients, N = 25 Prob > |r| under H0: Rho=0												
	W	R	H	_B	_B0	HR	RBI	SB	BB	SO	BA	BatAge
W	1.00000	0.68073 0.0002	0.42643 0.0335	0.33609 0.1005	-0.25139 0.2254	0.43712 0.0289	0.69049 0.0001	0.20443 0.3270	0.62224 0.0009	-0.19120 0.3599	0.39684 0.0495	0.33060 0.1065
R	0.68073 0.0002	1.00000	0.81682 <.0001	0.49888 0.0111	0.08285 0.6938	0.50110 0.0107	0.99422 <.0001	0.42158 0.0358	0.60335 0.0014	-0.42108 0.0361	0.78334 <.0001	0.29023 0.1593
H	0.42643 0.0335	0.81682 <.0001	1.00000	0.56498 0.0033	0.32178 0.1167	0.07337 0.7274	0.77714 <.0001	0.60803 0.0013	0.43803 0.0285	-0.70545 <.0001	0.97387 <.0001	0.27044 0.1910
_B	0.33609 0.1005	0.49888 0.0111	0.56498 0.0033	1.00000	0.24327 0.2413	-0.03521 0.8673	0.50670 0.0097	0.45403 0.0226	0.35228 0.0842	-0.36077 0.0764	0.50459 0.0101	0.12594 0.5486
_B0	-0.25139 0.2254	0.08285 0.6938	0.32178 0.1167	0.24327 0.2413	1.00000	-0.41114 0.0412	0.03654 0.8623	0.44973 0.0241	-0.17560 0.4011	-0.32707 0.1105	0.30404 0.1395	-0.18936 0.3646
HR	0.43712 0.0289	0.50110 0.0107	0.07337 0.7274	-0.03521 0.8673	-0.41114 0.0412	1.00000	0.56191 0.0035	-0.22850 0.2719	0.16537 0.4295	0.35267 0.0838	0.01792 0.9322	0.21720 0.2970
RBI	0.69049 0.0001	0.99422 <.0001	0.77714 <.0001	0.50670 0.0097	0.03654 0.8623	0.56191 0.0035	1.00000	0.38399 0.0581	0.59595 0.0017	-0.34911 0.0872	0.73544 <.0001	0.27212 0.1882
SB	0.20443 0.3270	0.42158 0.0358	0.60803 0.0013	0.45403 0.0226	0.44973 0.0241	-0.22850 0.2719	0.38399 0.0581	1.00000	0.22211 0.2859	-0.54605 0.0047	0.58599 0.0021	0.27698 0.1801
BB	0.62224 0.0009	0.60335 0.0014	0.43803 0.0285	0.35228 0.0842	-0.17560 0.4011	0.16537 0.4295	0.59595 0.0017	0.22211 0.2859	1.00000	-0.32003 0.1189	0.44929 0.0243	0.43358 0.0304
SO	-0.19120 0.3599	-0.42108 0.0361	-0.70545 <.0001	-0.36077 0.0764	-0.32707 0.1105	0.35267 0.0838	-0.34911 0.0872	-0.54605 0.0047	-0.32003 0.1189	1.00000	-0.78533 <.0001	-0.46564 0.0190
BA	0.39684 0.0495	0.78334 <.0001	0.97387 <.0001	0.50459 0.0101	0.30404 0.1395	0.01792 0.9322	0.73544 <.0001	0.58599 0.0021	0.44929 0.0243	-0.78533 <.0001	1.00000	0.31123 0.1299
BatAge	0.33060 0.1065	0.29023 0.1593	0.27044 0.1910	0.12594 0.5486	-0.18936 0.3646	0.21720 0.2970	0.27212 0.1882	0.27698 0.1801	0.43358 0.0304	-0.46564 0.0190	0.31123 0.1299	1.00000

Table 14 - Proc Reg: Runs, Hits, Home Runs

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	979.56942	326.52314	7.49	0.0014
Error	21	915.79058	43.60908		
Corrected Total	24	1895.36000			

Root MSE	6.60372	R-Square	0.5168
Dependent Mean	81.16000	Adj R-Sq	0.4478
Coeff Var	8.13666		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	55.89035	30.18421	1.85	0.0782
R	R	1	0.14151	0.05215	2.71	0.0130
H	H	1	-0.05152	0.03853	-1.34	0.1955
HR	HR	1	-0.02066	0.05760	-0.36	0.7234

Table 15 - Proc Reg: Runs, Hits, Runs Batted In

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	973.96065	324.65355	7.40	0.0014
Error	21	921.39935	43.87616		
Corrected Total	24	1895.36000			

Root MSE	6.62391	R-Square	0.5139
Dependent Mean	81.16000	Adj R-Sq	0.4444
Coeff Var	8.16154		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	49.59934	29.13251	1.70	0.1034
R	R	1	0.12576	0.23842	0.53	0.6034
H	H	1	-0.04210	0.03463	-1.22	0.2376
RBI	RBI	1	0.00137	0.22992	0.01	0.9953

Table 16 - Proc Reg: Runs, Hits, Walks

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1083.30086	361.10029	9.34	0.0004
Error	21	812.05914	38.66948		
Corrected Total	24	1895.36000			

Root MSE	6.21848	R-Square	0.5716
Dependent Mean	81.16000	Adj R-Sq	0.5103
Coeff Var	7.66200		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	38.37855	24.25249	1.58	0.1285
R	R	1	0.09867	0.03579	2.76	0.0118
H	H	1	-0.03680	0.02703	-1.36	0.1878
BB	BB	1	0.04475	0.02661	1.68	0.1075

Table 17 - Proc Reg: Runs, Home Runs, Runs Batted In

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	911.10436	303.70145	6.48	0.0028
Error	21	984.25564	46.86932		
Corrected Total	24	1895.36000			

Root MSE	6.84612	R-Square	0.4807
Dependent Mean	81.16000	Adj R-Sq	0.4065
Coeff Var	8.43533		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	18.98403	14.58645	1.30	0.2072
R	R	1	-0.03117	0.24481	-0.13	0.8999
HR	HR	1	0.01245	0.06068	0.21	0.8394
RBI	RBI	1	0.12123	0.26943	0.45	0.6574

Table 18 - Proc Reg: Runs, Home Runs, Walks

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1064.42420	354.80807	8.97	0.0005
Error	21	830.93580	39.56837		
Corrected Total	24	1895.36000			

Root MSE	6.29034	R-Square	0.5616
Dependent Mean	81.16000	Adj R-Sq	0.4990
Coeff Var	7.75054		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	11.28459	13.91163	0.81	0.4264
R	R	1	0.04531	0.02682	1.69	0.1060
HR	HR	1	0.04784	0.04141	1.16	0.2610
BB	BB	1	0.05532	0.02727	2.03	0.0554

Table 19 - Proc Reg: Runs, Runs Batted In, Walks

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1048.68150	349.56050	8.67	0.0006
Error	21	846.67850	40.31802		
Corrected Total	24	1895.36000			

Root MSE	6.34965	R-Square	0.5533
Dependent Mean	81.16000	Adj R-Sq	0.4895
Coeff Var	7.82362		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	10.08554	14.08903	0.72	0.4820
R	R	1	-0.10450	0.17432	-0.60	0.5553
RBI	RBI	1	0.17459	0.18212	0.96	0.3486
BB	BB	1	0.05025	0.02701	1.86	0.0769

Table 20 - Proc Reg: Hits, Home Runs, Runs Batted In

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	976.32110	325.44037	7.44	0.0014
Error	21	919.03890	43.76376		
Corrected Total	24	1895.36000			

Root MSE	6.61542	R-Square	0.5151
Dependent Mean	81.16000	Adj R-Sq	0.4458
Coeff Var	8.15108		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	50.35033	29.05077	1.73	0.0977
H	H	1	-0.04487	0.03654	-1.23	0.2331
HR	HR	1	-0.03597	0.06235	-0.58	0.5701
RBI	RBI	1	0.14670	0.05444	2.69	0.0136

Table 21 - Proc Reg: Hits, Home Runs, Walks

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1006.80439	335.60146	7.93	0.0010
Error	21	888.55561	42.31217		
Corrected Total	24	1895.36000			

Root MSE	6.50478	R-Square	0.5312
Dependent Mean	81.16000	Adj R-Sq	0.4642
Coeff Var	8.01476		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.36882	23.22492	0.02	0.9875
H	H	1	0.02060	0.01802	1.14	0.2658
HR	HR	1	0.08349	0.03683	2.27	0.0341
BB	BB	1	0.07114	0.02479	2.87	0.0092

Table 22 - Hits, Runs Batted In, Walks

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1084.08111	361.36037	9.35	0.0004
Error	21	811.27889	38.63233		
Corrected Total	24	1895.36000			

Root MSE	6.21549	R-Square	0.5720
Dependent Mean	81.16000	Adj R-Sq	0.5108
Coeff Var	7.65832		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	31.41678	23.25761	1.35	0.1911
H	H	1	-0.02798	0.02462	-1.14	0.2686
RBI	RBI	1	0.09406	0.03406	2.76	0.0117
BB	BB	1	0.04673	0.02626	1.78	0.0896

Table 23 - Home Runs, Runs Batted In, Walks

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1066.56178	355.52059	9.01	0.0005
Error	21	828.79822	39.46658		
Corrected Total	24	1895.36000			

Root MSE	6.28224	R-Square	0.5627
Dependent Mean	81.16000	Adj R-Sq	0.5003
Coeff Var	7.74057		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	10.92357	13.96639	0.78	0.4429
HR	HR	1	0.03972	0.04386	0.91	0.3754
RBI	RBI	1	0.05067	0.02968	1.71	0.1025
BB	BB	1	0.05455	0.02742	1.99	0.0598

Table 24 - Models

Models	F-Test Statistics	R^2 Adjusted	CV	T-Test Statistics
Wins = B0 + B1(Runs) + B2(hits) + B3(Homeruns)	F-Value: 7.49 P-Value: 0.0014	0.4478	8.13666	**Runs:** T-value: 2.71; P-Value: 0.0130 **Hits:** T-Value: -1.34; P-Value: 0.1955 **Home Runs:** T-Value: -0.36; P-Value: 0.7234
Wins = B0 + B1(Runs) + B2(hits) + B3(Runs Batted In)	F-Value: 7.40 P-Value: 0.0014	0.4444	8.16154	**Runs:** T-value: 0.53; P-Value: 0.6034 **Hits:** T-value: -1.22; P-Value: 0.2376 **RBI:** T-value: 0.01; P-Value: 0.9953
Wins = B0 + B1(Runs) + B2(hits) + B3(Walks)	F-Value: 9.34 P-Value: 0.0004	0.5103	7.66200	**Runs:** T-value: 2.76; P-Value: 0.0118 **Hits:** T-value: -1.36; P-Value: 0.1878 **Walks:** T-value: 1.68; P-Value: 0.1075
Wins = B0 + B1(Runs) + B2(Homeruns) + B3(Runs Batted In)	F-Value: 6.48 P-Value: 0.0028	0.4065	8.43533	**Runs:** T-value: -0.13; P-Value: 0.8999 **Home Runs:** T-value: 0.21; P-Value: 0.8394 **RBI:** T-value: 0.45; P-Value: 0.6574
Wins = B0 + B1(Runs) + B2(Homeruns) + B3(Walks)	F-Value: 8.97 P-Value: 0.0005	0.4990	7.75054	**Runs:** T-value: 1.69; P-Value: 0.1060 **Home Runs:** T-value: 1.16; P-Value: 0.2610 **Walks:** T-value: 2.03; P-Value: 0.0554
Wins = B0 + B1(Runs) + B2(Runs Batted In) + B3(Walks)	F-Value: 8.67 P-Value: 0.0006	0.4895	7.82362	**Runs:** T-value: -0.60; P-Value: 0.5553 **RBI:** T-value: 0.96; P-Value: 0.3486 **Walks:** T-value: 1.86; P-Value: 0.0769
Wins = B0 + B1(Hits) + B2(Homeruns) + B3(Runs Batted In)	F-Value: 7.44 P-Value: 0.0014	0.4458	8.15108	**Hits:** T-value: -1.23; P-Value: 0.2331 **Home Runs:** T-value: -0.58; P-Value: 0.5701 **RBI:** T-value: 2.69; P-Value: 0.0136
Wins = B0 + B1(Hits) + B2(Homeruns) + B3(Walks)	F-Value: 7.93 P-Value: 0.0010	0.4642	8.01476	**Hits:** T-value: 1.14; P-Value: 0.2658 **Home Runs:** T-value: 2.27; P-Value: 0.0341 **Walks:** T-value: 2.87; P-Value: 0.0092
Wins = B0 + B1(Hits) + B2(Runs Batted In) + B3(Walks)	F-Value: 9.35 P-Value: 0.0004	0.5108	7.65832	**Hits:** T-value: -1.14; P-Value: 0.2686 **RBI:** T-value: 2.76; P-Value: 0.0117 **Walks:** T-value: 1.78; P-Value: 0.0896
Wins = Bo + B1(Homeruns) + B2(Runs Batted In) + B3(Walks)	F-Value: 9.01 P-Value: 0.0005	0.5003	7.74057	**Home Runs:** T-value: 0.91; P-Value: 0.3754 **RBI:** T-value: 1.71; P-Value: 0.1025 **Walks:** T-value: 1.99; P-Value: 0.0598

Table 25 - Residual vs Predicted Value (Wins)

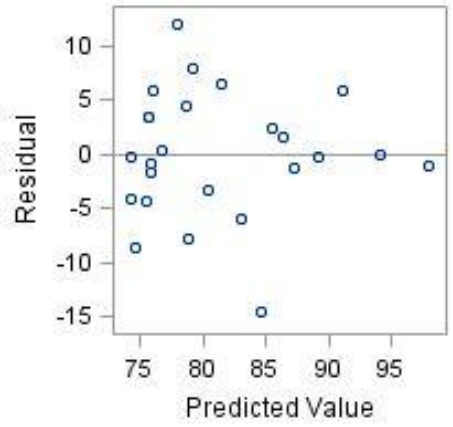

Table 26 - Normal Probability Plot

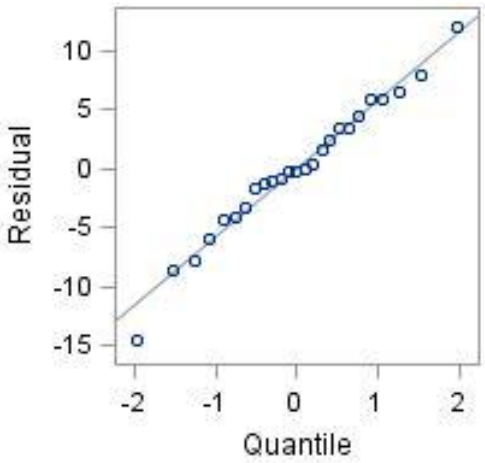

Table 27 - Descriptive Analysis

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
W	W	25	81.1600000	8.8866942	66.0000000	97.0000000
R	R	25	713.6400000	69.8032234	619.0000000	853.0000000
H	H	25	1404.20	81.9827218	1242.00	1553.00
HR	HR	25	163.4800000	36.5560301	95.0000000	242.0000000
_B	2B	25	277.1200000	21.4773524	228.0000000	323.0000000
_B0	3B	25	27.5200000	10.0917458	14.0000000	49.0000000
RBI	RBI	25	679.6000000	66.3525935	593.0000000	814.0000000
SB	SB	25	97.8800000	40.3074848	42.0000000	200.0000000
BB	BB	25	537.3200000	60.2416799	445.0000000	717.0000000
SO	SO	25	1147.92	148.8421759	928.0000000	1404.00
BA	BA	25	0.2549600	0.0122423	0.2340000	0.2790000
BatAge	BatAge	25	29.0320000	1.0490948	27.2000000	30.7000000

Table 28 - Parameter Estimates for Interaction and Quadratic Terms

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-134.71455	442.77542	-0.30	0.7651
H	H	1	0.62047	0.80587	0.77	0.4533
RBI	RBI	1	-0.73126	1.20249	-0.61	0.5522
BB	BB	1	-0.00360	0.86113	-0.00	0.9967
hits_RBI		1	0.00145	0.00108	1.34	0.1999
hits_Walks		1	0.00043213	0.00086126	0.50	0.6231
RBI_Walks		1	-0.00160	0.00197	-0.81	0.4293
hits_sq		1	-0.00065778	0.00042984	-1.53	0.1468
RBI_sq		1	-0.00027755	0.00078623	-0.35	0.7290
walks_sq		1	0.00051176	0.00089064	0.57	0.5741

SAS Code

```
proc contents data=sport; run;

proc print data=sport; run;

proc sgplot data=sport;
 scatter y=W x=R;
 run;
 proc sgplot data=sport;
 scatter y=W x=H;
 run;
proc sgplot data=sport;
 scatter y=W x=_B;
 run;
proc sgplot data=sport;
 scatter y=W x=_B0;
 run;
proc sgplot data=sport;
 scatter y=W x=HR;
 run;
proc sgplot data=sport;
 scatter y=W x=RBI;
 run;
proc sgplot data=sport;
 scatter y=W x=SB;
 run;
proc sgplot data=sport;
 scatter y=W x=BB;
 run;
proc sgplot data=sport;
 scatter y=W x=SO;
 run;
proc sgplot data=sport;
 scatter y=W x=BA;
 run;
proc sgplot data=sport;
 scatter y=W x=BatAge;
 run;

proc corr data=sport;
 var W R H _B _B0 hr rbi sb bb so ba batage;
 run;

proc means data=sport;
 var W R H HR _B _B0 rbi sb bb so ba batage;
run;

proc reg data=sport;
 model W = R H HR;
```

```
 run;
proc reg data=sport;
 model W = R H RBI;
 run;
proc reg data=sport;
 model W = R H BB;
 run;
proc reg data=sport;
 model W = R HR RBI;
 run;
proc reg data=sport;
 model W = R HR BB;
 run;
proc reg data=sport;
 model W = R RBI BB;
 run;
proc reg data=sport;
 model W = H HR RBI;
 run;
proc reg data=sport;
 model W = H HR BB;
 run;
proc reg data=sport;
 model W = H RBI BB;
 run;
proc reg data=sport;
 model W = HR RBI BB;
 run;

data sport2;
set sport;
hits_RBI = h*RBI;
hits_Walks = H*BB;
RBI_Walks = RBI*BB;
hits_sq = h**2;
RBI_sq = RBI**2;
walks_sq = BB**2;

proc reg data=sport2;
 model w = h RBI BB hits_rbi hits_walks RBI_walks hits_sq RBI_sq walks_sq;
 run;
```