Modeling Team Wins for the Los Angeles Angels (2000–2025)

Thomas Wright
10/25/2025

Executive Summary

Offensive performance is an important driver of success in professional baseball, which makes it a natural candidate for modeling of team outcomes. This analysis investigates which factors in batting performance best predict the total number of wins for the Los Angeles Angels between 2000 and 2025 seasons. Data were collected from Baseball Reference and was modeled in SAS using multiple regression with the dependent variable of interest team wins per season. The four predictor variables that were selected based on their high correlation with wins were hits, doubles, batting average and runs.

First a correlation analysis was performed on all variables in the dataset to identify the four batting metrics that were most strongly associated with total wins. Next, a scatterplot and histogram were used to visualize the distribution and nature of the association with wins for each of the predictors and to spot any potential outliers. Multiple regression analysis was conducted to test all possible combinations of two variable models, and the best model was selected for further analysis. After selection the remaining two predictors were added to a three variable model and each model was assessed again. And finally, after selecting the best overall model an interaction model and quadratic model were fit to test for any possible interaction or non-linear effects between the predictors.

The two variable model using hits and batting average produced the best fit overall with an adjusted r^2 of 0.54 and a root MSE of 6.9, meaning the model predicts total season wins within approximately $\pm$ 7 games. Adding a third predictor or including higher order terms all increased the variability of the model and reduced the overall explanatory power when compared to the two variable model selected. The results show that offensive performance explains a moderate portion of overall team wins per season but other factors not related to batting performance also play a substantial role in total wins.

Exploratory Data Analysis

The dataset used in this analysis was obtained from Baseball Reference, which archives year-by-year team batting statistics. Data were collected for the Los Angeles Angels for all seasons from 2000 to 2025. The 2020 season was excluded from the dataset due to the COVID-19 pandemic, which resulted in a shortened 60 game season and is not representative

of overall batting performance. After cleaning there are a total of 25 observations for all variables representing the 25 full seasons of play.

The first exploratory step was to conduct a correlation analysis on all the offensive variables in the dataset with the outcome variable wins. The results of this analysis lead to the selection of four potential predictors for regression modeling which were hits, batting average, doubles, and runs.

Figures A1-A4 in Appendix A display histograms and scatterplots for each selected predictor in relation to team wins. Each variable shows a generally linear positive association with wins meaning seasons with higher offensive output in hits, batting average, doubles, and runs all tended to result in more wins across the 25 seasons analyzed. Hits and batting average show a much clearer linear association when compared to doubles and runs. Both doubles and runs showed greater variability near the middle, meaning that when doubles and runs were about average, this resulted in seasons with both unusually high and unusually low number of total wins. For example, the 2008 season was the best season during the observed period with 100 total wins, but the team only tallied 274 doubles, which is exactly the median number of doubles. Whereas in the 2019 season the team hit 268 doubles but only managed a total of 72 wins which is in the bottom quartile for the 25 season period (See Table A4).

All four variables show roughly symmetric distributions without any extreme outliers. Hits (Figure 1) and batting average (Figure A2) shows mild left-skewness, reflecting that the team more frequently had high offensive performance than unusually low performance. Doubles (Figure 2) is the most symmetric of the four, centered near 270, meaning doubles remained consistent across all seasons. Runs (Figure A4) shows mild right-skewness, meaning fewer very high scoring seasons compared to lower scoring seasons. The distributions show that all of the predictors are approximately normal and are acceptable for regression modeling.

Figure 1: Relationship and distribution of Hits as a predictor of team Wins

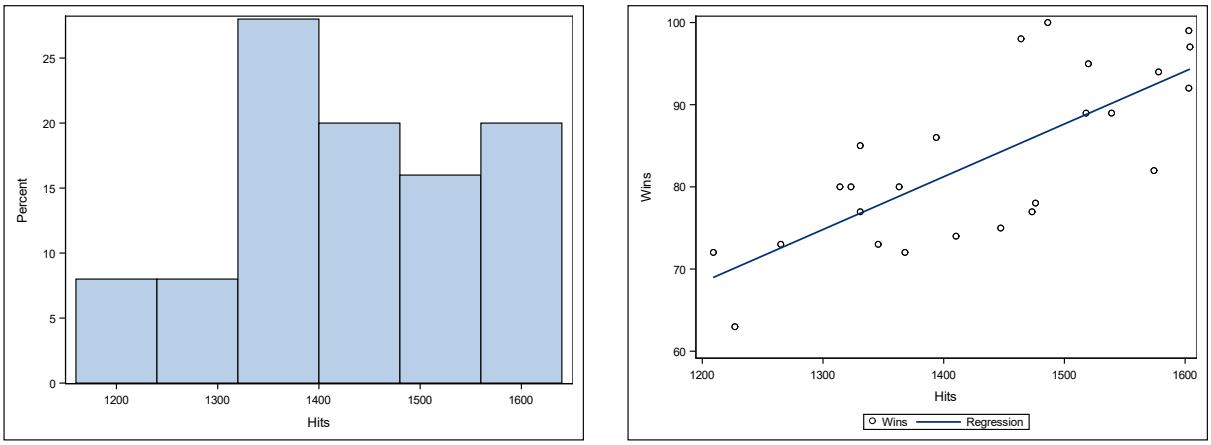
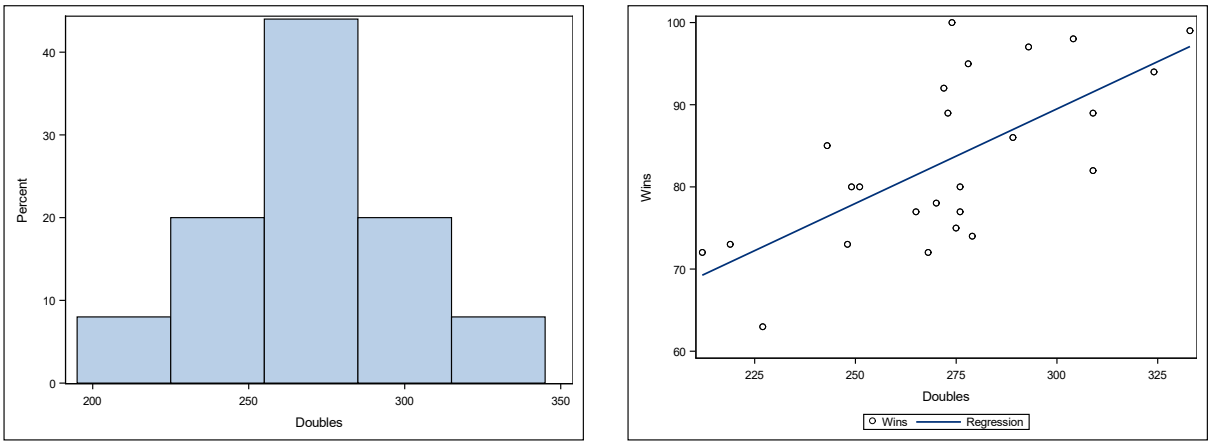


Figure 2: Relationship and distribution of Doubles as a predictor of team Wins



Methodology

All analyses were conducted using SAS. The first step was a correlation analysis using PROC CORR to identify which batting metrics were most strongly associated with total team wins. After the correlation analysis, descriptive statistics for the selected predictors were created using PROC MEANS, and the histograms and scatterplots were created using PROC SGPLOT to visualize variable distributions and relationships. The visualizations were used to assess linearity of the predictors in relation to the outcome variable, visualize the spread of the variables and detect any potential outliers.

Multiple regression modeling was then conducted using PROC REG where all combinations of two variable models were compared. After selecting the best two variable model this was followed by the addition of a third predictor variable to evaluate if there could be any incremental improvements made to the model. One interaction term and one quadratic term were also tested to identify any potential nonlinear or combined effects. Model comparison and selection was evaluated using adjusted r^2, Root Mean Square Error, Coefficient of Variation, and the overall F-statistic. For the final selected model, coefficient estimates along with their t-values and p-values were used to interpret the strength and significance of each predictor to explaining variation in team wins.

Results

Correlation Analysis

The full correlation matrix for all team batting statistics is presented in Appendix A (Tables A1 and A2). Table 1 below summarizes the correlations for the four selected predictors and the dependent variable, wins. Each of the predictors showed a strong positive correlation to wins with hits ($r = 0.756$), batting average ($r = 0.736$), doubles ($r = 0.680$), and runs ($r = 0.659$). This makes clear that increased batting performance tends to result in more

wins. Something of note however, was that all the predictors were also highly correlated with one another, with the most extreme one being hits and batting average ($r = 0.993$). This near-perfect correlation makes sense since batting average is calculated as hits divided by at-bats, meaning the two variables share essentially the same information about team performance.

Table 1: Correlation Matrix for Selected Predictors and Wins

	Wins	**Hits**	**Doubles**	**BatAvg**	**Runs**
Wins	1.000	0.756	0.680	0.736	0.659
Hits	0.756	1.000	0.884	0.993	0.929
Doubles	0.680	0.884	1.000	0.871	0.896
BatAvg	0.736	0.993	0.871	1.000	0.933
Runs	0.659	0.929	0.896	0.933	1.000

Two Variable Model Comparison

Multiple regression was conducted on all possible two predictor combinations of the four selected variables from the correlation analysis: hits, batting average, doubles, and runs. The goal was to determine which pair of predictors best explained variation in total team wins. Table 2 summarizes the relevant statistics for each two variable model, which included the Root Mean Square Error, Adjusted r^2, Coefficient of Variation, F-statistic, and the overall model p-value. All models were highly statistically significant with $p \leq 0.003$, meaning that combinations of batting statistics explain a portion of the variation in team wins.

Of the six combinations tested, the hits + batting average model performed the best on all relevant statistics and had the lowest root MSE (6.84), the highest adjusted r^2 (0.55), and the lowest coefficient of variation (8.22%). This means the model accounts for the most variation in team wins with the least prediction error, making it the best overall fit. Other models, such as hits + doubles and hits + runs, also performed similarly well but were marginally worse across the board.

Table 2: Two Variable Regression Models in Predicting Total Wins

Predictors	**Root MSE**	**Adj r^2**	**Coeff Var**	**F**	$p_{\textbf{Model}}$
Hits + BatAvg	6.838	0.5525	8.22	15.82	< 0.0001
Hits + Doubles	6.940	0.5391	8.34	15.03	< 0.0001
Hits + Runs	6.983	0.5333	8.39	14.71	< 0.0001
BatAvg + Doubles	7.129	0.5137	8.57	13.68	0.0001
BatAvg + Runs	7.196	0.5044	8.65	13.22	0.0002
Doubles + Runs	7.418	0.4735	8.92	11.79	0.0003

Three Variable Model Comparison

Next in the analysis was to test if the best two variable model could be improved by adding a third predictor using the two remaining variables, doubles or runs. The results summarized in Appendix Table A6, show that neither model were able to provide a marginal improvement over the best two variable model. Both three variable models were statistically significant, but just as with all previous two variable models, they were both marginally worse across the board in both explanatory power in terms of a lower adjusted r^2, around 0.53, and in variability with higher root MSE around 7 and coefficient of variation of 8.4. This means the two variable model that included hits and batting average already explain most of the variation in total wins and adding doubles or runs were mostly redundant and only increased model complexity.

Interaction and Quadratic Models

Finally, two higher-order models were fit to test for any potential nonlinear and interaction effects between the predictors. The first model included an interaction term between hits and batting average and the second included a quadratic term for hits. The variable hits was selected for the quadratic term instead of batting average because batting average is already a scaled variable rather than a pure count. Since batting average is bounded between 0 and 1, squaring this term would only further compress the scale rather than reveal any non-linear effect. While hits on the other hand is on a much larger scale and would be more likely to show a non-linear effect and is easier to interpret. Just as the previous three variable models, adding either an interaction term or a quadratic term was unable to provide a marginal improvement over the best two variable model and again both performed marginally worse across the board compared to the two variable model (Appendix Table A7).

Final Model Evaluation

The best fitting model selected from the regression analysis was the two variable model using hits and batting average as predictors of total team wins. The regression equation is:

$$\widehat{W} = 22.25 + 0.1609(\text{Hits}) - 653.96(\text{BatAvg})$$

Interestingly, while the overall model was highly statistically significant ($F = 15.82$, $p < 0.0001$), neither hits ($p = 0.12$) nor batting average ($p = 0.34$) was significant on its own. This implies that the two variables largely capture the same aspect of team performance, which is reflected in their extremely high correlation ($r = 0.993$).

The intercept ($\beta_0 = 22.2465$) represents the predicted number of wins for a team with zero hits and a batting average of zero, which is not a useful interpretation in practice since it's not possible to win a game and never hit the ball. The coefficient for hits ($\beta_1 = 0.1609$) indicates that, holding batting average constant, each additional team hit per season is associated with an increase of about 0.16 wins. However, this should be interpreted with caution since hits individually was not statistically significant and shares enormous overlap with batting average. The negative coefficient for batting average ($\beta_2 = -653.96$) contradicts the positive effect for hits. This implies that lower batting average increases wins while more

hits also increases wins, which is nonsensical because batting average is derived from hits (hits/at-bats). Furthermore, batting average was originally selected because it was strongly positively correlated with wins, again the model contradicts previous analysis. Therefore, the batting average coefficient should not be interpreted in this model.

The model's adjusted r^2 of 0.55 means that roughly 55% of the variation in total wins across seasons is explained by these two batting statistics. The Root MSE (6.84) shows that the model's predictions differ from actual wins by about ± 7 games on average and the coefficient of variation was 8.22%. The overall model appears to be adequate for predicting team wins since more than half of the variation is accounted for in the model and the coefficient of variation is less than 10%, meaning that the error in predicting total team wins of ± 7 is reasonable.

Overall, this model provides a good fit for the data and captures much of the relationship between team batting performance and total wins, even though the individual predictors were not statistically significant.

Table 3: Regression Output for Final Model: Wins = Hits + Batting Average

Variable	**Estimate (β)**	**Std. Error**	**t Value**	**p Value**
Intercept	22.2465	35.8589	0.62	0.5414
Hits	0.1609	0.0999	1.61	0.1217
BatAvg	-653.9606	671.6898	-0.97	0.3408
Model Fit Statistics				
Root MSE		6.8383		
Adj r^2		0.5525		
Coeff Var		8.22		
F Value		15.82		
Model p Value		< 0.0001		

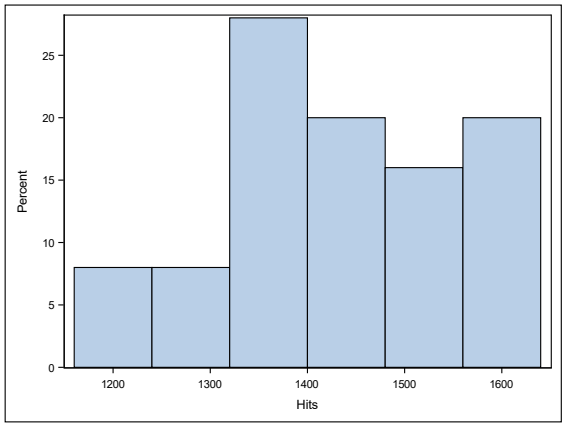
Conclusion

The regression analysis showed that hits and batting average, when taken together, were good predictors of team wins, explaining over half of the variation in total wins across the 25 seasons analyzed. The overall model was statistically significant, but interpretation of the individual predictors is limited since neither coefficient was significant on its own.

For future analysis, a simpler model using only one of the two predictors, such as a simple linear regression with hits could provide better insight into exactly how hits translates into wins. It could also be interesting to include offensive statistics that are not directly the result of hitting the ball, such as stolen bases, to see if that could explain additional variation in total wins. Other available data such as pitching statistics could also provide a more complete picture of overall team performance. Lastly, using the full historical dataset rather than restricting the analysis to the past 25 seasons, could improve the reliability of the model's predictions.

Appendix A: Tables and Figures

Figure A1: Relationship and distribution of Hits as a predictor of team Wins



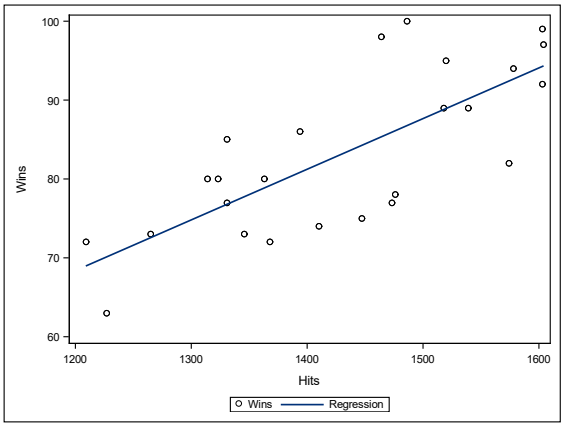
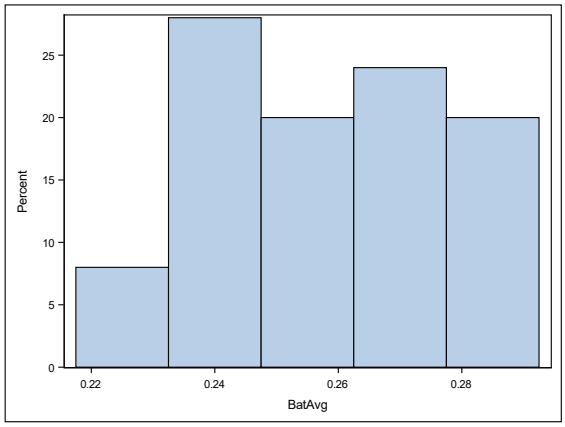


Figure A2: Relationship and distribution of BatAvg as a predictor of team Wins



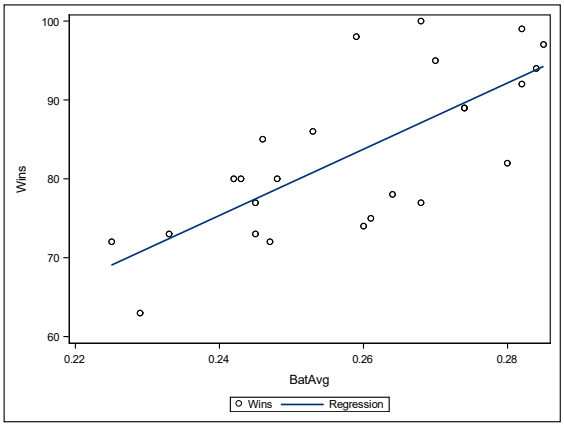
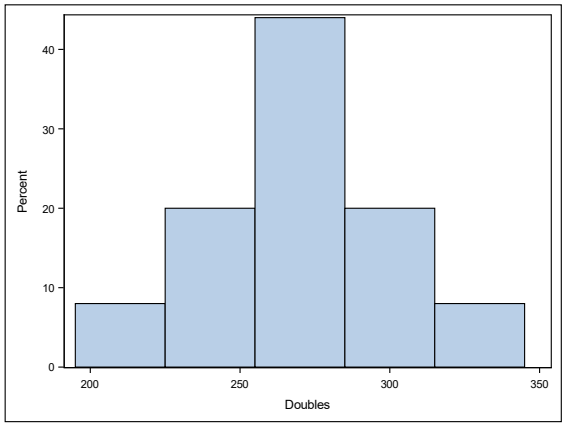


Figure A3: Relationship and distribution of Doubles as a predictor of team Wins



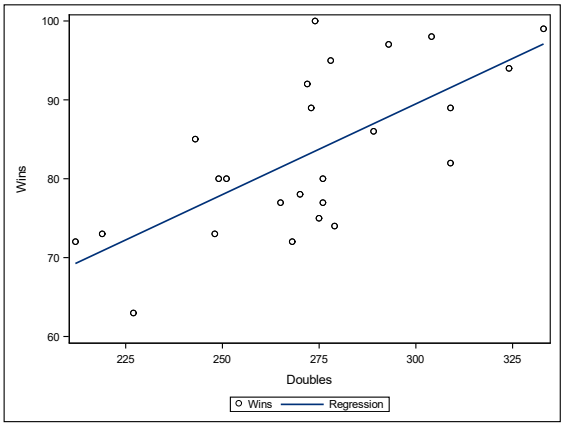
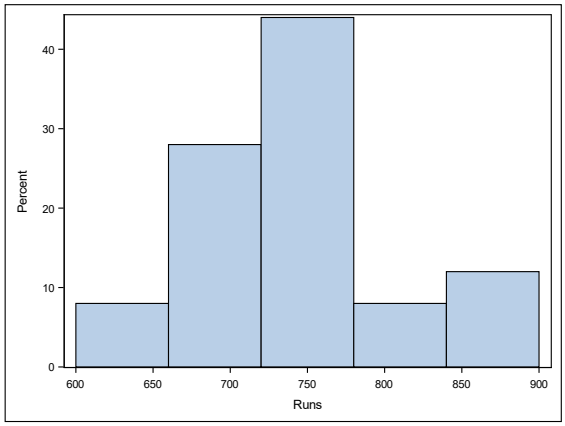


Figure A4: Relationship and distribution of Runs as a predictor of team Wins



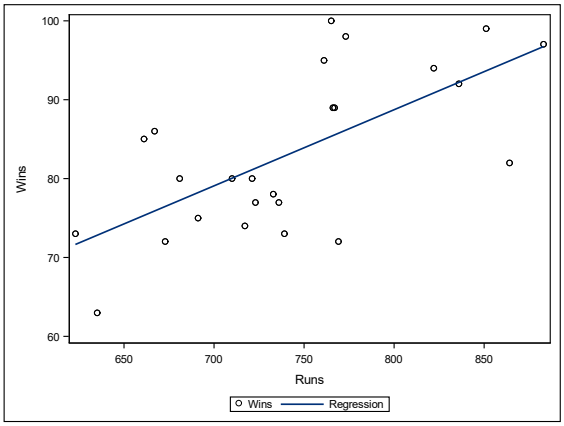


Table A1: Correlation Matrix for Wins and Offensive Variables (Part 1)

Pearson Correlation Coefficients, N = 25												
	W	R/G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS
W	1.00000	0.65942	0.55341	0.73128	0.65945	0.75639	0.68014	0.42370	-0.46353	0.62878	0.46670	0.45306
R/G	0.65942	1.00000	0.85508	0.80889	0.99998	0.86631	0.73765	0.46499	-0.05619	0.99438	0.33302	0.37143
PA	0.55341	0.85508	1.00000	0.93071	0.85416	0.82729	0.74270	0.63469	-0.07972	0.85425	0.24842	0.38269
AB	0.73128	0.80889	0.93071	1.00000	0.80761	0.89175	0.77981	0.72651	-0.27631	0.79696	0.35967	0.43311
R	0.65945	0.99998	0.85416	0.80761	1.00000	0.86555	0.73702	0.46376	-0.05542	0.99436	0.33260	0.37088
H	0.75639	0.86631	0.82729	0.89175	0.86555	1.00000	0.84790	0.63623	-0.41954	0.84520	0.50462	0.59405
2B	0.68014	0.73765	0.74270	0.77981	0.73702	0.84790	1.00000	0.45452	-0.47381	0.72816	0.36041	0.56589
3B	0.42370	0.46499	0.63469	0.72651	0.46376	0.63623	0.45452	1.00000	-0.19681	0.44333	0.16977	0.32331
HR	-0.46353	-0.05619	-0.07972	-0.27631	-0.05542	-0.41954	-0.47381	-0.19681	1.00000	0.00405	-0.52140	-0.61650
RBI	0.62878	0.99438	0.85425	0.79696	0.99436	0.84520	0.72816	0.44333	0.00405	1.00000	0.30422	0.34397
SB	0.46670	0.33302	0.24842	0.35967	0.33260	0.50462	0.36041	0.16977	-0.52140	0.30422	1.00000	0.77136
CS	0.45306	0.37143	0.38269	0.43311	0.37088	0.59405	0.56589	0.32331	-0.61650	0.34397	0.77136	1.00000
BB	-0.15881	0.44085	0.51251	0.19658	0.44086	0.13596	0.17630	0.03499	0.51369	0.47856	-0.17309	-0.04530
SO	-0.61857	-0.54576	-0.52111	-0.61484	-0.54447	-0.80731	-0.77081	-0.36543	0.65876	-0.51346	-0.57733	-0.72620
BA	0.73586	0.85000	0.77520	0.83430	0.84936	0.99325	0.84135	0.59012	-0.44000	0.82756	0.51553	0.60905
OBP	0.61353	0.92249	0.86191	0.81327	0.92201	0.94907	0.80868	0.53260	-0.23446	0.91078	0.43528	0.54265
SLG	0.43282	0.85952	0.77869	0.69333	0.85942	0.74057	0.59366	0.52066	0.28072	0.88316	0.11875	0.18506
OPS	0.52904	0.92646	0.85224	0.77702	0.92619	0.86496	0.71380	0.54762	0.06387	0.93558	0.26914	0.35902
E	-0.15448	0.13422	0.27045	0.17577	0.13527	0.27157	0.30729	0.18944	0.02685	0.16932	0.03470	0.34009
DP	0.17872	0.35099	0.23738	0.14473	0.35075	0.26547	0.25491	0.15182	0.10679	0.37634	0.37416	0.31491
Fld%	0.31290	-0.01832	-0.14516	-0.02289	-0.01945	-0.11504	-0.16053	-0.08783	-0.10890	-0.05575	0.08464	-0.22157
BatAge	0.45318	0.17484	0.02845	0.14721	0.17486	0.13821	0.16312	-0.07473	-0.19350	0.16020	0.27446	0.13960

Table A2: Correlation Matrix for Wins and Offensive Variables (Part 2)

Pearson Correlation Coefficients, N = 25										
	BB	SO	BA	OBP	SLG	OPS	E	DP	Fld%	BatAge
W	-0.15881	-0.61857	0.73586	0.61353	0.43282	0.52904	-0.15448	0.17872	0.31290	0.45318
R/G	0.44085	-0.54576	0.85000	0.92249	0.85952	0.92646	0.13422	0.35099	-0.01832	0.17484
PA	0.51251	-0.52111	0.77520	0.86191	0.77869	0.85224	0.27045	0.23738	-0.14516	0.02845
AB	0.19658	-0.61484	0.83430	0.81327	0.69333	0.77702	0.17577	0.14473	-0.02289	0.14721
R	0.44086	-0.54447	0.84936	0.92201	0.85942	0.92619	0.13527	0.35075	-0.01945	0.17486
H	0.13596	-0.80731	0.99325	0.94907	0.74057	0.86496	0.27157	0.26547	-0.11504	0.13821
2B	0.17630	-0.77081	0.84135	0.80868	0.59366	0.71380	0.30729	0.25491	-0.16053	0.16312
3B	0.03499	-0.36543	0.59012	0.53260	0.52066	0.54762	0.18944	0.15182	-0.08783	-0.07473
HR	0.51369	0.65876	-0.44000	-0.23446	0.28072	0.06387	0.02685	0.10679	-0.10890	-0.19350
RBI	0.47856	-0.51346	0.82756	0.91078	0.88316	0.93558	0.16932	0.37634	-0.05575	0.16020
SB	-0.17309	-0.57733	0.51553	0.43528	0.11875	0.26914	0.03470	0.37416	0.08464	0.27446
CS	-0.04530	-0.72620	0.60905	0.54265	0.18506	0.35902	0.34009	0.31491	-0.22157	0.13960
BB	1.00000	0.10306	0.12242	0.40359	0.52703	0.50299	0.30141	0.34379	-0.32134	-0.13094
SO	0.10306	1.00000	-0.83110	-0.72597	-0.38963	-0.55447	-0.26511	-0.16070	0.11633	-0.25852
BA	0.12242	-0.83110	1.00000	0.95248	0.72974	0.85969	0.28759	0.28457	-0.13491	0.13523
OBP	0.40359	-0.72597	0.95248	1.00000	0.83322	0.94507	0.33881	0.37098	-0.20679	0.06099
SLG	0.52703	-0.38963	0.72974	0.83322	1.00000	0.96792	0.35062	0.39843	-0.24811	-0.01016
OPS	0.50299	-0.55447	0.85969	0.94507	0.96792	1.00000	0.36521	0.40686	-0.24551	0.02115
E	0.30141	-0.26511	0.28759	0.33881	0.35062	0.36521	1.00000	0.16541	-0.97562	-0.43787
DP	0.34379	-0.16070	0.28457	0.37098	0.39843	0.40686	0.16541	1.00000	-0.08977	-0.18653
Fld%	-0.32134	0.11633	-0.13491	-0.20679	-0.24811	-0.24551	-0.97562	-0.08977	1.00000	0.48117
BatAge	-0.13094	-0.25852	0.13523	0.06099	-0.01016	0.02115	-0.43787	-0.18653	0.48117	1.00000

Table A3: Correlation Matrix for Selected Predictors and Wins

	Wins	**Hits**	**Doubles**	**BatAvg**	**Runs**
Wins	1.000	0.756	0.680	0.736	0.659
Hits	0.756	1.000	0.884	0.993	0.929
Doubles	0.680	0.884	1.000	0.871	0.896
BatAvg	0.736	0.993	0.871	1.000	0.933
Runs	0.659	0.929	0.896	0.933	1.000

Table A4: Summary Statistics for Selected Variables

Variable	**Min**	**Q1**	**Mean**	**Median**	**Q3**	**Max**
Hits	1209	1331	1430.64	1447	1520	1604
BatAvg	0.225	0.245	0.25868	0.260	0.274	0.285
Doubles	212	251	272.64	274	289	333
Runs	623	691	742.68	736	769	883
Wins	63	75	83.20	80	92	100

Table A5: Model Comparison for Two Variable Regression Models in Predicting Total Wins

Predictors	**Root MSE**	**Adj r^2**	**Coeff Var**	**F**	$p_{\textbf{Model}}$
Hits + BatAvg	6.838	0.5525	8.22	15.82	< 0.0001
Hits + Doubles	6.940	0.5391	8.34	15.03	< 0.0001
Hits + Runs	6.983	0.5333	8.39	14.71	< 0.0001
BatAvg + Doubles	7.129	0.5137	8.57	13.68	0.0001
BatAvg + Runs	7.196	0.5044	8.65	13.22	0.0002
Doubles + Runs	7.418	0.4735	8.92	11.79	0.0003

Table A6: Model Comparison for Three Variable Regressions

Predictors	**Root MSE**	**Adj r^2**	**Coeff Var**	**F**	$p_{\textbf{Model}}$
Hits + BatAvg + Doubles	6.956	0.5370	8.36	10.28	0.0002
Hits + BatAvg + Runs	6.998	0.5314	8.41	10.07	0.0003

Table A7: Model Comparison for Interaction and Quadratic Regressions

Predictors	**Root MSE**	**Adj r^2**	**Coeff Var**	**F**	$p_{\textbf{Model}}$
Hits + BatAvg + Hits $\times$ BatAvg	6.998	0.5313	8.41	10.07	0.0003
Hits + BatAvg + Hits^2	6.999	0.5313	8.41	10.07	0.0003

Table A8: Regression Output for Final Model: Wins = Hits + Batting Average

Variable	**Estimate (β)**	**Std. Error**	**t Value**	**p Value**
Intercept	22.2465	35.8589	0.62	0.5414
Hits	0.1609	0.0999	1.61	0.1217
BatAvg	-653.9606	671.6898	-0.97	0.3408
Model Fit Statistics				
Root MSE		6.8383		
Adj r^2		0.5525		
Coeff Var		8.22		
F Value		15.82		
Model p Value		< 0.0001		