Statistics Group Project

Baseball Dataset

Group 5
Mantek Singh Bhatia
Elodie Schreiber
Domonic Bodnar
Alec Borkosky

Index

I. Introduction

1. Data Set used in this Project

The data set we will be using for this project is a baseball data set with a total of 45 records. Our response variable is the batting average and the predictors we think that will be useful in predicting this batting average are runs scored/times batted, doubles/times batted, triples/times batted, home runs/times batted and strike outs/time batted.

Our response variable batting average is defined as the average runs a batter will score in each game.
Runs scored/times batted can be defined as how many runs a batter scores, each time he comes out to bat.
Doubles/times batted can be defined as how many doubles a batter scores, each time he comes out to bat.
Triples/times batted can be defined as how many triples a batter scores, each time he comes out to bat.
Home runs/times batted can be defined as how many home runs the batter scores, each time he comes out to bat.
Strike outs/times batted can be defined as the average number of times the batter will be struck out, each time he comes out to bat.

It is expected that Runs scored/ time batted, Doubles/ times batted, Triples/times batted, and Home runs/times batted will have a positive relationship with the batting average and that strike outs/times batted will have a negative relationship with the batting average.

The goal of the study is to find a regression model that can explain the maximum variation in our response variable (batting average) in relationship to the predictor variables. We hope to find the least Multicollinearity between the predictors, a high R^2 value and a regression equation that will make our model a good fit for the data we have. We will find the confidence interval and prediction interval for a particular set of values of the predictors. Moreover, we will be predicting a value of the response variable by assuming the value of the set of predictors and determine whether the predicted response variable value falls between the confidence interval and the prediction interval.

2. Introducing the Concepts used in this Project

Multiple Regression establishes a relationship between one dependent variable and two or more independent variables. Response variable is another name for the dependent variable. It is the variable that is being predicted. Predictor variables are independent variables used to predict the response variable and explain its changes.

The Regression equation is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots. + \beta_n X_n + \epsilon$

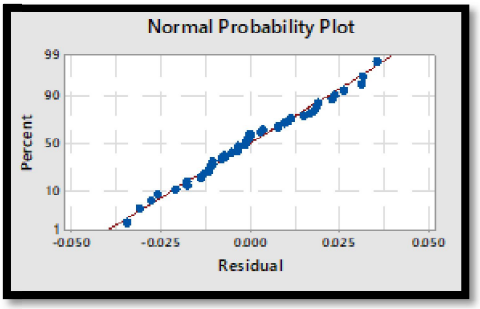
Coefficients: Y is the response variable
$X_1, X_2, \ldots\ldots, X_n$ are the predictor variables
β_0: Y intercept of the model, unknown regression parameter
β_1: the regression parameter for X_1, or the slope of X_1
β_2: the regression parameter for X_2, or the slope of X_2

Error (ϵ) describes the effect on Y of all factors other than the values of the independent variables. Error is the random fluctuations that cause the value of Y to deviate from the mean level. Assumptions about the Error term:

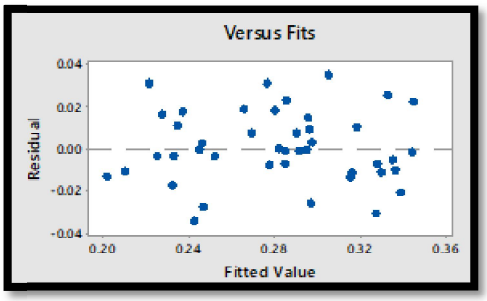
$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

We assume that each error has a normal distribution with the same variance, and errors are independent.

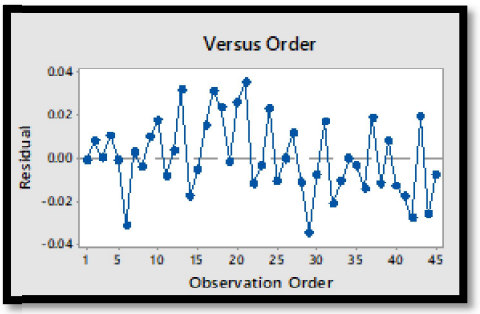
a) Normality is checked by using a normal probability plot of residuals. If they for a straight line, the error terms are considered to be normally distributed.



b) Constant variance assumption may be checked by plotting the residuals against the fitted vales and each error term. Vertical scatter on each plot should be roughly the same as you move across horizontal axis if the assumption is reasonable.



c) Independence of the error terms may be checked by plotting the residuals against the observation order and each error term. If the error terms are randomly distributed along the zero, they are considered Independent.



Confidence Interval (CI) is used for estimating the population mean value of the dependent variable, Y, corresponding to the specific values of the independent variables, X_1, X_2, X_3,...., X_n. It predicts that we are "x%" confident that the mean value of Y will fall into the interval.

Prediction Interval (PI) is used when we want to predict one particular value of the dependent variable, Y, corresponding to a specific value of the independent variables, X_1, X_2, X_3,...., X_n. It predicts that we are "x%" confident that a single value of Y will fall into the interval.

The confidence interval estimate of the expected value of Y will be narrower than the prediction interval for the same given value of the independent variables, X_1, X_2, X_3,...., X_n. at the same confidence level.

Multicollinearity exists among the independent variables in a regression if these independent variables are related to or dependent upon each other. It is measured through the correlation matrix, variance inflation factors.
It is moderately strong if VIF is greater than 5, severe if VIF is greater than 10. Ideally, VIF should be under 2.
Another indications for Multicollinearity is if the F statistic for the overall regression is large and significant (p-value is small) but the T statistic for individual coefficients are small and probably insignificant (p values are high).

Standard error of estimate is the point estimate of σ, measure of the accuracy of predictions made with regression lines.

$$s = \sqrt{\frac{SSE}{n - (k + 1)}}$$

If the ratio percentage of S/Mean (of the response variable) is less than 10%, a model is considered a good fit.

Coefficient of determination (R^2) is the proportion of the total variation in n observed values (where n is the total number of samples considered) of the dependent variable that is explained by the overall regression model.

$$R^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}$$

Adjusted R^2 is sometimes used to avoid overestimating the importance of the independent variables. This is the adjusted value of R^2 for the number of independent variables (or degrees of freedom).

Stepwise Regression Analysis method is an iterative model building technique used for selecting important predictor variables. Stepwise regression begins by considering all of the one-independent-variable models and choosing the model for which the p-value related to the independent variable in the model is the smallest. If this p-value is less than α_{entry}, an α value

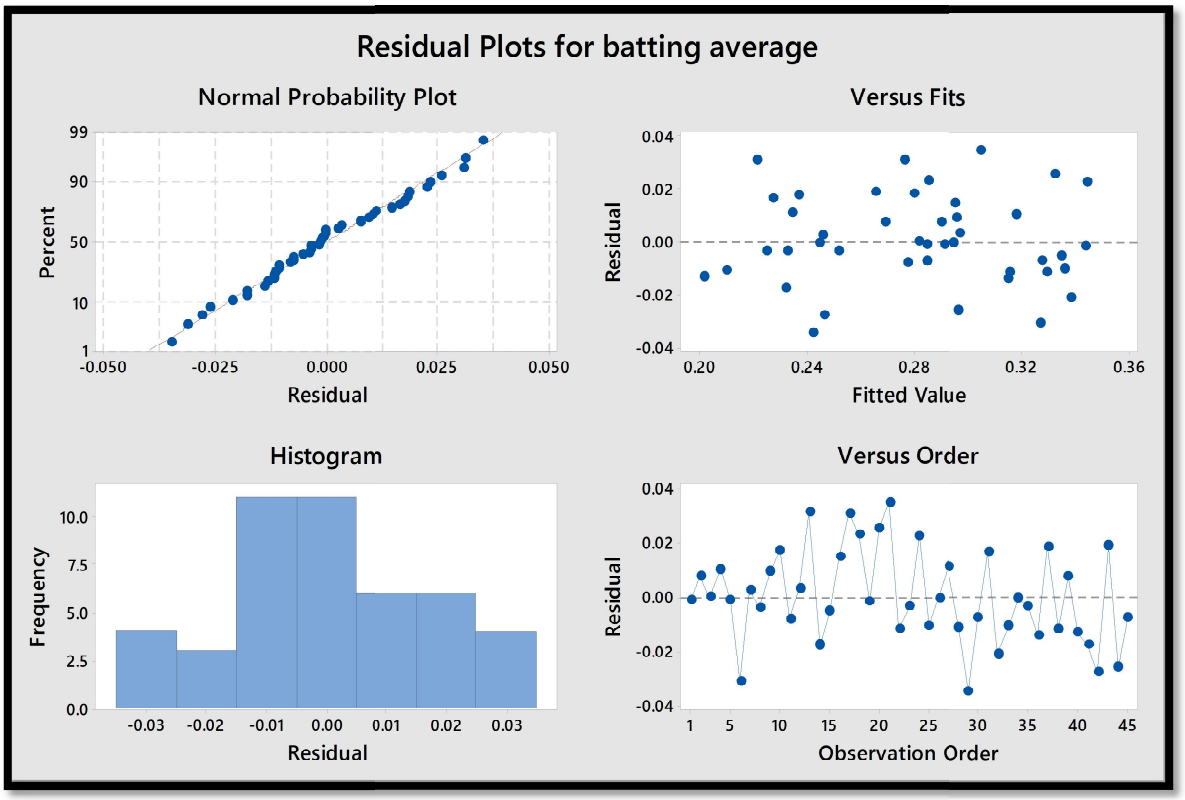
for “entering” a variable, the independent variable is the first variable entered into the stepwise regression model and stepwise regression continues. Stepwise regression then considers the remaining independent variables not in the stepwise model and chooses the independent variable which, when paired with the first independent variable entered, has the smallest p-value. If this p-value is less than α_{entry}, the new variable is entered into the stepwise model. Moreover, the stepwise procedure checks to see if the p-value related to the first variable entered into the stepwise model is less than α_{stay}, an α value for allowing a variable to stay in the stepwise model. This is done because Multicollinearity could have changed the p-value of the first variable entered into the stepwise model. The stepwise procedure continues this process and concludes when no new independent variable can be entered into the stepwise model. It is common practice to set both α_{entry} and α_{stay} equal to 0.05 or 0.10.

Best Subset Regression Analysis Method compares all possible models that can be created based upon an identified set of predictors. Shows the 2 (this number can be changed up to 5) best models for each predictor variable.

II. Analysis

1. Checking Assumptions

1.1. Normality, Variance and Independence of error terms



The first plot, i.e., Normal Probability Plot of the residual is a straight line. Therefore, our assumption of normality of error terms is valid. The second plot, i.e., the versus fits plot is somewhat symmetrical around the residual equals zero value. Therefore, our assumption that all errors terms have equal variance and zero mean is valid. The fourth plot, i.e., the versus order plot has random plot of all residuals around the residual equals zero line. Therefore, our assumption that all error terms are independent of each other is also valid.

Therefore, our assumption of $\varepsilon_i \overset{iid}{\sim} \mathbf{N(0, \sigma^2)}$ is met.

2. Checking for Multicollinearity

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	0.073275	0.014655	47.96	0.000
runs scored/times batted	1	0.005072	0.005072	16.60	0.000
doubles/times batted	1	0.003061	0.003061	10.02	0.003
triples/times batted	1	0.000350	0.000350	1.15	0.291
home runs/times batted	1	0.000798	0.000798	2.61	0.114
strike outs/times batted	1	0.009231	0.009231	30.21	0.000
Error	39	0.011916	0.000306		
Lack-of-Fit	37	0.011916	0.000322		*
Pure Error	2	0.000000	0.000000		
Total	44	0.085191			

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.1832	0.0171	10.69	0.000	
runs scored/times batted	0.447	0.110	4.07	0.000	3.05
doubles/times batted	0.991	0.313	3.16	0.003	1.54
triples/times batted	0.622	0.581	1.07	0.291	2.35
home runs/times batted	0.274	0.169	1.62	0.114	2.05
strike outs/times batted	-0.2846	0.0518	-5.50	0.000	1.53

The high F-value of the Regression and P-value < α suggests that the model is a good fit. The low VIF values (around 2), of the predictors are indicators that the Multicollinearity in the model is insignificant.

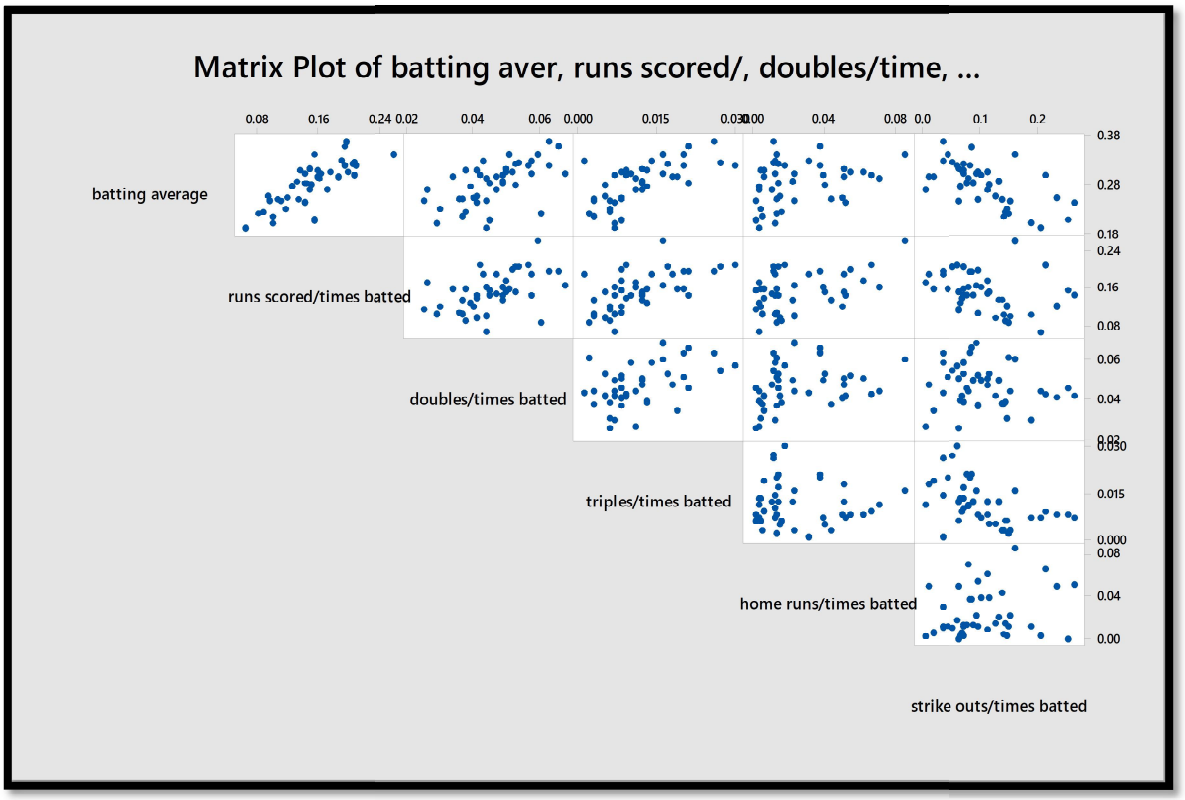
Correlation: batting average, runs scored/times batted, ... times batted

Correlations

	batting average	runs scored/time	doubles/times ba	triples/times ba
runs scored/time	0.825			
	0.000			
doubles/times ba	0.609	0.517		
	0.000	0.000		
triples/times ba	0.662	0.599	0.487	
	0.000	0.000	0.001	
home runs/times	0.336	0.496	0.299	-0.037
	0.024	0.001	0.046	0.808
strike outs/time	-0.621	-0.366	-0.157	-0.487
	0.000	0.013	0.304	0.001

	home runs/times
strike outs/time	0.197
	0.194

Cell Contents
Pearson correlation
P-Value



The Pearson correlation above tells us that the correlation between predictors is weak. However, the correlation between the response and individual predictors is moderately high.

The matrix plot suggests the correlation between the predictors moderately low and between the response and each predictor is high. Therefore, our model is appropriate for the regression analysis.

3. Regression Analysis

3.1. Stepwise Regression Analysis

Regression Analysis: batting average versus runs ... outs/times batted

Stepwise Selection of Terms

Candidate terms: runs scored/times batted, doubles/times batted, triples/times batted, home runs/times batted, strike outs/times batted

	------Step 1-----		------Step 2------		------Step 3------	
	Coef	P	Coef	P	Coef	P
Constant	0.1501		0.1983		0.1694	
runs scored/times batted	0.8641	0.000	0.7227	0.000	0.5749	0.000
strike outs/times batted			-0.2574	0.000	-0.2643	0.000
doubles/times batted					1.119	0.001
S		0.0251575		0.0202238		0.0176521
R-sq		68.05%		79.84%		85.00%
R-sq(adj)		67.31%		78.88%		83.91%
R-sq(pred)		65.19%		76.47%		80.83%
Mallows' Cp		48.07		17.22		4.81

α to enter = 0.15, α to remove = 0.15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0.072416	0.024139	77.47	0.000
runs scored/times batted	1	0.016670	0.016670	53.50	0.000
doubles/times batted	1	0.004403	0.004403	14.13	0.001
strike outs/times batted	1	0.010568	0.010568	33.91	0.000
Error	41	0.012776	0.000312		
Lack-of-Fit	39	0.012776	0.000328	*	*
Pure Error	2	0.000000	0.000000		
Total	44	0.085191			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0176521	85.00%	83.91%	80.83%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.1694	0.0151	11.19	0.000	
runs scored/times batted	0.5749	0.0786	7.31	0.000	1.54
doubles/times batted	1.119	0.298	3.76	0.001	1.37
strike outs/times batted	-0.2643	0.0454	-5.82	0.000	1.16

Regression Equation

batting average = 0.1694 + 0.5749 runs scored/times batted + 1.119 doubles/times batted - 0.2643 strike outs/times batted

Fits and Diagnostics for Unusual Observations

Obs	batting average	Fit	Resid	Std Resid	
21	0.34000	0.30503	0.03497	2.04	R
29	0.20700	0.24171	-0.03471	-2.16	R

R Large residual

3.2 Best Subset Regression Analysis

Best Subsets Regression: batting average versus runs ... s/times batted

Response is batting average

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	runs scored/times batted	doubles/times batted	triples/times batted	home runs/times batted	strike outs/times batted
1	68.1	67.3	65.2	48.1	0.025158	X				
1	43.8	42.5	37.6	115.8	0.033376			X		
2	79.8	78.9	76.5	17.2	0.020224	X				X
2	72.6	71.3	68.7	37.4	0.023575	X	X			
3	85.0	83.9	80.8	4.8	0.017652	X	X			X
3	80.7	79.3	76.1	16.8	0.020032	X		X		X
4	85.6	84.2	80.4	5.1	0.017511	X	X		X	X
4	85.1	83.6	80.0	6.6	0.017829	X	X	X		X
5	86.0	84.2	79.9	6.0	0.017480	X	X	X	X	X

3.3 Analysis of both Regression methods

The Stepwise Regression Analysis method runs until step 3 to find the appropriate regression model for our data, According to this method, only 3 of the 5 predictors are significant to explain the variation is the response variable. The analysis of the Stepwise is further corroborated by the Best Subset Regression Analysis method.

The Best Subset Regression Analysis method generates the indicators like R^2, R^2(adj), and C value for all possible subsets of predictors. Then, based on these indicators, we decide which model would explain the most variation in the response variable.

Runs scored/times batted, doubles/times batted, and strike outs/times batted are deemed significant, whereas, triples/times batted and home runs/times batted are considered insignificant in our model. The Stepwise method also calculates indicators like the Coefficient of Determination (R^2), Regression Equation, Standard Error of Estimate (S) and the Coefficients for our model, which are analyzed in latter sections.

4. Regression Equation

Regression Equation

batting average = 0.1694 + 0.5749 runs scored/times batted + 1.119 doubles/times batted - 0.2643 strike outs/times batted

4.1 Significant and Insignificant Predictors

According to our regression model, Runs scored/times batted, doubles/times batted and strikeouts/times batted are the significant predictors. While, triples/times batted and home runs/times batted are considered as insignificant predictors. This is due to the p values of the insignificant predictors being greater than α, therefore, they are deemed insignificant.

4.2 Least Square Line

$Y(hat) = 0.1694 + 0.5749X_1 + 1.119X_2 - 0.2643X_3$

where $Y(hat)$ = batting average
X_1 = runs scored/times batted
X_2 = doubles/times batted
X_3 = strike outs/times batted

4.3 Regression Equation

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

where Y = Response variable
β_0 = y intercept
β_1 = coefficient of the first predictor or slope of X_1
β_2 = coefficient of the second predictor or slope of X_2
β_3 = coefficient of the third predictor or slope of X_3
ϵ = error term in Y

In our model, the regression equation is

$Y = 0.1694 + 0.5749X_1 + 1.119X_2 - 0.2643X_3$

where Y = batting average
X_1 = runs scored/times batted
X_2 = doubles/times batted
X_3 = strike outs/times batted

4.4 Analysis of coefficients

Predictor	Coefficients	Significance
Constant	0.1694	The batting average is atleast 0.1694 even when the value of other predictors is 0. This can be possible due to external predictors that we are not considering in our model
Runs scored/times batted	0.5749	The batting average increases by an average of 0.5749 if the runs scored/times batted increases by 1 unit
Doubles/times batted	1.119	The batting average increases by an average of 1.119 if the doubles/times batted increases by 1 unit
Strike outs/times batted	- 0.2643	The batting average decreases by an average of -0.2643 if the strike outs/times batted increases by 1 unit

5. Prediction of a Response using some values of Predictors

In this section we make a prediction for the response variable by assuming values of the predictors.

Assume, runs scored/times batted = 0.15
Doubles/times batted = 0.045
Strike outs/times batted = 0.14

Now, we can predict the batting average using the values of these predictors.

Our equation is,

Batting average = 0.1694 + 0.5749 runs scored/times batted + 1.119 doubles/times batted - 0.2643 strike outs/times batted

= 0.1694 + (0.5794 x 0.15) + (1.119 x 0.045) - (0.2643 x 0.14)

= 0.269663

Therefore, batting average is 0.269663 when the values of the predictor variables are above mentioned.

6. Coefficient of Determination (R^2)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0176521	85.00%	83.91%	80.83%

$R^2 = 85\%$	This states that 85% of the variation in the Batting Average can be explained by the variation in the independent variables. The independent variables are the significant predictors we consider in our model
R^2 (adj) = 83.91%	This states that after being adjusted for the number of independent variables (or degrees of freedom), 83.91% of the variation in the Batting Average can be explained by the variation in the independent variables. The independent variables are the significant predictors we consider in our model

Since both of these are above 50%, it indicates that our model is a good fit.

7. Standard Error of Estimate

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0176521	85.00%	83.91%	80.83%

Descriptive Statistics: batting average

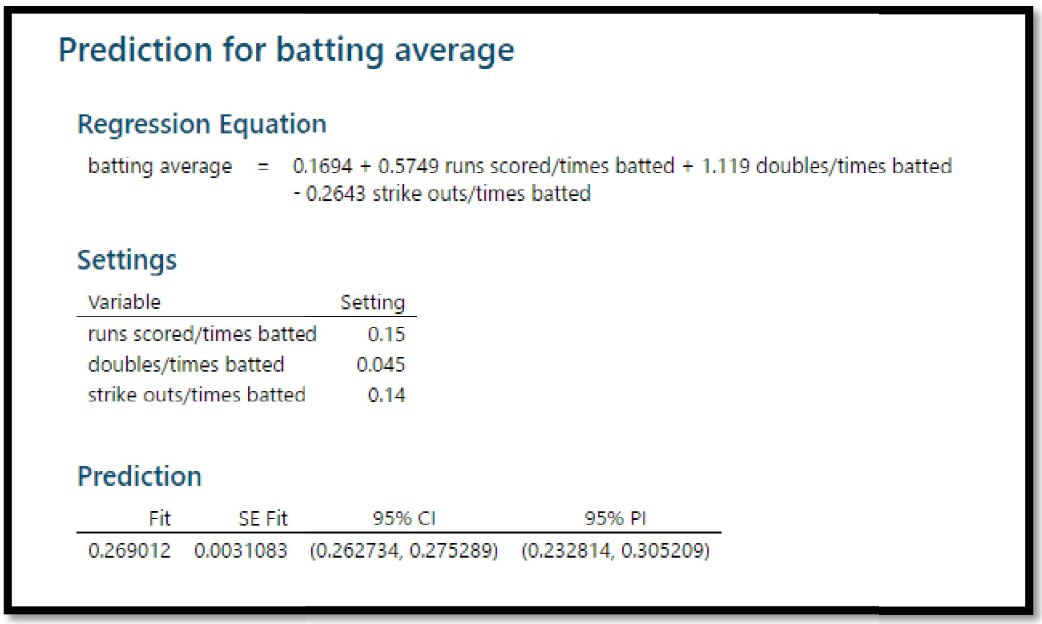
Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
batting average	45	0	0.28047	0.00656	0.04400	0.18800	0.24650	0.29000	0.30900

Variable	Maximum
batting average	0.36700

S is the Standard Error of Estimate. It is an estimate of the standard deviation of the error term.
In our model, S = 0.0176521.
Mean of Batting Average = 0.28047
Coefficient of Variation (CV) = S / Mean = 0.0176521 / 0.28047 = 0.06294
Since, CV = 6.294% < 10%, we can confirm that our linear model fit the data well.

8. Confidence Interval and Prediction Interval



Prediction for batting average

Regression Equation

batting average = 0.1694 + 0.5749 runs scored/times batted + 1.119 doubles/times batted - 0.2643 strike outs/times batted

Settings

Variable	Setting
runs scored/times batted	0.15
doubles/times batted	0.045
strike outs/times batted	0.14

Prediction

Fit	SE Fit	95% CI	95% PI
0.269012	0.0031083	(0.262734, 0.275289)	(0.232814, 0.305209)

8.1 Interpretation of the Prediction Interval (PI)

In our model, if we assume, runs scored/times batted = 0.15, Doubles/times batted = 0.045, Strike outs/times batted = 0.14, then we are 95% confident that a Batting Average will be between 0.232814 and 0.305209.

8.2 Interpretation of the Confidence Interval (CI)

In our model, if we assume, runs scored/times batted = 0.15, Doubles/times batted = 0.045, Strike outs/times batted = 0.14, then we are 95% confident that the average Batting Average will be between 0.262734 and 0.275289

These statements are corroborated by the prediction of Batting Average we made in Section 5 above. The predicted value of Batting Average lie in both the CI as well as the PI. We intentionally kept the value of the predictors same in this Section and Section 5 to validate these statements.

The confidence interval estimate of the expected value of Y will be narrower than the prediction interval for the same given value of the independent variables, X_1, X_2, X_3,...., X_n. at the same confidence level. In our model, we prove this statement to be true as well.

III. Summary and Conclusion

1. Summary

The data appears to cooperate perfectly with the formation of a regression model, and a wide variety of tests performed further prove this. The regression model meets all 4 of the assumptions made when performing regression analysis. These assumption include the error terms being normally distributed, having equal variances, having a mean of zero, and all terms being independent of one another. These assumptions are confirmed by the Versus Order, Versus Fit, and other graphs shown in section 1.1.

Section 2 makes it clear that the most significant input variables in determining batting average are runs scored/times batted, doubles/times batted, and strike outs/times batted. These variables all had extremely low p-values that all fall below the alpha of .05 (.05 is the alpha value because we tested the variables at 95% confidence). These same 3 input variables also had the three lowest F statistics, which only further confirms the same information. One thing to note from these tests is that the VIF, which shows the multicollinearity between input variables, is relatively low.

This independence of the input variables is then proven in the Matrix Plot. As one can see, the plots of 2 of the input variables don't appear to have any real relationship with one another, as the points in each graph are scattered throughout. The Matrix Plot also proves how the same 3 input variables mentioned in the previous paragraph are significant indicators of the output. The first 2 inputs, runs scored/times batted and doubles/times batted, have a clear and linear relationship with the output of batting average. At the top right of the matrix one can also see how strike outs/times batted inversely impact a player's batting average. It is interesting to note that by using the Matrix plot to examine the significance of each input variable on the output it is not as easy to see whether or not triples/times batted and home runs/times batted affect batting average, or it is at least not as easy to tell as in the tests performed in section 2 of the analysis.

The regression equation came to be $Y = 0.1694 + 0.5749X_1 + 1.119X_2 - 0.2643X_3$, where Y = batting average, X_1 = runs scored/times batted, X_2 = doubles/times batted, X_3 = strike outs/times batted. The coefficient for X_3 is negative because strike outs/times batted has a negative correlation to the output. Using this equation we were able to find an estimate output value that was consistent with the rest of the sample data. Using the same input values for the tests in Section 8 of the analysis, we found the estimated value also fell within the 95% Confidence Interval and the 95% Prediction Interval.

The final value to note is the adjusted R^2 value of 83.91%. This value states that 83.91% of the variation in the Batting Average can be explained by the variation in the independent variables. Considering professionals are often satisfied with an adjusted R^2 value of 50%, this value shows that our regression model is a clear fit.

2. Conclusion

While a well-fitting regression model was found, one must still be somewhat surprised by what the model relays. Of course it makes sense that hits and runs scored improve a player's batting average, as proven by the adjusted R^2 value, but not to the degree that one would expect. We initially believed that home runs/times batted would have a much more significant impact on a player's batting average, but there appears to only be a slight correlation between the response variable and the predictor, if any. We certainly did not expect doubles hit/times batted to have the highest correlation to batting average. Perhaps this is because doubles are the easiest hits to achieve, at least in the context of this data set. Triples/times batted and home runs/times batted may increase batting average by more on a single hit basis, but considering how much harder they are to achieve they don't play nearly as much of a role in increasing batting average as doubles do. To conclude, if a baseball player truly wants to improve their batting average they are best off not swinging for the fences, and instead making sure they do not strike out and just manage to get on base.