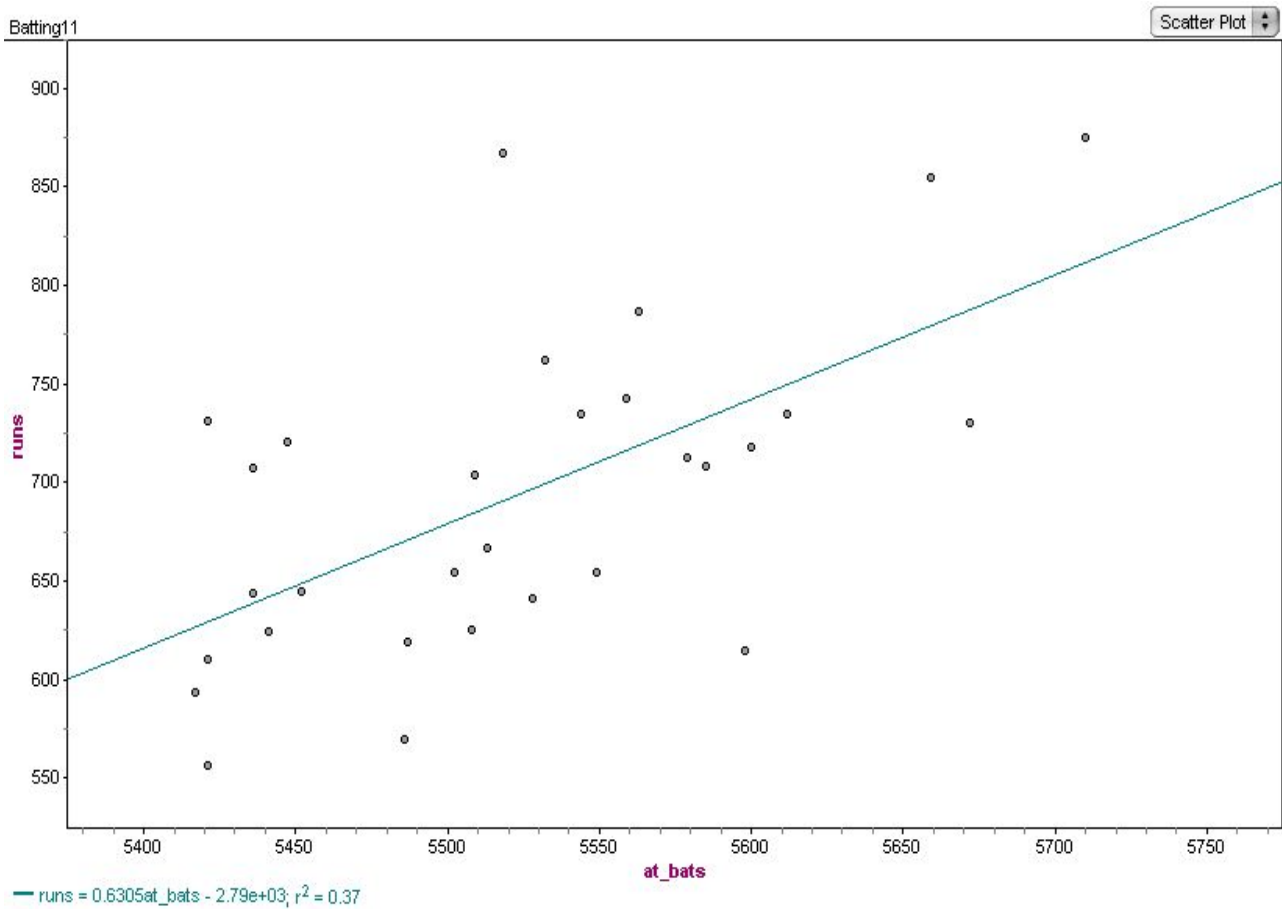
Ryan Yang
Professor Zanontian
404904494
Oct 8 2018

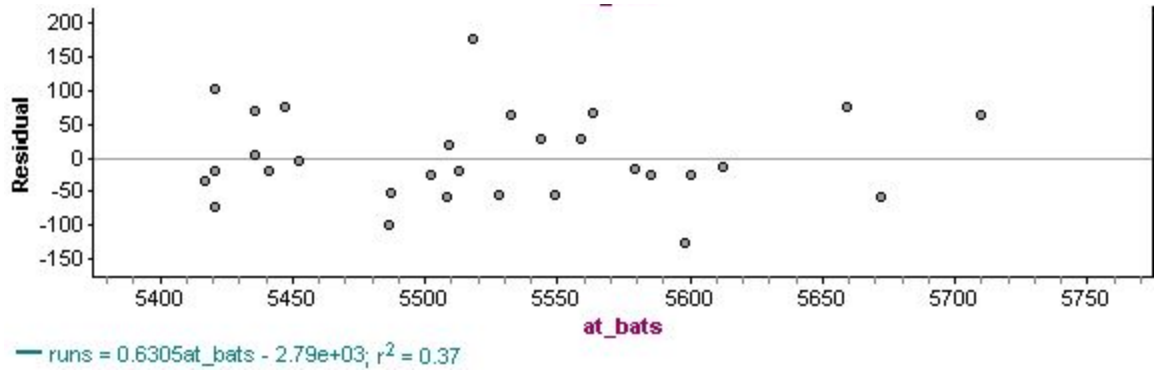
Lab 2

Question 1) Based on the graph with “at_bats” as the x variable predicting the y variable “runs”, we can see that based on the linear regression line that there is only a weak positive linear trend (slope is only 0.63) between at bats and number of runs; thus there is only a slight positive correlation between at_bats and runs.



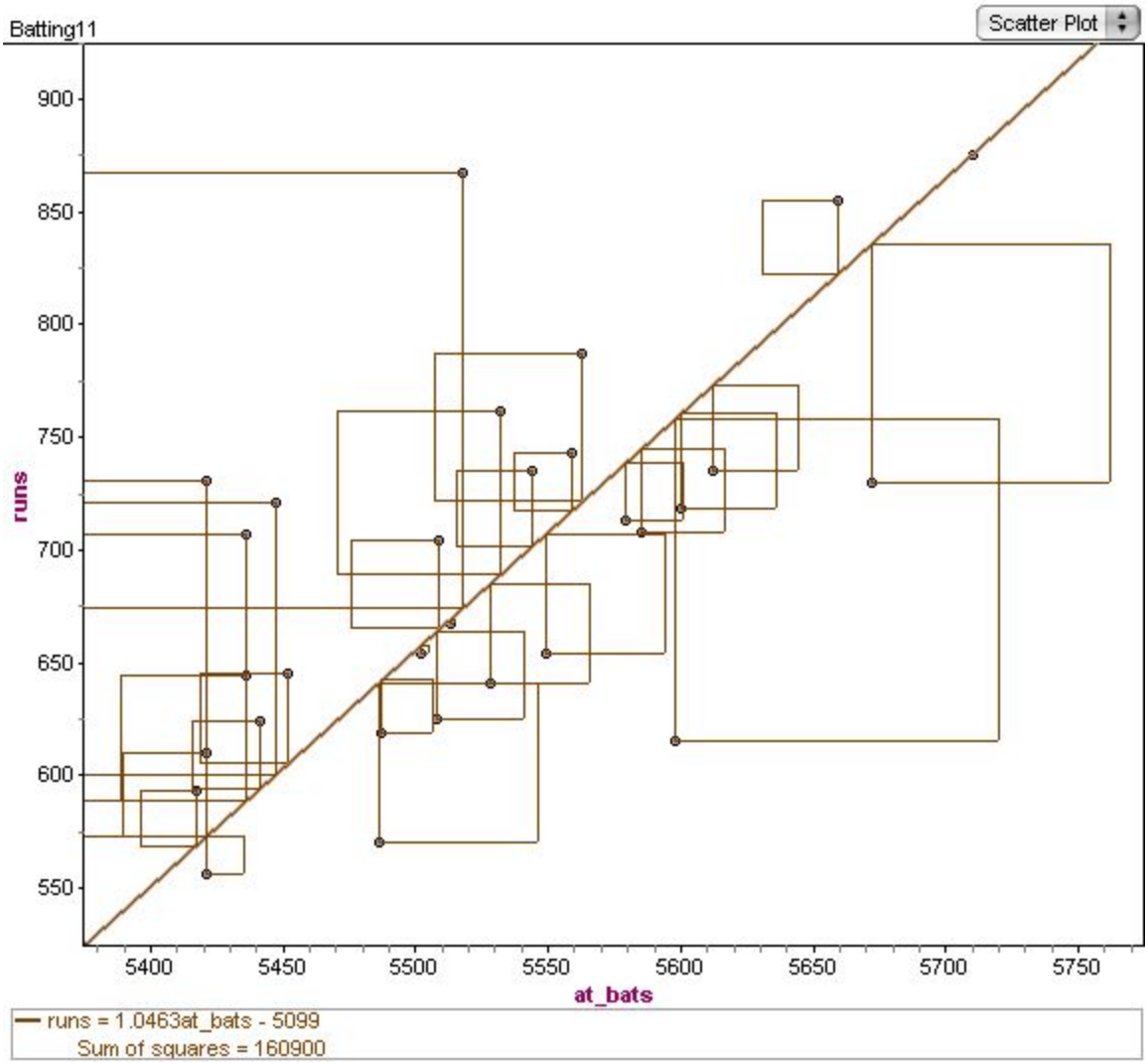
Question 2)

Based on the residual plot, we can see that there is a linear trend since there is no curve in the line and points are scattered randomly above and below the line.



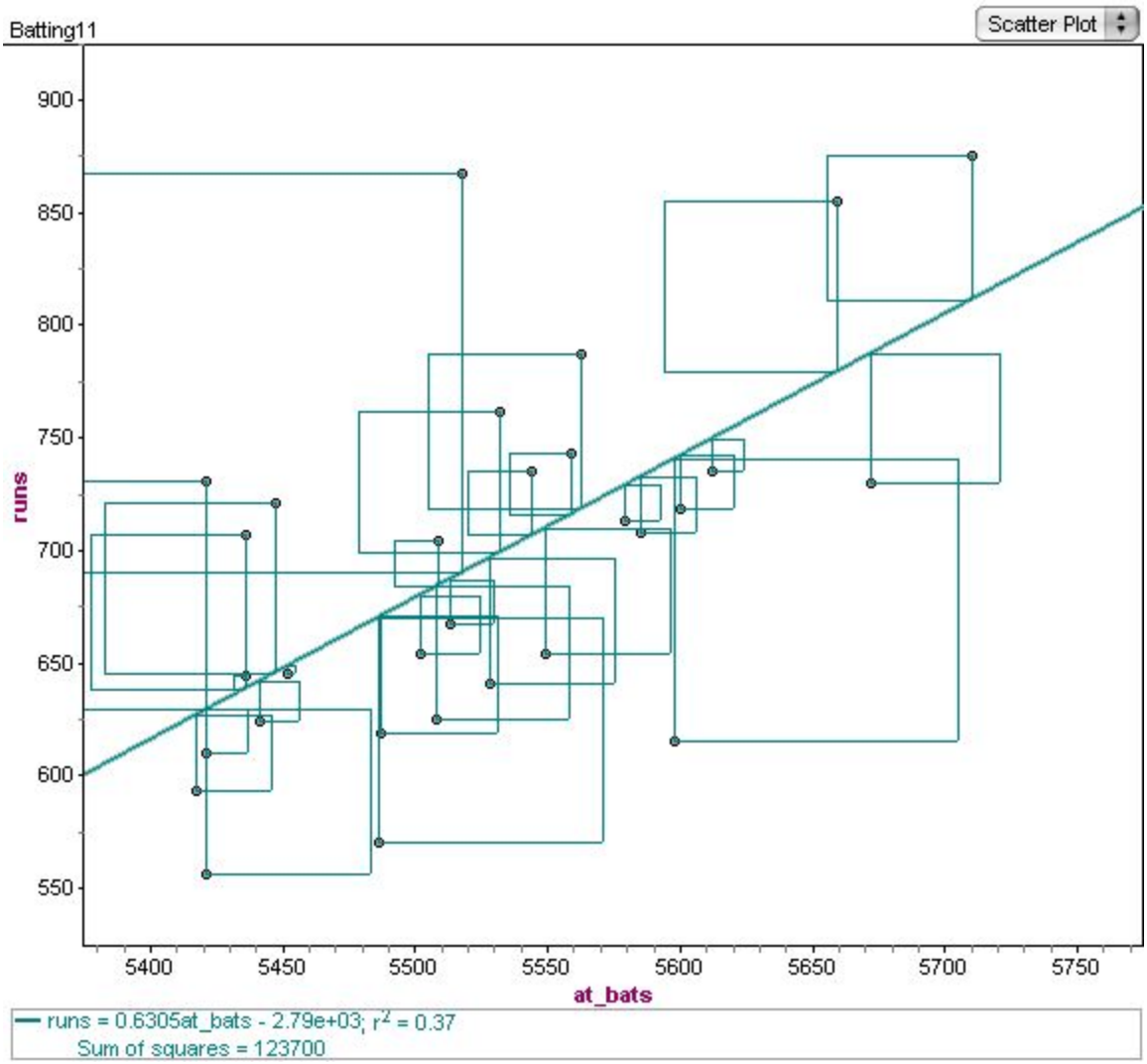
Question 3)

The Sum of squares error decreases as the line is moved to better fit the data as sum of squares is calculated by adding up the squared differences between the actual data point and the predicted data point on the line of best fit. Thus, if the line best fits the data, there will be a smaller sum of squares value. When I moved the line to anywhere besides the picture shown below, the sum of squares increased.



Question 4)

The graph with the least square line through the points has a lower sum of squares, meaning the least least square line better fits the data in comparison to the movable line that I tried to fit the data with. In comparison to the graph above, while both are linear and positively increasing, the line of best fit created using the least square line has a lower positive slope.



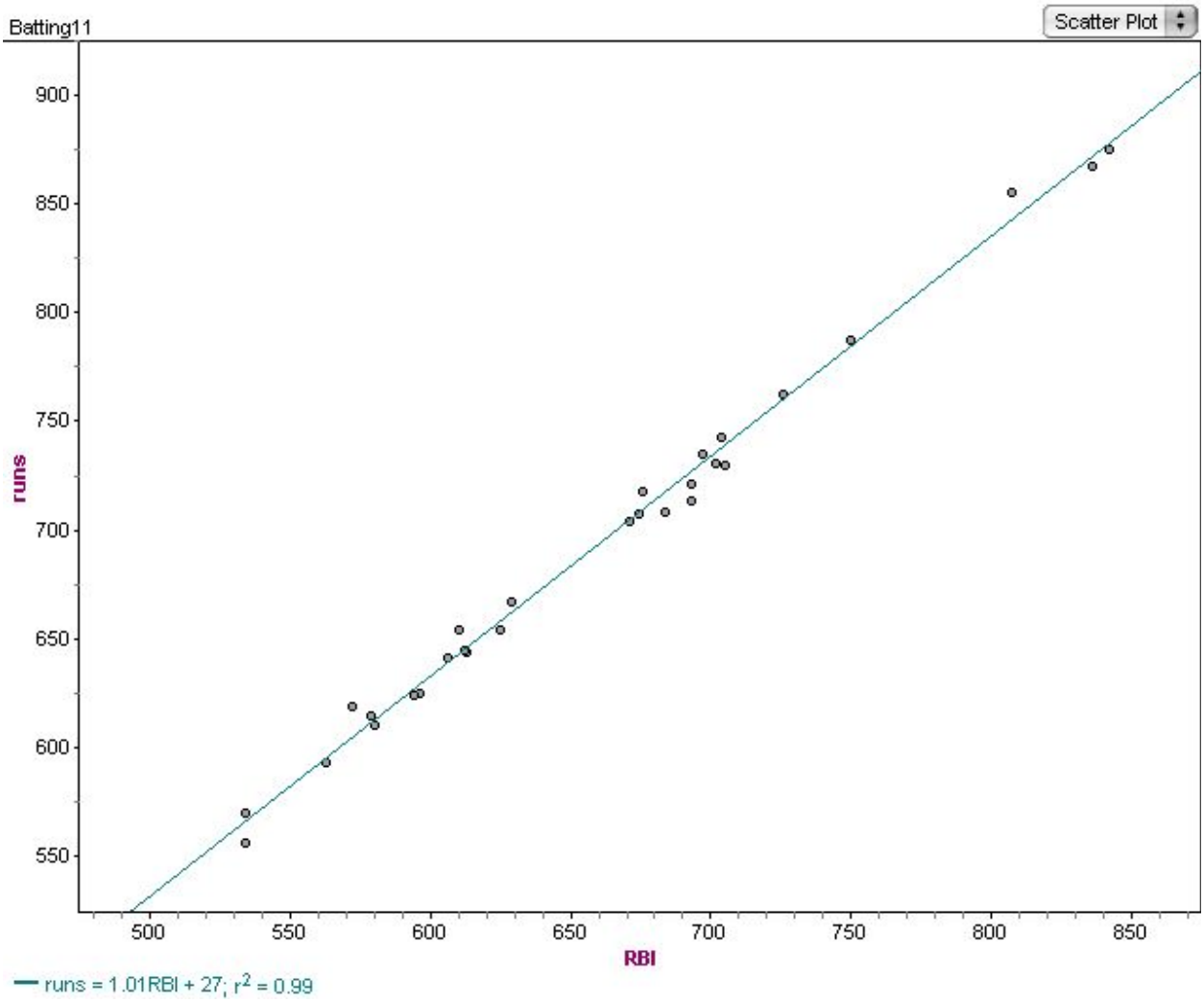
Question 5)

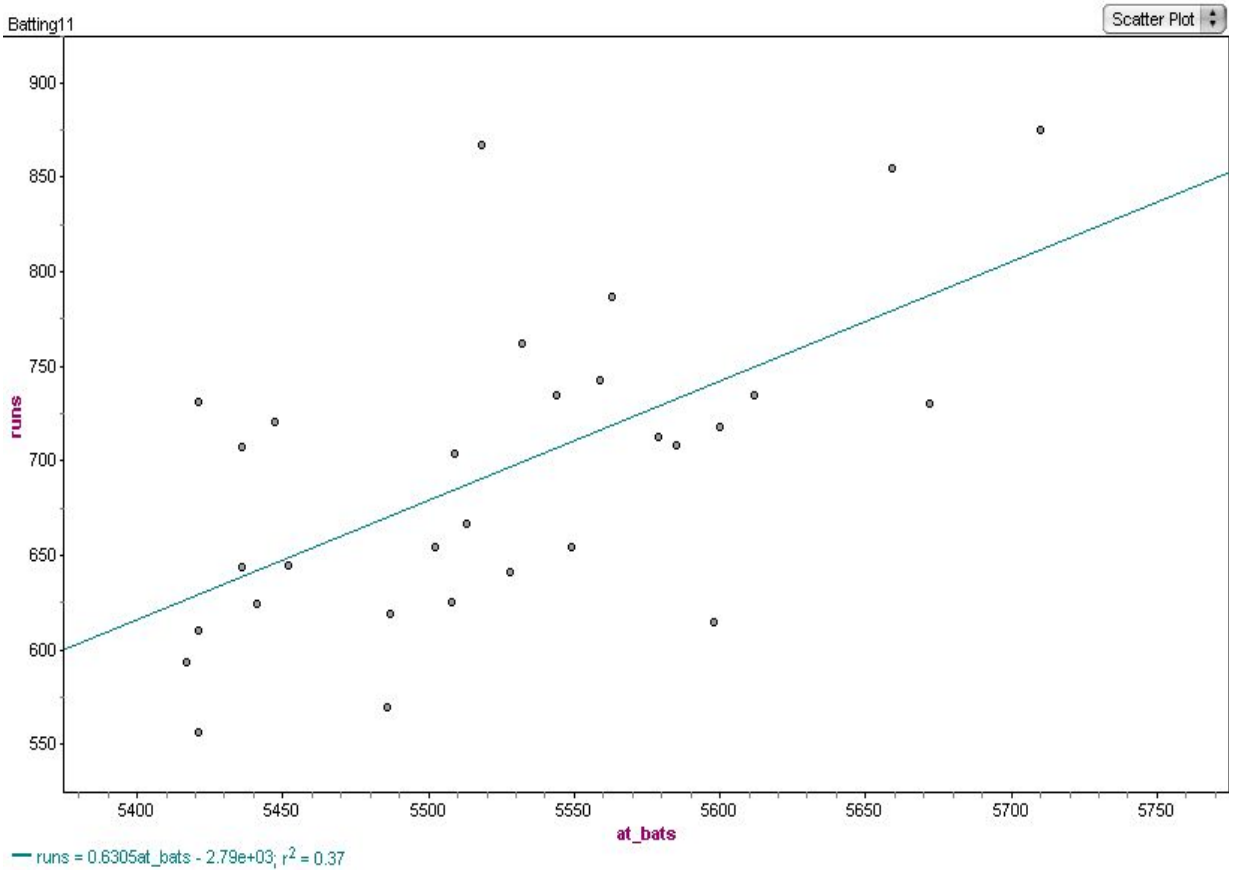
0.6305*(5508) -2790 = 682.794
Based on the data, the predicted runs for a team that had 5508 at_bats would be 682.794 based off the least square line in the data set. The LA angels had 667 runs, Florida Marlins had 625 runs, Chicago white sox had 654 runs, and the Cleveland Indians had 704 runs.
Thus, the sum of squares or typical error would be 15.794^2 + 57.794^2 + 28.794^2 + 21.206^2 = 4868.385

Team	Far off from actual number of runs
LA Angels	667 - 682.794 - = -15.794
Florida Marlins	625 - 682.794 = -57.794
Chicago White Sox	654- 682.794 = -28.794
Cleveland Indians	704 - 682.794 = 21.206

Question 6)
I believe that Runs Batted In will have the lowest sum of squares using the least square line, due to the fact that these variables are almost completely related since other than batting the ball to score a run, there aren't

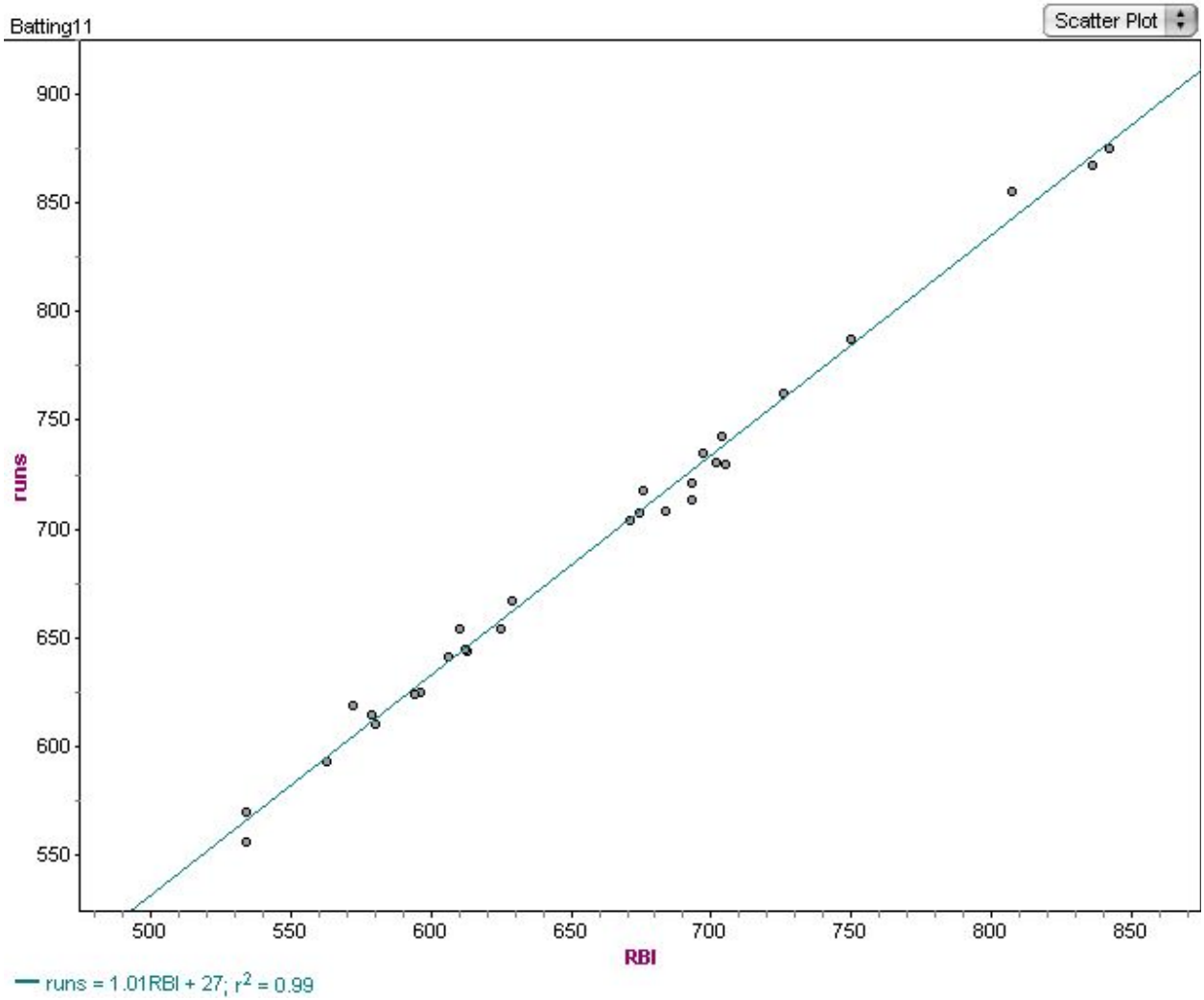
that many ways to score a run; for instance, getting hit by the ball will only allow you to run to first base which I believe doesn't happen often. At first glance, the relationship seems to be linear and compared to the first graph with variables runs and at bats, this graph has a much stronger positive linear trend and much higher coefficient of determination of 0.99 in comparison to 0.37 in the first graph.

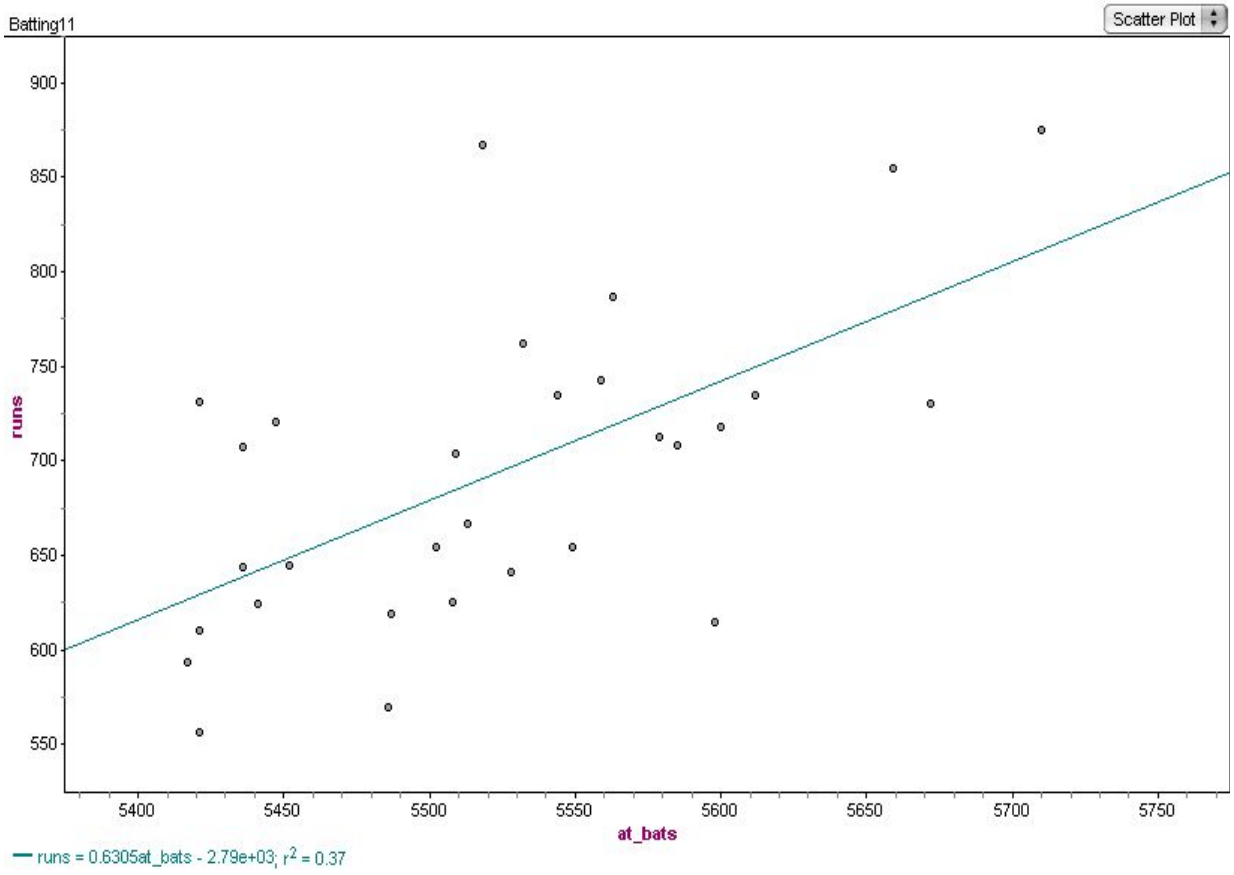




Question 7)

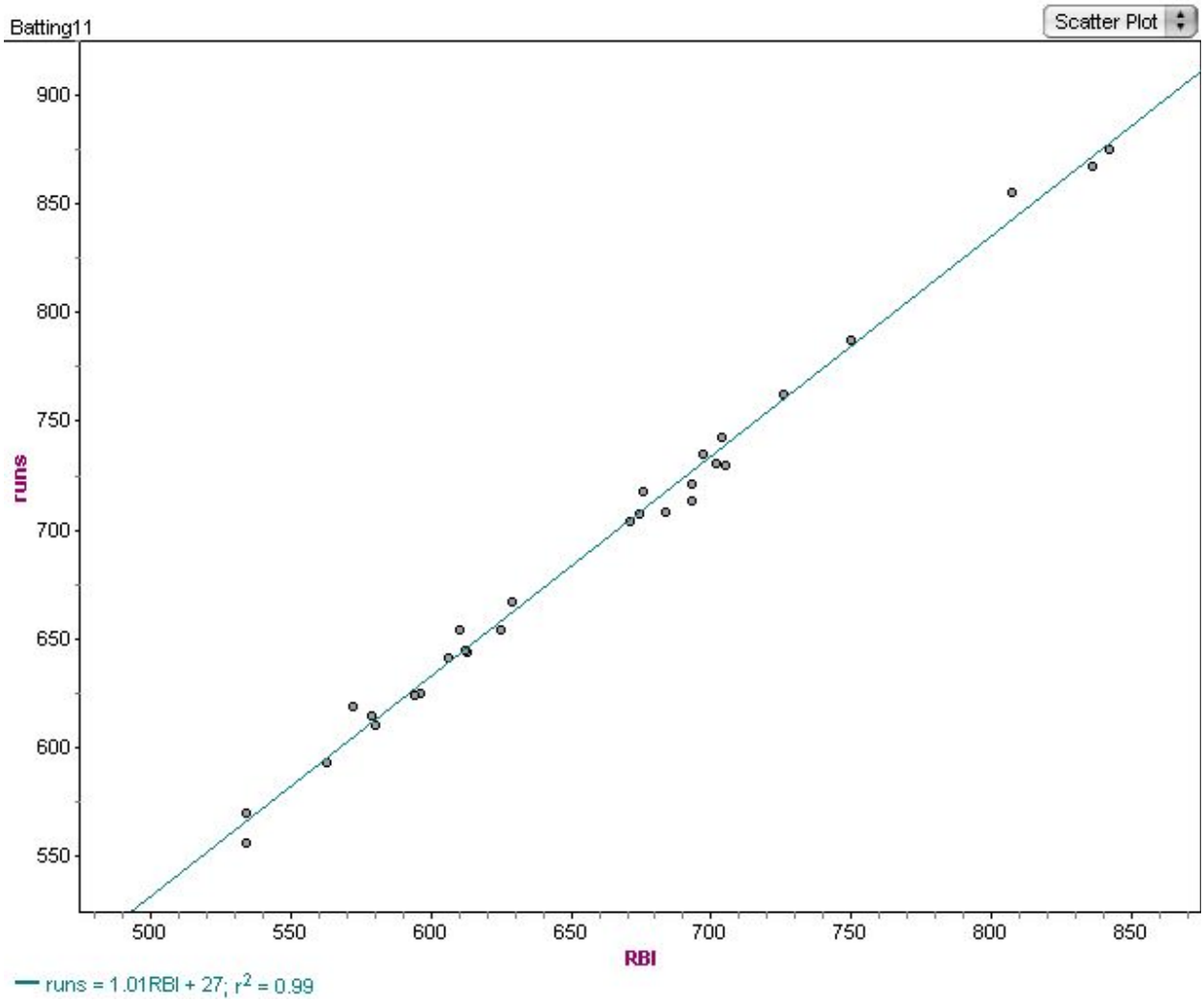
Slope is the Coefficient of determination multiplied by (standard deviation of y variable divided by standard deviation of x variable). Thus, if we know the slope, we can predict the number of runs based off of x variable if we know the equation of the regression line (the slope plus the y intercept). For the first graph, the slope was 0.63 so we know that for every 1 bat, it corresponds to 0.63 runs plus the y intercept, allowing us to predict the success of the team in terms of number of runs. For the second graph, the slope is 1.01 for the second graph which suggests an almost 1:1 ratio meaning every 1 RBI corresponds to a 1.01 increase in the number of runs plus the y intercept..

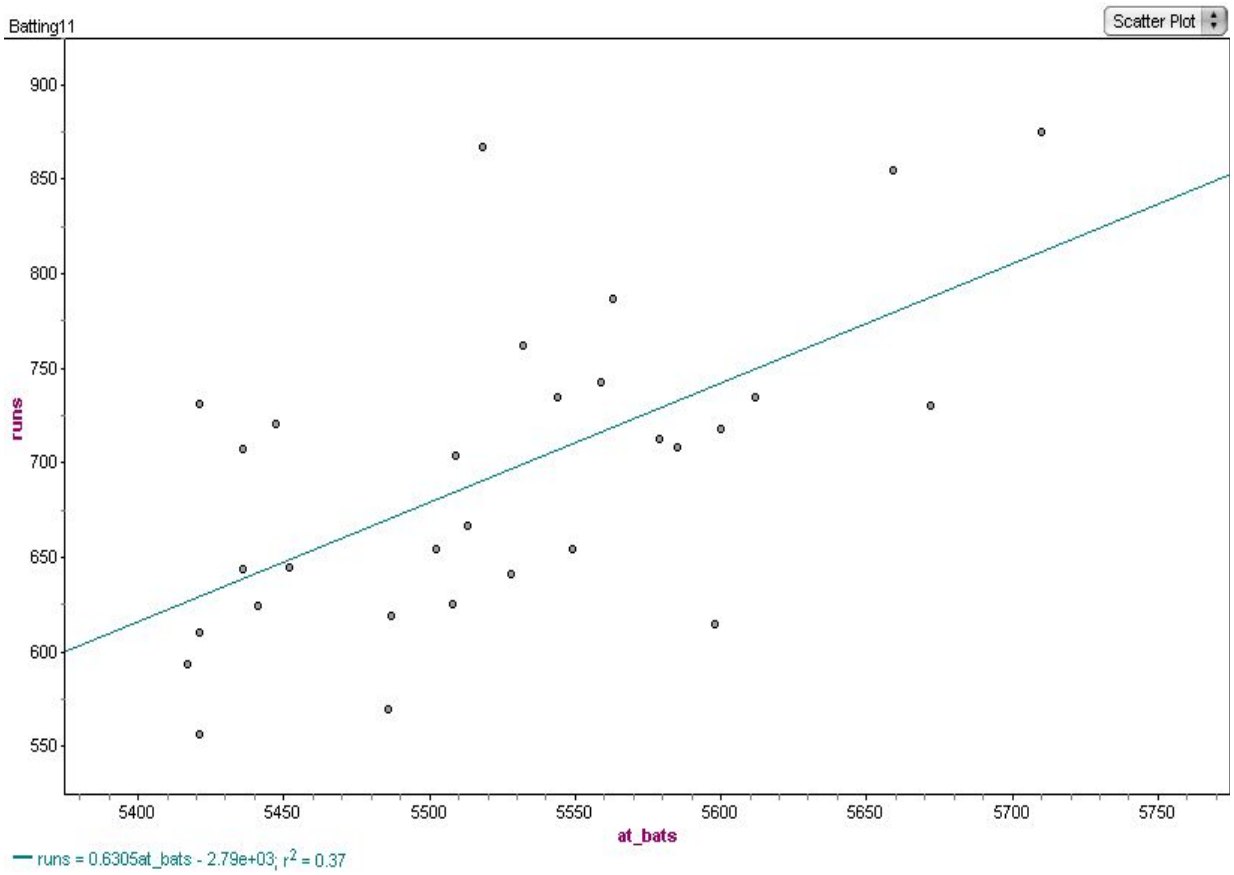

Batting11
Scatter Plot
runs
RBI
900
850
800
750
700
650
600
550
500
550
600
650
700
750
800
850
runs = 1.01RBI + 27; r2 = 0.99


Batting11
Scatter Plot
runs
at_bats
900
850
800
750
700
650
600
550
5400
5450
5500
5550
5600
5650
5700
5750
runs = 0.6305at_bats - 2.79e+03; r2 = 0.37

Question 8)

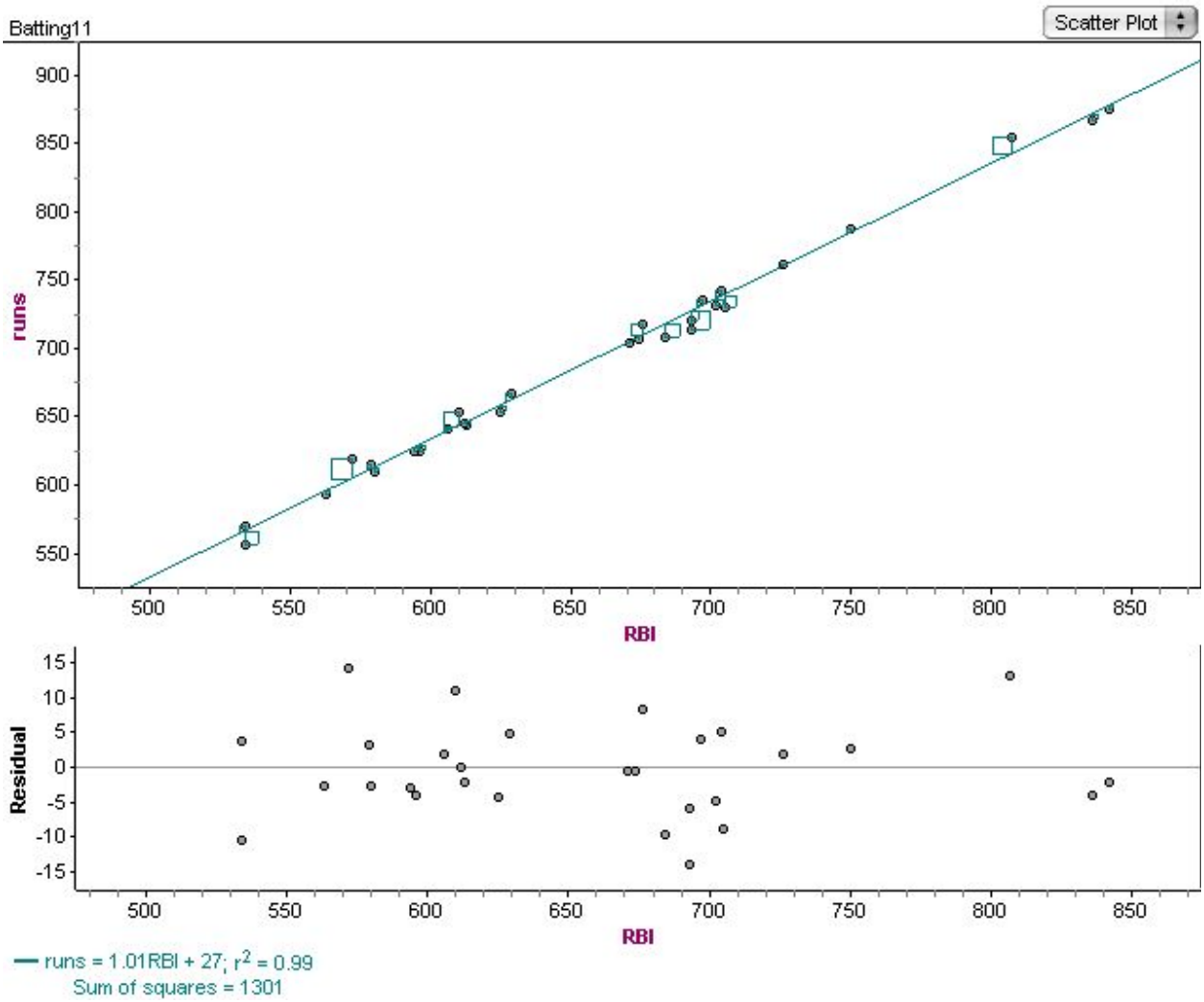
The r^2 value for my second graph with x variable RBI has a value of 0.99 which is much closer to 1 than r^2 value of 0.37 which is the number for the first graph with the x variable at_bats. The r^2 value represents the correlation coefficient, which is percentage of variance explained by x variable. Thus, my r^2 variable seems better at predicting runs than at_bats since it is closer to 1 as the coefficient of determination better predicts the y variable the closer it is to a value of 1 and less when it is closer to a value of 0.





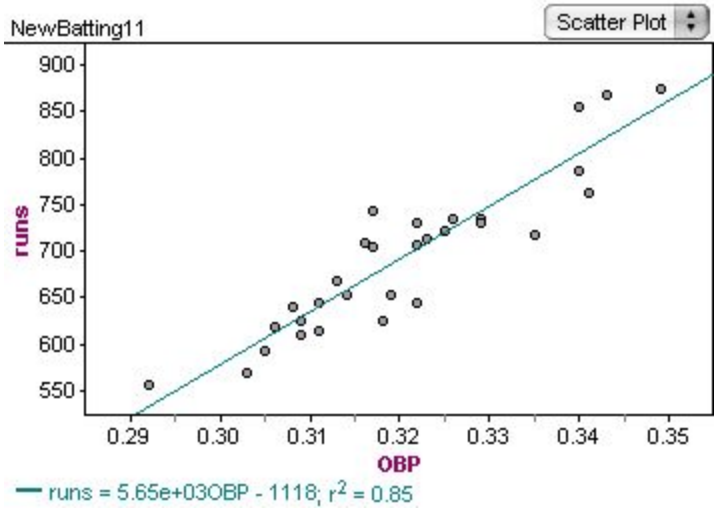
Question 9)

The variable that best predicts runs is RBI as it has the highest coefficient of determination with runs in comparison to all other variables when I tried making graphs (I didn't paste all the graphs that I tried since there were so many) using all the variables in fathom. The graph has a positive linear trend, has a very small sum of square value of 1301 in comparison to the first graph which had a value of 123700. Also, based on the residual plot, we see that the graph is linear due to scattered points and a lack of curves in the line.This variable predicts runs very well due to a strong coefficient of determination closest to the value 1(0.99). This variable matches what I Initially thought in question 6 as this was the variable I used in question 6. I am not surprised since runs batted in are usually the most common way to get runs as there is only a few other ways to get a run(such as getting hit by the pitcher with the ball, which should be very rare).

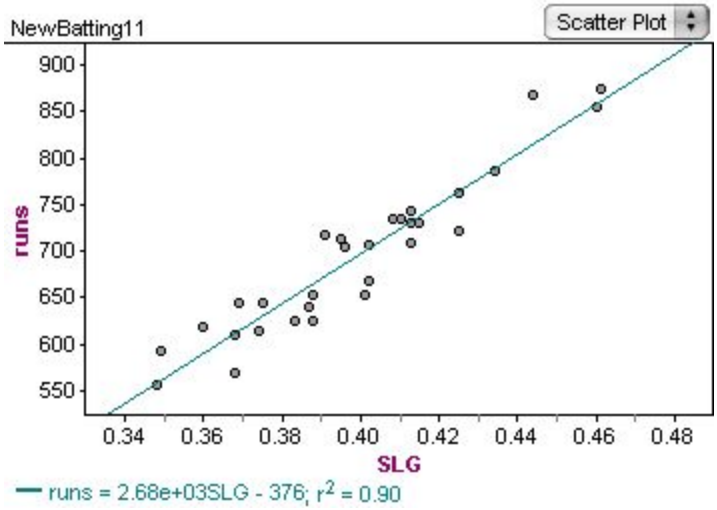


Question 10)

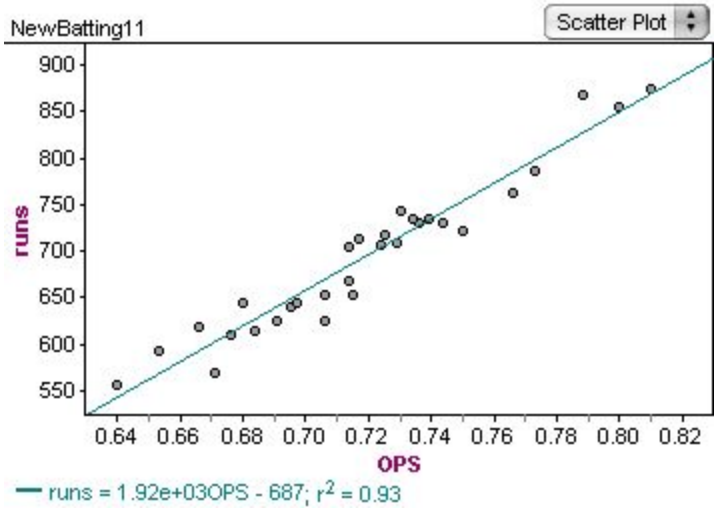
The newer variables in NewBatting11 Collection highlighted that the researchers were successful in finding better variables in predicting number of runs scored. The Three variables all had a coefficient of determination higher than 0.8, meaning a strong coefficient of determination. Of all the variables analyzed, RBI still seems to be the best predictor of runs as it has the coefficient of determination closest to 1 which is 0.99 which is higher than 0.85, 0.9 and 0.93 in the new variables. This result makes sense as RBI should be the variable that best predicts runs since runs batted in are usually the most common way to get runs as there is only a few other ways to get a run(such as getting hit by the pitcher with the ball, which should be very rare). However, based solely on the new variables, the OPS variable best predicts the number of runs since it has the coefficient of determination closest to 1 compared to all other new variables. This makes sense since slugging percentage or the total number of bases / at_bats and on base percentage should correlate with number of runs as getting to a base usually means you run to it unless by the off chance you get hit by the pitcher(which should be very rare).



runs = 5.65e+03OBP - 1118; r^2 = 0.85



runs = 2.68e+03SLG - 376; r^2 = 0.90



runs = 1.92e+03OPS - 687; r^2 = 0.93