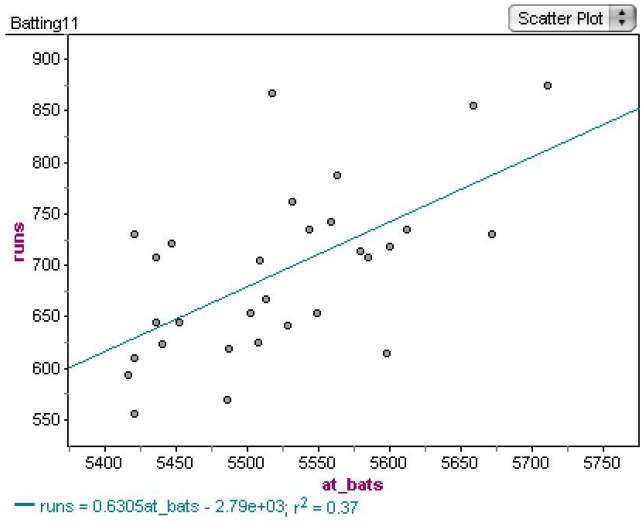
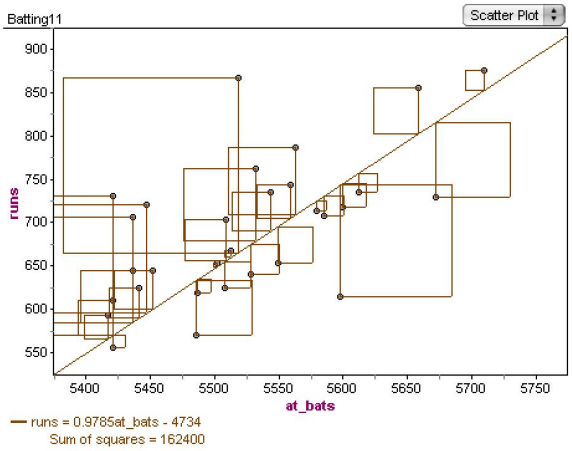
Yiqiao Jin
UID: 305107551
Section 1C

1.

Based on the linear regression line, there is only a weak positive linear trend between at bats and number of runs; thus there is only a weak positive correlation between at_bats and runs.

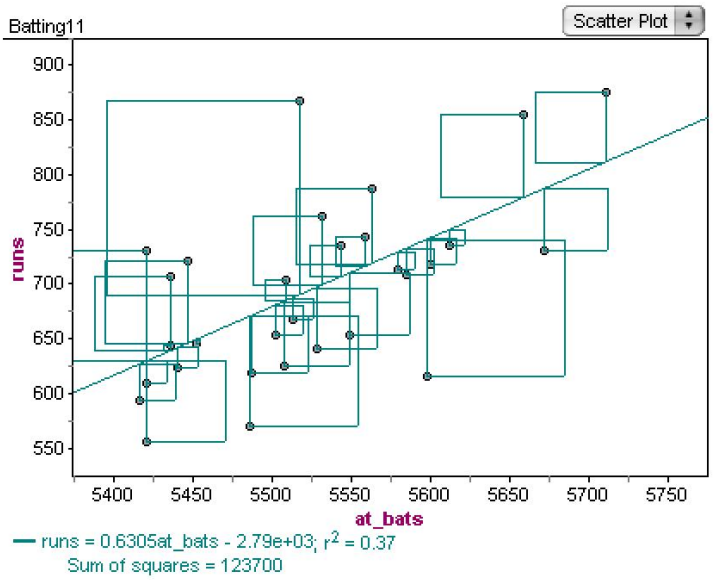


3.



The sum of squares will be smaller when the line better fits the data.
This is because sum of squares is calculated by adding up the squared differences between the actual y value and the predicted y value by the regression line.

4.



The sum of square is smaller for this least square line.

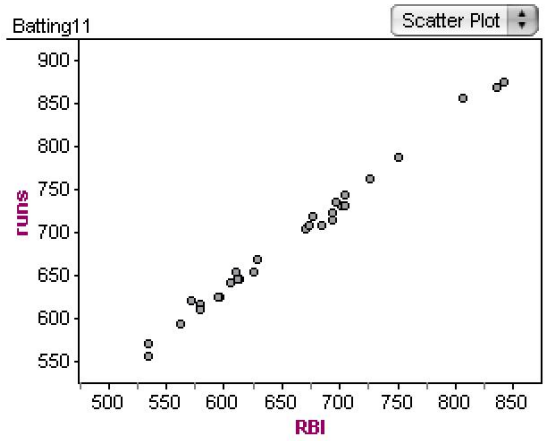
5.
Based on the data, the predicted runs for a team that had 5508 at_bats would be:
0.6305*5508-2790 = 682.794

The actual number of runs for the four teams are:
Chicago White Sox: 654 runs
Cleveland Indians: 704 runs
Florida Marlins: 625 runs
L.A Angles: 667 runs

So the differences between the predicted and the actual values are calculated as:
L.A Angles: 667 - 682.794 - = -15.794
Florida: 625 - 682.794 = -57.794
Chicago White Sox: 654- 682.794 = -28.794
Cleveland Indians: 704 - 682.794 = 21.206

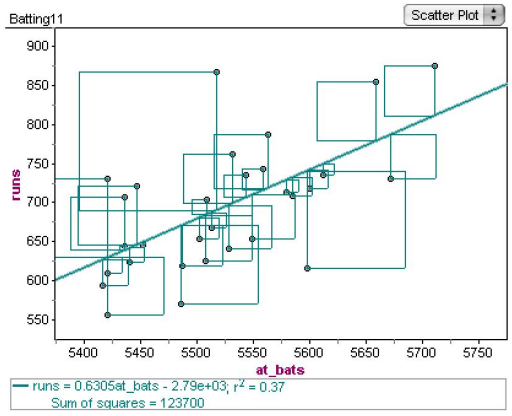
Thus, the variance, or the sum of squares, would be 15.794^2 + 57.794^2 + 28.794^2 + 21.206^2 = 4868.386

6.
I believe that Runs Batted In will have the lowest sum of squares using the least square line, because 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. The relationship appears to be linear. 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.

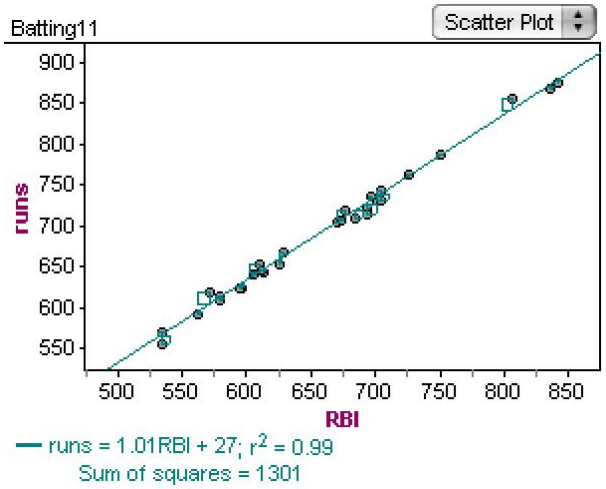


7.

Slope is coefficient of determination * (standard dev of y / standard dev of x)
More specifically, slope = r*(sy/sx).
The slope is 0.63. It means for every 1-unit increase of at_bats, we would expect number of runs increase by 0.63. With the slope and and y-intercept, we can derive the predicted y value (in this case, number of runs) with an x value(in this case, the value of at_bats)



8.



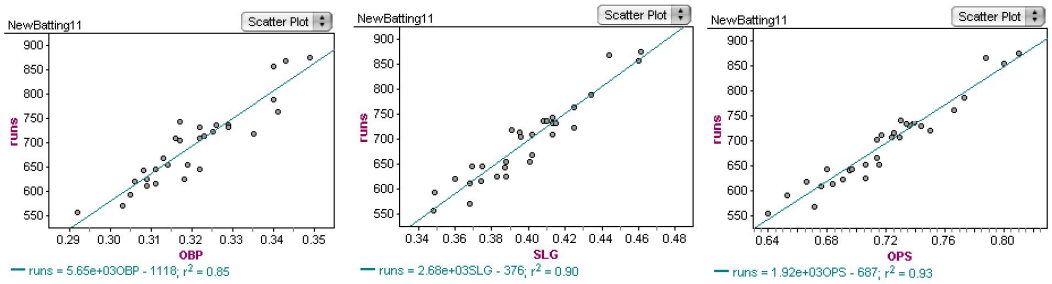
r^2 is coefficient of determination. It represents that r^2 of the variance can be explained

by x. The variable I choose - RBI - seems to better predict runs than at_bats because It has an r^2=0.99, which is greater than 0.37, the coefficient of determination of at_bats. It means 99% of variance in number of runs can be explained by RBI

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. It has a sum of square that equals 1301, which is significantly lower than 123700, the sum of square of at_bats. Based on the residual plot, I observe that the relation is linear. The variable RBI 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).

10.



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 of 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). 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.

Summary

Concepts including slope, intercept, coefficient of determination, correlation coefficient, linear regression line (or least square line) are all covered in the lab as well as homework. I have seen these concepts in the textbook, lectures and homework.