Bivariate Regression

Open the *Baseball.txt* data set, a collection of batting statistics of 331 baseball players who played in the American League in 2002. Suppose we are interested in whether there is a relationship between batting average and the number of home runs a player hits. Some fans might argue, for example, that those who hits lots of home runs also tend to make a lot of strike outs, so that their batting average is lower. Let us check it out, using a regression of the number of home runs against the player's batting average (hits divided by at bats). Because baseball batting averages tend to be highly variable for low numbers of bats, we restrict our data to set of those players who had at least 100 at bats for the 2002 season. This leaves us with 209 players. Use this data set for answering the following:

1. Construct a scatter plot of *home runs* versus *batting average*.
2. Informally, is there evidence of a relationship between the variables?
3. What would you say about the variability of the number of home runs, for those with higher batting averages?
4. Perform a regression of *home runs* on *batting averages*. Obtain a normal probability plot of the standardized residuals from this regression. Does the normal probability plot indicate acceptable normality, or is there skewness? If skewed, what type of skewness?
5. Construct a plot of the residuals versus the fitted values. What pattern do you see? What does this indicate regarding the regression assumptions?
6. Take the natural log of *home runs*, and perform a regression of *in home runs* on *batting average*. Obtain a normal probability plot of the standardized residuals from this regression. Does the normal probability plot indicate acceptable normality?
7. Construct a plot of the residuals versus the fitted values. Do you see strong evidence that the constant variance assumption has been violated?
8. Write the population regression equation for the model. Interpret the meaning of the values of β_0 and β_1.
9. State the regression equation (from the regression results) in words and numbers.
10. Interpret the value of the *y*-intercept b_0.
11. Intercept the value of the slope b_1.
12. Estimate the number of *home runs* (not *in home runs*) for a player with a batting average of 0.300.
13. What is the size of the typical error in predicting the number of *home runs*, based on the player's *batting average*?
14. What percentage of the variability in the *in home runs* does *batting average* account for?
15. Perform the hypothesis test for determining whether a linear relationship exists between the variables.
16. Construct and interpret a 95% confidence interval for the unknown true slope of the regression line.
17. Calculate the correlation coefficient. Construct a 95% confidence interval for the population correlation coefficient. Interpret the result.
18. Construct and interpret a 95% confidence interval for the mean number of home runs for all players who had a batting average of 0.300.
19. Construct and interpret a 95% confidence interval for a randomly chosen player with a 0.300 batting average. Is this prediction interval useful.
20. List the outliers. What do all these outliers have in common? For Orlando Palmerio, explain why he is an outlier.