HW 9: Regression Diagnostics

Instructions: Work must be shown to receive full credit. You may work with others on the homework, but you must write and turn in your own copy. **This does not mean that you can simply copy someone else's work!!** Also, make sure your homework is neat, stapled, and all answers are written in complete sentences!! Come and see me if you have any questions.

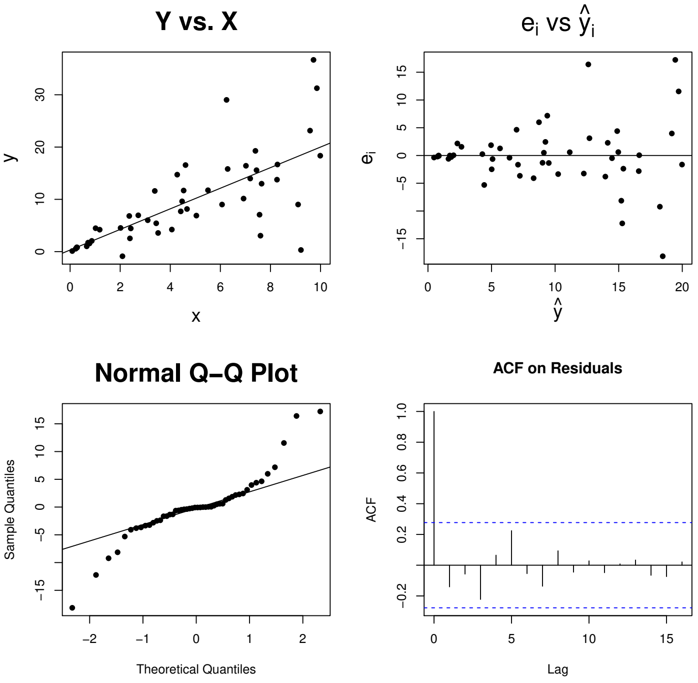
On problems that require the use of R, PLEASE give me the RELEVANT R code and output to for each problem so I can assess partial credit. I may take off for including unnecessary R output. If one problem refers back to output from another problem, make sure to cite that output in your answer. Incorrect one-sentence answers will get little or no credit.

NOTE: If a problem asks you to perform a hypothesis test, make sure to give the hypotheses, test statistic, p-value, and a conclusion in the terms of the problem. Also, if the problem asks you to perform a confidence interval, make sure to interpret the confidence interval.

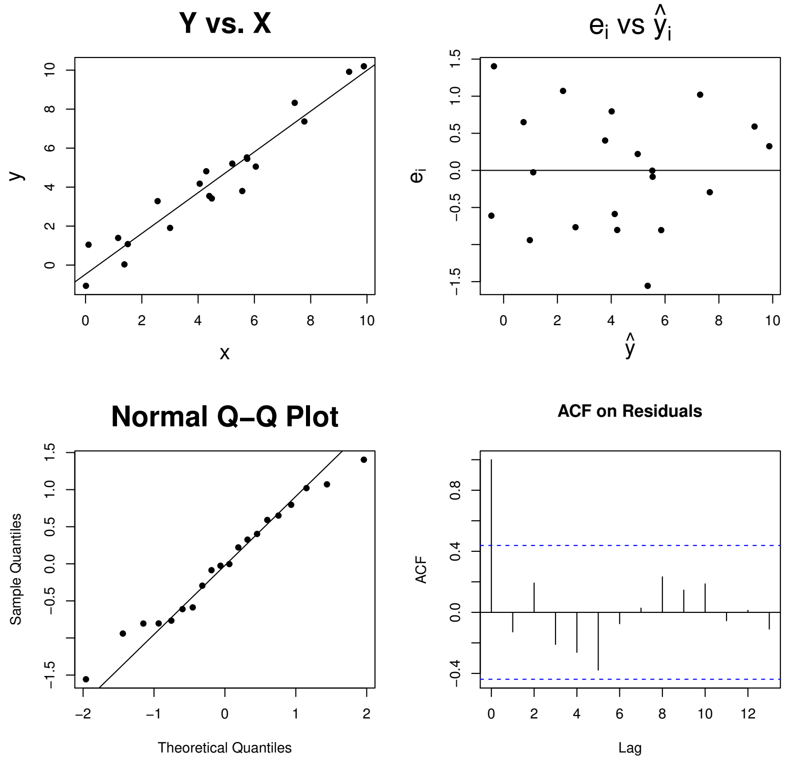
"By Hand" Problems: For hypothesis tests, you may use R to find the p-value. For confidence intervals, you may use R to find the multiplier.

1. Given each set of output, describe any assessments that can be made regarding appropriateness of the model, constant variance, normality, and independence. Clearly state which plot(s) you are looking at the make each conclusion. If there are violations, state some recommendations for addressing those violations.

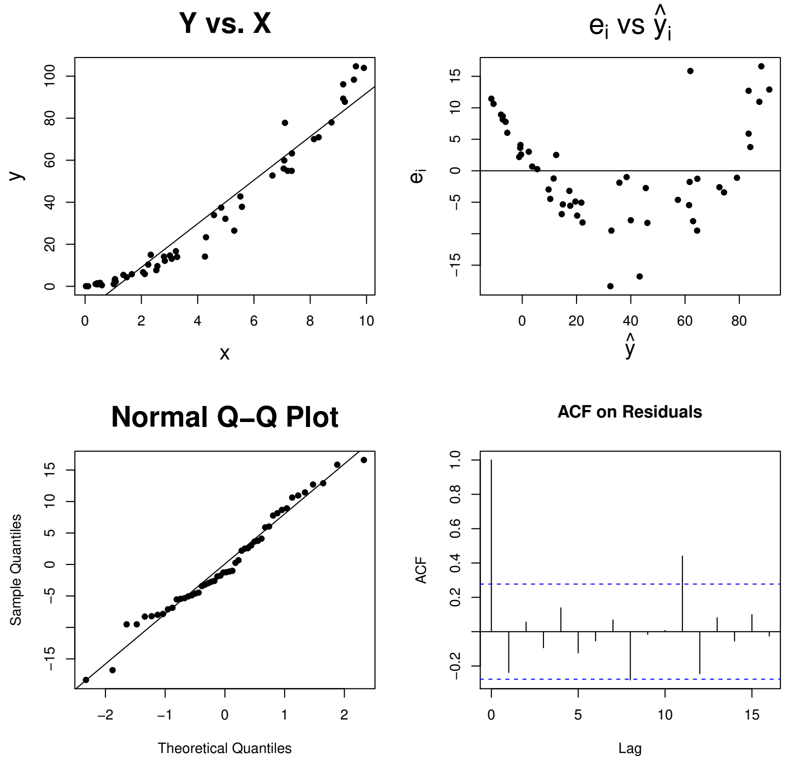
   (a) **Scenario A**



(b) Scenario B



(c) Scenario C



“R” Problems:

2. During WWII baseball was still played, though many players served in the military. Suppose we are interested in predicting batting averages towards the the end of the war (`BA44`) using the prior season’s batting averages (`BA43`). The data are provided on Moodle in the `ww2baseball.xlxs` file.

   (a) Using `R`, load the data and obtain the regression line for predicting 1944 batting averages using 1943 batting averages.

   (b) Using the `R`, obtain a well labeled scatterplot of the data and overlay the regression line.

   (c) Using `R`, obtain a scatterplot of the residuals versus fitted values with a horizontal line at zero.

   (d) What can we tell about our model assumptions based on the plots in (b) and (c)? Be specific.

   (e) Using `R`, obtain a normal-QQ plot. What can we conclude from the plot?

   (f) Using `R`, obtain an autocorrelation plot (we will assume time order is known for practice). What can we conclude from the plot?

3. Dr. Fribance teaches MCSI classes at CCU. One of her classes in particular involves collecting data on the boat. It is well known that increasing temperatures allow for increased conductivity in water. It is important to study this in different settings and types of water (ex. Hayashi, 2004). Using the student collected data consisting of 55 observations, what kind of conclusions can we draw on the effects of temperature on water conductivity? From the Moodle page, load the `FribanceStation1.csv` data set collected from Station 1 (south of Georgetown) in 2017. The data are in fact presented in time order. Thoroughly assess the assumptions of simple linear regression. Clearly provide all relevant `R` code and output used in addressing each assumption. Be organized and complete in your response!