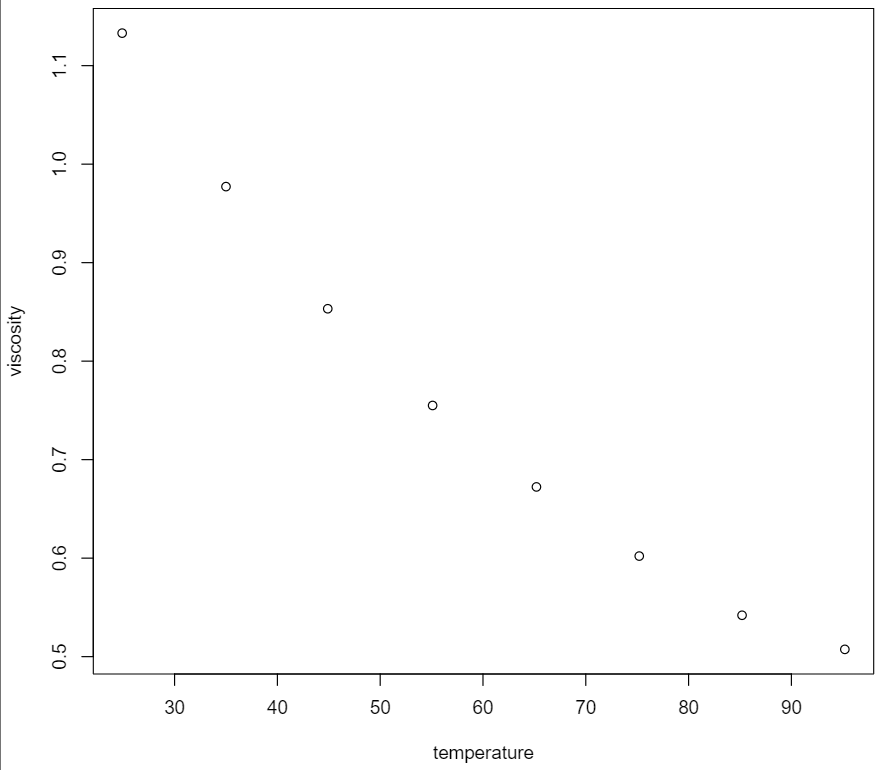
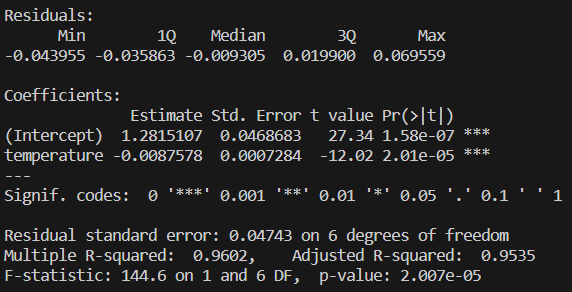
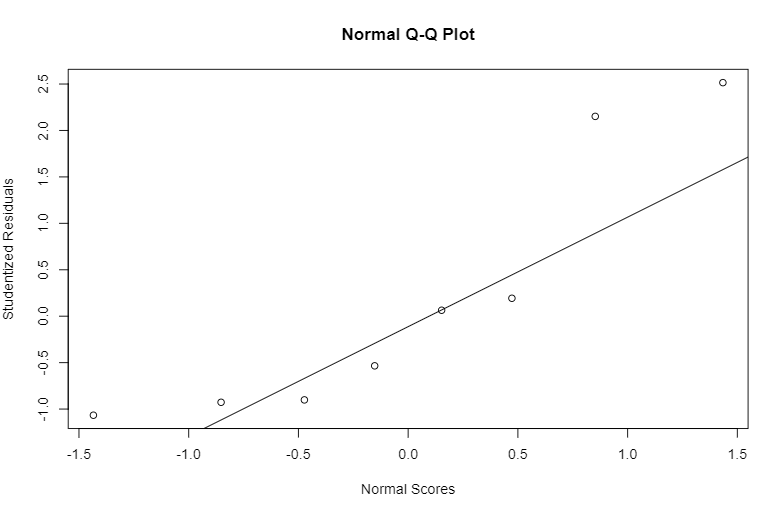
STAT641 Regression Analysis

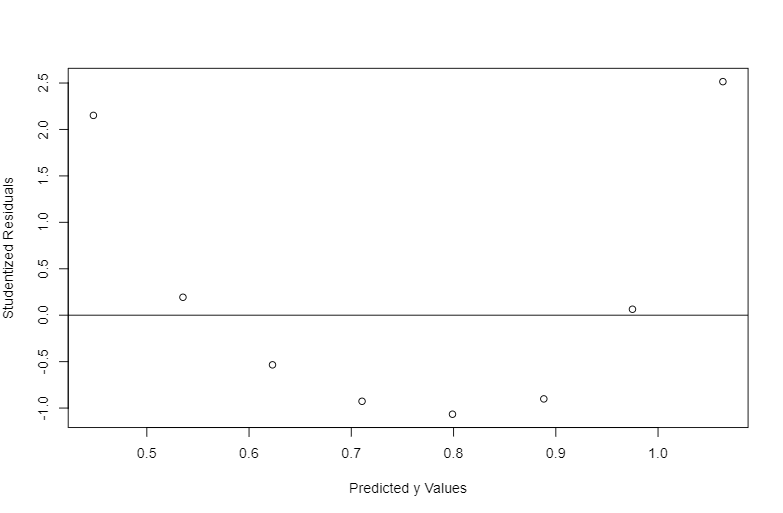
Homework #5

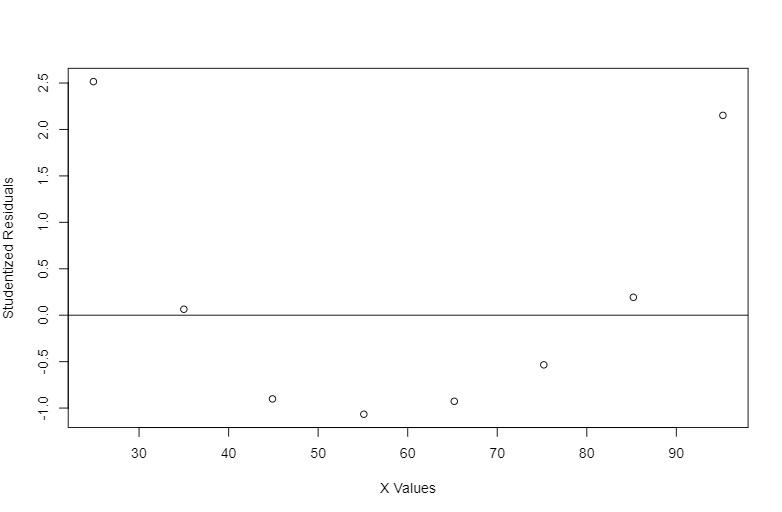
1. 

A linear model seems like it could be a good rough estimate for the data within this range of X, but a better estimate at a higher degree of polynomial could probably be found.

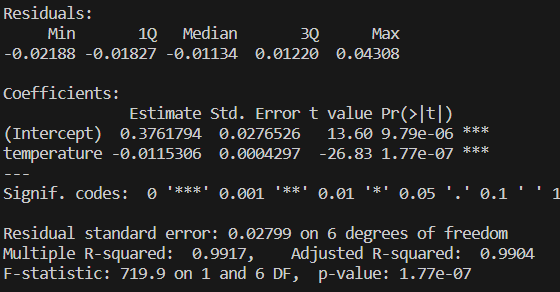
1. 

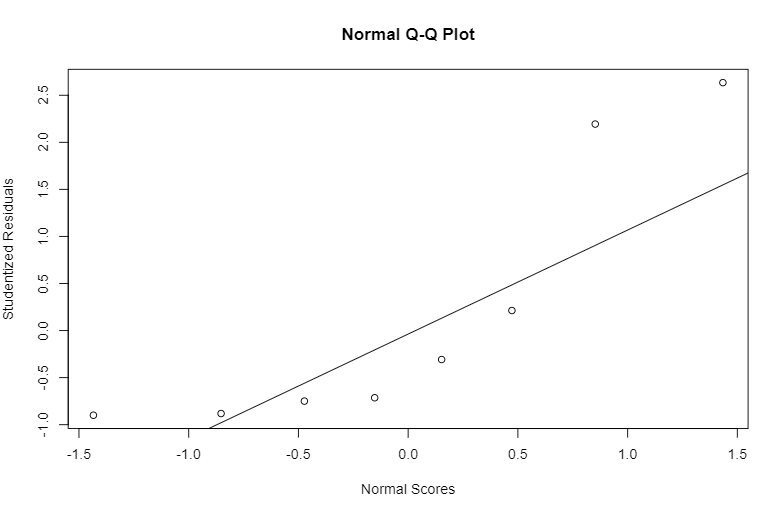


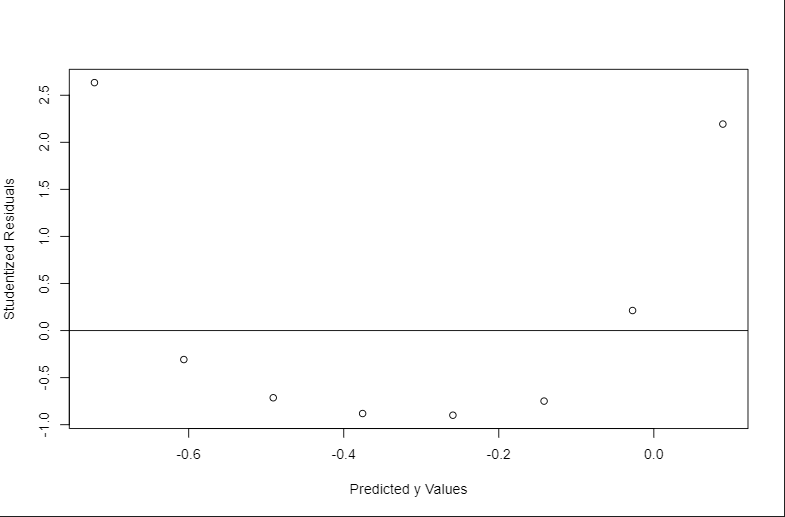


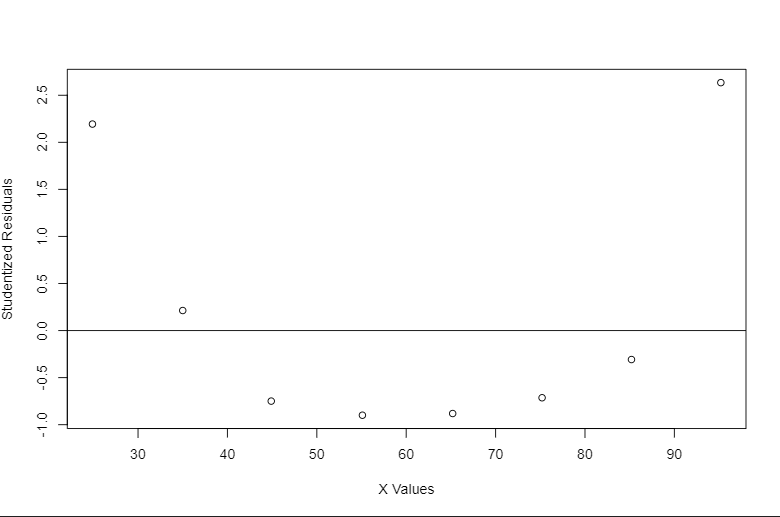


Although the summary statistics suggest a good fit, the residual plots undeniably indicate that a linear model is inadequate for this data.

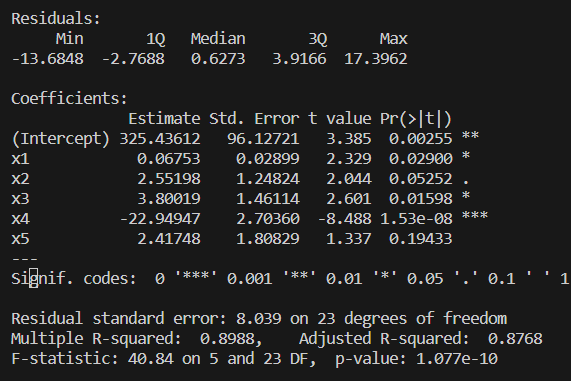
1. 

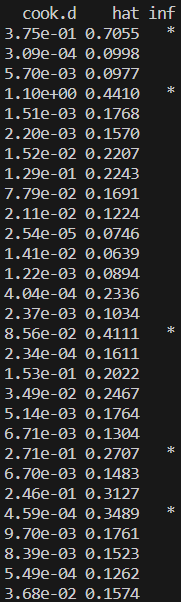
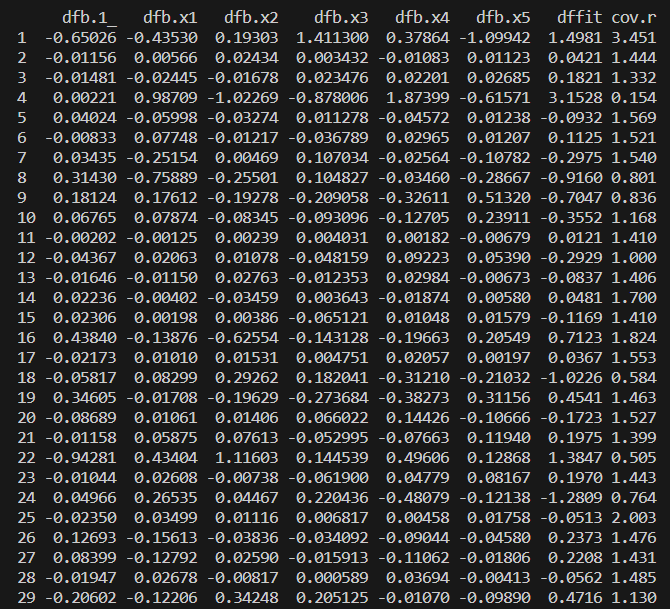






The summary statistics for the exponential model suggest a better fit (especially based on R2, but, again, the residual graphs show an undeniable pattern that indicate that the exponential model is also not a good fit for the data.

1. 



Based on the model that incorporates all of the data (even the regressors that are determined to not contribute significantly to the model), there are 5 influential points in the data: 1, 4, 16, 22, and 25.

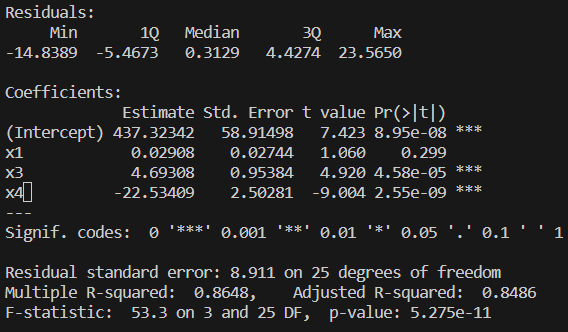
Point 1 is identified to be influential based on its hat value, several of its DFBETAS values, and its DFFITS value.

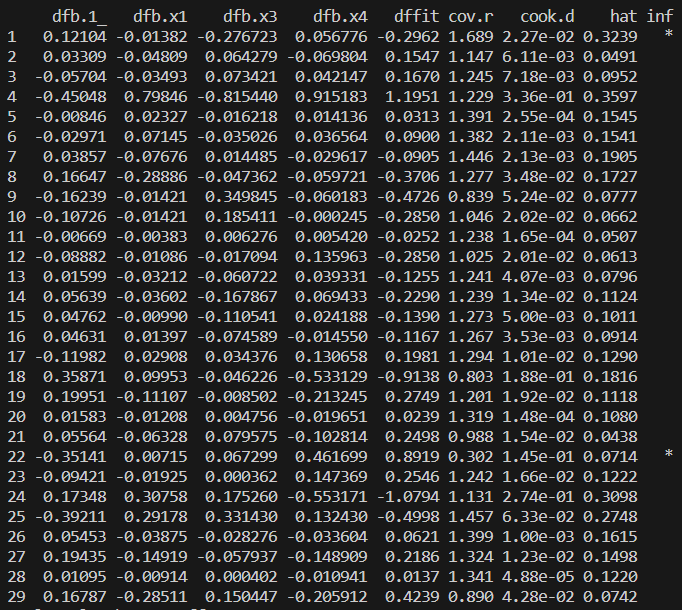
Point 4 is identified to be influential based on its hat value, its Cook’s D value, several of its DFBETAS values, and its DFFITS value.

Point 16 is identified to be influential based on only two of its DFBETAS values.

Point 22 is identified to be influential based on several of its DFBETAS values and its DFFITS value.

It is unclear to me why point 25 is identified to be influential as the hat value, Cook’s D value, all DFBETAS values, and DFFITS value all fall within the cutoffs identified within the textbook.





Based on the model that only incorporates the regressors that were originally determined to contribute significantly to the model, only points 1 and 22 are determined to be influential points in the data.

Point 1 is identified to be influential based on its hat value and its DFFITS value.

Point 22 is identified to be influential based on one of its DFBETAS values and its DFFITS value.

I intend to focus on environmental data. I will analyze water quality data pertaining to 20 reservoirs over the course of 31 years (1987-2018) published by the US EPA and associated with the following publication:

Smucker, N., J. Beaulieu, C. Nietch, and J. Young. Increasingly severe cyanobacterial blooms and deep water hypoxia coincide with warming water temperatures in reservoirs. GLOBAL CHANGE BIOLOGY. Blackwell Publishing, Malden, MA, USA, 27(11): 2507-2519, (2021).

All the data can be found at <https://catalog.data.gov/dataset/1987-2018-cyanobacteria-and-water-quality-data-for-20-reservoirs> in 13 excel files. The response variable that I will focus on will be cyanobacteria density (cells/mL) and the regressors that I will focus on will be nutrient density (units vary depending on the nutrient) and surface temperature (°C).