Homework 6

STAT 462 (Fall 2020)

Clearly label your answers to each question and each sub-question. Your answers MUST be uploaded to Canvas as a <HWx_Yourfirtname.nb.html> file by the deadline.

1. Many different interest groups such as the lumber industry, ecologists, and foresters benefit from being able to predict the volume of a tree just by knowing its diameter. One classic data set (shortleaf.txt) reported by C. Bruce and F. X. Schumacher in 1935 concerned the diameter (in inches) and volume (in cubic feet) of 70 shortleaf pines.

A researcher is interested in learning about the relationship between the diameter and volume of shortleaf pines.

(i). Identify the response variable and explanatory variable for the problem

(ii). Draw a scatter plot to show how volume of a tree and its diameter are associated. Comment on your observations. Provide any outputs you might have used.

(iii). Fit a regression line for the problem, write down the estimated equation (define any terms you might have used), and mark the estimated line on the scatter plot in part (2). Provide all outputs. Interpret the estimated parameters clearly in the problem context.

(iv). Obtain suitable diagnostics for the estimated model in part (iii). Clearly state your observations. Provide any outputs you might have used.

(v). Identify (a) the point with highest residual, (b) the point with highest leverage, and (c) the point with highest cook's distance. Suppose a friend of the researcher suggested that there is an influential point in the data, and should be investigated. Do you agree with this comment? Explain your reasoning.

(vi). Another friend of the researcher suggested, perhaps the diagnostics observed in part (iv) is an artifact of lack of linearity between the two variables of interest. He proposed the researcher should transform the variable(s) and re-fit the model. Using the discussions we had in the class about possible starting points for transformations, how would you proceed (i.e. would you transform the response variable, explanatory variable or both? What transformations would you use?)? Clearly explain.

(vii). Use log transformation on the explanatory variable, and re-draw the scatter plot in part (i). Refit the regression model using the transformed data, and mark the estimated line on the scatter plot. Comment on your observations. Provide any outputs you might have used. Does the transformation appear to have improve the linearity?

(viii). Use log transformation on the response variable as well. Obtain a scatter plot between the two transformed variables. Refit the regression model using the transformed data, and mark the estimated line on the scatter plot. Comment on your observations. Provide any outputs you might have used. Does the transformation appear have to improve the linearity?

(ix). Obtain suitable diagnostics for the estimated model in part (viii). Clearly state your observations. Provide any outputs you might have used.

(x). Identify the point with highest cook's distance (using results in part ix). Refit the model in part (viii) excluding the identified point. Obtain suitable diagnostics and comment on your observations. Provide any outputs you might have used.

(xi). The researcher decided to proceed with the model obtained in part (x). Write down the estimated regression line (define all your terms), interpret the estimated parameters. Provide any outputs.

(xii). Test the for the strength of the linear relationship between two variables.