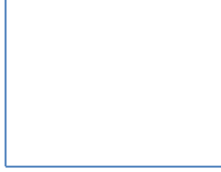


HW 2

1. Draw data sets for the following scenarios:

a. A point that has large leverage but does not have a large effect on the fitted model.



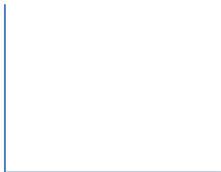
b. A point that has small leverage but has a large effect on the fitted model.



c. A point that has large leverage but has a large effect on the fitted model.



d. A point that has small leverage but does not have a large effect on the fitted model.



2. It has been theorized that developing countries cut down their forests to pay off foreign debt. The data on debt, deforestation, and population are in the included sas code.

(a) Convert this scientific question into a statistical one by positing a multiple regression model. Include the formal notation for the model here.

(b) Examine the assumptions for the model you stated in (a). Discuss whether they seem to be met or not.

(c) If the assumptions are not satisfactorily met, iterate through altering the model and checking it for assumptions. After you are done, state the form for the final model here (if it is the same one stated in (a), then just copy and paste it here).

(d) Using the model from (c), check for any high leverage points. If there are any, discuss whether they are worrisome in your analysis.

(e) Write up a scientific conclusion and scope of inference answering the original scientific question (note that the phrasing of the scientific question is somewhat vague and could be operationalized in a few different ways, some of which are not possible to answer with this data set).

3. Taking the data set from 2., add a new “observation” to the data set that would have high leverage in the main effects model of debt and population but would not extreme in neither debt nor population marginally. Record the observation and its (approximate) diagnostic values here (Cooks D, leverage, studentized residual, dfbetas).

4. In class exercise. Record your answers below. Make sure you completely answer each item below, including plots, a discussion of what you did, and statistical conclusions.
- a) Does metabolism differ by gender?
 - b) Do differences in gender persist when controlling for gender?
 - c) Do we need to account for alcoholism?

a)

- i. Write down the multiple regression model for this question. Estimate this model.
- ii. Write down the model as a two-sample t-test. Estimate this model.
- iii. Are the assumptions met? If not, what can we do instead? Interpret your results.

b)

- i. What does a large Cook's D indicate?
- ii. What does a large leverage indicate?
- iii. What does a large dfbeta indicate?
- iv. Fit the interaction model with gender and gastric, using "Male" as the reference category
- v. Use the above diagnostics to produce some plots to identify any influential points. Use them to decide how to proceed. Decide on a model for answering question b).

c) Compare the fully interactive model for (gender gastric) to the fully interactive model for (gender gastric alcoholic) to decide if alcoholic should be included.