Assignment 4

1. Cook's distance is defined as:

   $$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{pMS_{Res}}.$$

   Show that

   $$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}.$$

2. Consider the data in the files assign4_i.csv and assign4_ii.csv.
   a. For the data in assign4_i.csv, using the unit length scaling, what are the standardized regression coefficients?
   b. Are the estimates of the coefficients for x_1 and x_2 independent? Why or why not?
   c. Report the VIFs.
   d. For the data in assign4_ii.csv, using the unit length scaling, what are the standardized regression coefficients?
   e. Are the estimates of the coefficients for x_1 and x_2 independent? Why or why not?
   f. Report the VIFs.
   g. The responses for the two data sets were generated from the same model, but the x's are different (i.e., the designs). Which of the two designs is preferable? Why?

3. **Power study:** An important task for statisticians is to help experimenters decide upon sample sizes for an experiment. This is most frequently done by simulation. Suppose an experimenter has decided on the settings for a two-variable problem, and these are the same as the x's in assign4_i.csv. The experimenter views regression coefficients of 2 (or more) to be of practical interest. Suppose it is known from previous studies that the process experimental standard deviation is σ=20. Using a significance level of α=0.05, find the number of replicates of the experiment design (i.e., repeats of the nine x settings) that is needed to detect **both** variables as significant, if they both have an effect size of 2, with power of 0.80 (sometimes called 80% power). (Recall, power = probability of rejecting the null hypothesis, when the null hypothesis is false).

4. An experiment was performed to investigate the amount of a drug that has been absorbed by the kidneys of a lab's guinea pigs. Twenty guinea pigs were randomly selected. Because larger animals were thought to potentially absorb more of a given does, the actual dose given was approximately proportional to the body weight of the animal (the dose was difficult to administer, so the target proportion was only approximated). The response variable was the percentage of the dose of the drug that was actual administered to the guinea pig that was absorbed by the kidneys. The data can be found in guinea_pig.csv.

    a. Construct a *pairs plot*. (i) Comment on the relationship between the predictors. Do you observe any issues? (ii) Comment on the relationship between the predictors and the response variable.
    b. Fit a linear regression model to these data. Test the appropriate hypotheses to see if there is evidence that the three predictors are linearly related to the response. For which of the predictors, if any, is there evidence of a significant linear relationship with the response variable?
    c. Assess the model adequacy by constructing residual plots (use the studentized residuals) for you plots. Summary your findings with respect to the assumptions for the linear regression model.
    d. Compute the leverages for the observations. Do any of the observations have a large h_{ii} ?
    e. Construct the Cook's distances. (i) Do any of the observations have a relatively large Cook's distance? (ii) Looking at the results of this question as well as parts b, and c., are any of the observations potentially influential.
    f. Is there evidence of collinearity?
    g. Refit the model without the most influential observation. For which of the predictors, if any, is there evidence of a significant linear relationship with the response variable? What has changed and why?