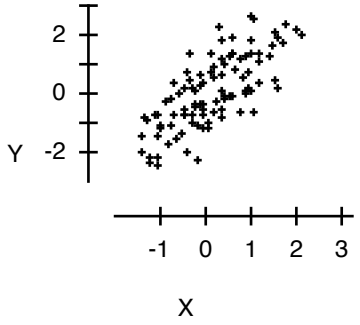
1. Residual Plots (8 points)

(a) Sketch a picture of a residual plot that shows constant variance and linearity.
(b) Sketch a picture of a residual plot that shows non-constant variance and linearity.
(c) Sketch a picture of a residual plot that shows constant variance and nonlinearity.
(d) Sketch a picture of a residual plot that shows non-constant variance and nonlinearity.

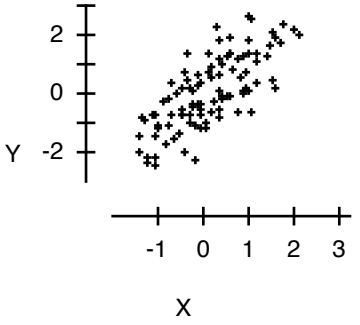
2. Leverage and Cook's Distance (8 points)

On the scatterplots below, sketch a new point or points meeting the stated criteria.

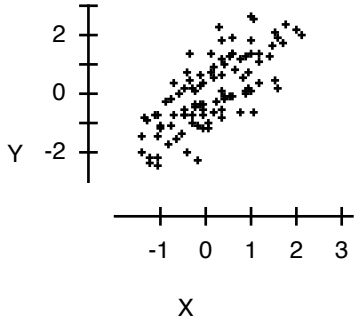
(a) A point with high leverage and high Cook's distance.



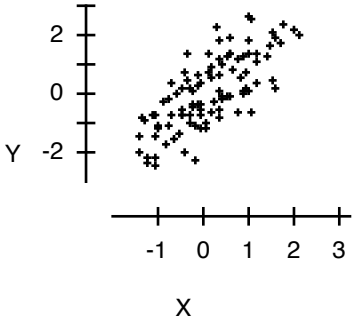
(b) A point with high leverage but low Cook's distance.



(c) A point with neither high leverage nor high Cook's distance.



(d) Two points that are influential when taken together, even though neither alone has high Cook's distance.



3. Short Answer (8 points)

For each of the following statements, indicate whether the statement is true or false, and explain your answer briefly.

(a) A predictor X_3 can be statistically significant in a multiple regression even though the correlation of Y and X_3 (by itself) is .10.

(b) A predictor X_3 can be statistically *not* significant in a multiple regression even though the correlation of Y and X_3 (by itself) is .80.

(c) If your dataset has an outlier, transforming Y to log(Y) can reduce the leverage of this point.

(d) If your dataset has an outlier, transforming Y to log(Y) can reduce the Cook's distance of this point.

4. Used Airplane Prices (22 points)

During the 1980s, US production of small light aircraft declined dramatically; production of new piston-powered single-engine aircraft dropped 91% from 1979 to 1989. As a result, there is an active market for used aircraft of this category. The purpose of this assignment is to use data on this market to develop a model that can predict and explain the asking price of a particular model of aircraft: the Warrior, manufactured by Piper Aircraft.

The Piper Warrior is a four-passenger airplane with either a 150- or 160-horsepower engine. Introduced in the mid 1970s, it is a modification of the earlier Piper Cherokee.

What factors affect the price of a used Piper Warrior? Some factors might include the age of the airframe (i.e., the aircraft body, excluding instruments and engine), the number of hours the aircraft has been flown (a measure of wear on the engine, among other things), and the time since the engine has been overhauled or replaced. Older airplanes may require a new paint job; for an airplane this can cost several thousand dollars.

A Warrior can be equipped with a wide range of features that affect its price. These include radios and electronic navigational and piloting instruments, also called avionics. In new aircraft, the difference between airplanes equipped with basic avionics compared to one fully equipped with the latest high tech gadgets can easily exceed the cost of the aircraft itself. Two particular kinds of avionics that could represent considerable investments over the basic package are abbreviated DME (distance measuring equipment) and LORAN (long-range navigation based on satellite communication.

The data here were transcribed from an issue of *Trade-A-Plane*, a national publication advertising used aircraft for sale, in the early 1990s. The variables recorded include the `year` of the aircraft, `TT` (total flight time in hours), `SMOH` (hours since major overhaul), `DME`, `LORAN`, `HP` (engine horsepower), `paint` (new or recent paint job), and `price`. Not all information may be available for all aircraft. The variables `DME`, `LORAN`, and `paint` are categorical variables that indicate whether the corresponding item was mentioned as being present in the ad. The `price` is given in thousands of dollars.

(a) Read in the dataset from Moodle using the following commands:

```
# read in dataset
setwd("~/Fill in your pathname here")    # set working directory
data <- read.csv("airplanes.csv")
```

There should be 25 rows in the dataset; if you get a 26th row of missing values, delete it using the command `data <- data[-26,]`. Now define all the variables we will need:

```
# define variables
price <- data[,"price"]
year <- data[,"year"]
TT <- data[,"TT"]
SMOH <- data[,"SMOH"]
LORAN <- data[,"LORAN"]
DME <- data[,"DME"]
HP <- data[,"HP"]
paint<- data[,"paint"]
```

(b) Make a scatterplot matrix of the entire dataset using the `pairs` command. Notice that several of the variables are categorical, so scatterplots involving them are not informative. Re-do the scatterplot matrix using only the quantitative variables and hand it in (the `cbind` command may be helpful here). Comment on any notable features or patterns (or lack of thereof).

(c) What is notable about the variable `HP`? What should you do with it?

(d) Which plane appears to be an outlier? Can you guess why it might be an outlier? For the purposes of this assignment, let's delete this point from any further analyses since we don't know its correct value. Go ahead and delete this row from `data`. *Important*: now that you've deleted a row, you need to re-define all the variables accordingly. Re-issue the commands under "`define variables`" from part **(a)**. Make a new scatterplot matrix of the quantitative variables and hand it in.

(e) Fit a model with all the X variables included and save it as `fit1`. What is the value of R-squared? What is the conditional SD? What is the interpretation of the latter quantity? Which variables appear significant, and which do not? Hand in the output from the `summary(fit1)` command.

(f) Notice that `SMOH` has several missing values. Given that these data are taken from advertisements, why do you think these values might be missing? What might the actual values be? Make an educated guess as to what the values might be and fill them in, and re-run your model in part **(e)** but call it `fit2`. Hand it in these results. Is there any substantial change in the results?

(g) Let's go ahead and remove `SMOH` from the model. Re-run the model without `SMOH` and save it as `fit3`. All predictors should now be significant. Hand in these results.

(h) It seems plausible that `TT` and `year` might be collinear: older planes may have been flown more. Based on the scatterplot matrix, does this appear to be true?

(i) We should also check whether `DME`, `LORAN`, and `paint` are collinear. However, these variables are categorical, so you can't just make a scatterplot. Can you think of a way to check if these variables are associated? Hand in your output and describe what you find.

(j) Make an added-variable plot of each predictor variable. To do this, first install the car package. (If you are using regular R, go to the `Packages & Data` menu, choose `Package Installer`, and type `car` into the search box. Select `car` from the search results and click on `Install Selected`. If you are using RStudio, select `Packages` in the lower right window, and then click on the `Install` tab. Type `car` into the `Packages` box and click `Install`.) Now type `library(car)` to load the package into your R session. Then use the `avPlots(fit3)` command to make added-variable plots of each predictor, and hand in the plots. Do the added-variable plots support keeping all five predictors in the model?

(k) For our current model, do the regression assumptions appear to be satisfied? Make a residual plot (residuals vs predicted), and an NPP with the reference line added. Hand in these plots and comment on whether you think the assumptions are satisfied or whether there may be cause for concern.

(l) Summarize your findings from this model. In particular, what do the regression coefficients mean? What is the conditional SD, and what does it represent? What is the R-squared?