MLDS 401/IEMS 404-1 (Fall 2023): Lab 3 – 10/10/2023

4.5 **(Research expenditures data)** Refer to Exercise 3.11 on modeling research expenditures of the top 30 engineering schools using the number of faculty and the number of PhD students as predictor variables. The two scatter plots are shown in Figures 4.11 and 4.12 with each data point labeled by the abbreviated name of the university. Identify the outliers and influential observations in the data using appropriate diagnostic statistics. Provide plausible explanations for why these universities are flagged.

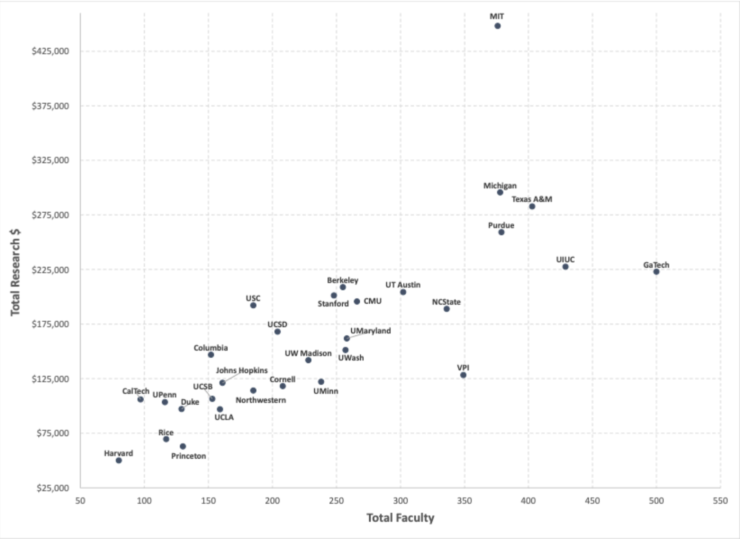

Figure 4.11 Plot of research expenditures (in millions of dollars) versus number of faculty for the top 30 graduate engineering programs

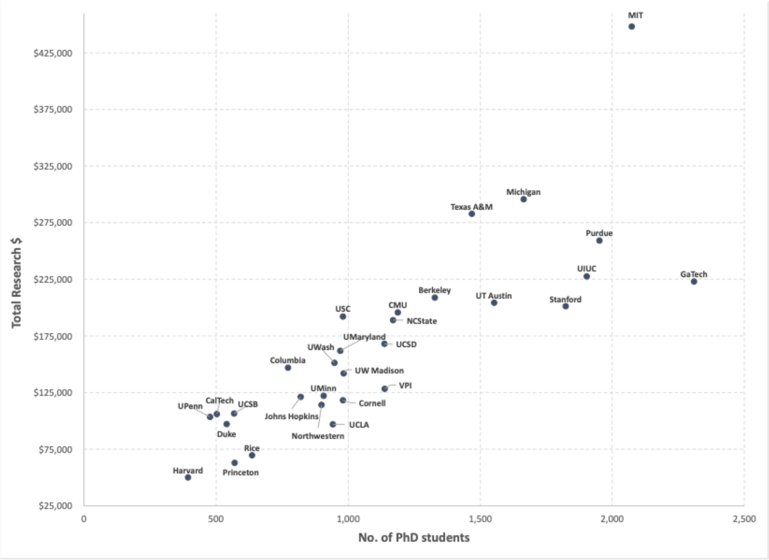

Figure 4.12 Plot of research expenditures (in millions of dollars) versus number of PhD students for the top 30 graduate engineering programs

4.8 (Woodbeam data: Influential observations) Table 4.4 gives data on the specific gravity (x_1), moisture content (x_2) and strength (y) of wood beams.

a) Make a scatter plot of the two predictors. Which observation appears to be influential?

b) Is that observation influential using the leverage rule $h_{ii} > 2(p+1)/n$ and Cook's distance rule $D_i > f_{p+1,n-(p+1),\alpha}$ for $1-\alpha = 0.10$ (i.e., 10% confidence ellipsoid).

c) Fit the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ from all data, and compare it with the fit obtained after omitting the influential observation. Does the fit change much?

Table 4.4 Woodbeam strength data

Observation No.	Specific Gravity	Moisture Content	Strength	Observation No.	Specific Gravity	Moisture Content	Strength
1	0.499	11.1	11.14	6	0.528	9.9	12.60
2	0.558	8.9	12.74	7	0.418	10.7	11.13
3	0.604	8.8	13.13	8	0.480	10.5	11.70
4	0.441	8.9	11.51	9	0.406	10.5	11.02
5	0.550	8.8	12.38	10	0.467	10.7	11.41

Source: Draper and Stoneman (1966), Table 1.

4.10 **(Multivariate linear dependency)** Webster, Gunst and Mason (1974) gave the data shown in Table 4.5 on four predictor variables. The data are constructed such that x_1, x_2, x_3, x_4 add up to 10 in all 12 observations except the first one where they add up to 11. Thus there is an almost exact linear dependence among the four observations.

Table 4.5 Data illustrating multivariate linear dependence

No.	x_1	x_2	x_3	x_4	No.	x_1	x_2	x_3	x_4
1	8	1	1	1	7	2	7	0	1
2	8	1	1	0	8	2	7	0	1
3	8	1	1	0	9	2	7	0	1
4	0	0	9	1	10	0	0	0	10
5	0	0	9	1	11	0	0	0	10
6	0	0	9	1	12	0	0	0	10

Source: Webster, Gunst and Mason (1974), Table 1.

a) Calculate the correlation matrix among the four predictors. Check that no correlation exceeds 0.5 in absolute value. Thus correlations do not provide an indication of multicollinearity.

b) Calculate the four VIFs and check that they all exceed 150 with the maximum equal to 289.23. Thus VIFs indicate serious multicollinearity.

4.11 (Gas mileages of cars: Multicollinearity and variance stabilizing transformation) The file `mpg.csv` contains data on gas mileages (`mpg`) of 392 cars and their number of cylinders, piston displacement, horsepower, weight and acceleration. We want to build a predictive model for `mpg` based on these five predictors.

a) Calculate the correlation matrix between the five predictors. Does it indicate presence of multicollinearity in the data?

b) Fit a full regression model with all five predictors. How is multicollinearity reflected in this fit?

c) To improve the model suppose we drop `displacement` from the above model since it is least significant. In what ways does the resulting model improve upon the full model?

d) Make the normal Q-Q plot and the fitted values plot for the above fit. Note that the spread of the residuals in the fitted values plot is roughly proportional to the square of the fitted values, which suggests the inverse transformation. Use the inverse transformation `gp100m = 100/mpg` (the gallons of fuel per 100 miles) as the dependent variable and rerun the regression and make the normal Q-Q and the fitted values plots. Has this transformation helped to remove the flaws of the previous model? Does this transformation change VIFs?

e) Calculate the estimated mpg of a car with 6 cylinders, 105 HP, 3000 lbs weight and 15 miles/sec^2 acceleration using the above fitted model.

4.12 (Acetylene data: Multicollinearity statistics) Table 4.6 gives data from Marquardt and Snee (1975) on conversion of n-heptane to acetylene (y) as a function of three reaction conditions: reactor temperature (x_1), ratio of H_2 to n-heptane (x_2) and contact time (x_3).

The following full second degree model is to be fitted to the data:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \\ &\quad + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \varepsilon. \end{aligned}$$

a) Plot the three predictor variables against each other. Also calculate the correlation coefficients between them. Do you see any indications of multicollinearity?

b) Calculate the VIFs for all the terms in the above model. Comment on your results.

c) Center x_1, x_2, x_3 by subtracting the mean of each predictor from its values. Compute the remaining terms (pairwise products and squares) from these centered values. Now calculate the VIFs for all the terms. Compare the results with those from (b). Has centering made the multicollinearity problem less severe?

Table 4.6 Acetylene data

x_1 Reactor Temperature (°C)	x_2 Ratio of H_2 to n-heptane (mole ratio)	x_3 Contact Time (sec)	y Conversion of n-heptane to Acetylene (%)
1300	7.5	0.0120	49.0
1300	9.0	0.0120	50.2
1300	11.0	0.0115	50.5
1300	13.5	0.0130	48.5
1300	17.0	0.0135	47.5
1300	23.0	0.0120	44.5
1200	5.3	0.0400	28.0
1200	7.5	0.0380	31.5
1200	11.0	0.0320	34.5
1200	13.5	0.0260	35.0
1200	17.0	0.0340	38.0
1200	23.0	0.0410	38.5
1100	5.3	0.0840	15.0
1100	7.5	0.0980	17.0
1100	11.0	0.0920	20.5
1100	17.0	0.0860	29.5

Source: Marquardt and Snee (1975)