3.13 **(Research expenditures data)** Research expenditures is an important factor in the algorithm used by *US News & World* to rank graduate engineering programs. It carries 25% weight (15% for total research expenditures and 10% for research expenditures per faculty). The file `Research.csv` gives data on research expenditures in millions of $ (Research), number of faculty (Faculty) and number of PhD students (PhD) in top 30 US Universities according to *US News & World* 2017 rankings. The data are taken from ASEE profiles. We want to build a predictive model for research expenditures as a function of number of faculty and number of PhD students.

a) Make a matrix scatter plot and compute the correlation matrix of all three variables. Comment on the relationships between the variables.

b) Fit a regression model of Research versus Faculty and PhD. From this model note that PhD is a significant predictor of Research but Faculty is not. Why can't research expenditures be increased simply by increasing the number of PhD students? Given that faculty with more grants fund more PhD students (i.e., the causal arrow is Faculty $\rightarrow$ PhD) explain the apparently anomalous result obtained.

c) Calculate the partial correlation coefficients between Research and each predictor controlling for the other predictor and their t-statistics. Check that these t-statistics are the same as those given by the regression analysis.

3.14 **(Sales data)** Consider the following data on sales (y) of a company in 10 sales regions. The predictors are: the number of salesmen (x_1) and the amount of sales expenditures in millions of dollars (x_2).

Table 3.10 Salary Data Variables

Variable	Explanation
Salary	Annual salary in $
YrsEm	No. of years employed with the company
PriorYr	No. of years of prior experience
Educ	No. of years of education after high school
Super	No. of people supervised
Gender	M = Male, F = Female
Dept	Advertising, Engineering, Purchase, Sales

Source: McKenzie and Goldman (1999, Temco Data Set)

No.	x_1	x_2	y	No.	x_1	x_2	y
1	31	1.85	4.20	6	49	2.80	7.42
2	46	2.80	7.28	7	31	1.85	3.36
3	40	2.20	5.60	8	38	2.30	5.88
4	49	2.85	8.12	9	33	1.60	4.62
5	38	1.80	5.46	10	42	2.15	5.88

Source: Tamhane and Dunlop (2000), Example 11.7.

a) Calculate the correlation matrix $\boldsymbol{R}$ between x_1 and x_2 and the correlation vector $\boldsymbol{r}$ between y and x_1, x_2. From these bivariate correlations calculate the partial correlations $r_{yx_1|x_2}$ and $r_{yx_2|x_1}$.

b) Calculate the standardized regression coefficients $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ from $\boldsymbol{R}$ and $\boldsymbol{r}$ for the model. How do they compare with the partial correlation coefficients $r_{yx_1|x_2}$ and $r_{yx_2|x_1}$?

c) Check that you get the same values for $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ from the unstandardized LS estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ by scaling them appropriately. Calculate $\widehat{\beta}_0$.

d) Compare $(\widehat{\beta}_1, \widehat{\beta}_2)$ with $(\widehat{\beta}_1^*, \widehat{\beta}_2^*)$. Which variable is a better predictor of sales and why?

4.4 (College GPA and entrance test scores: Checking normality and homoscedasticity) Refer to Example 3.16 in which we fitted the model

$$\text{GPA} = \beta_0 + \beta_1 \text{Verbal} + \beta_2 \text{Math} + \beta_3 \text{Verbal}^2 + \beta_4 \text{Math}^2 + \beta_5 \text{Verbal} \times \text{Math} + \varepsilon.$$

The regression coefficients are given in the R output in that example.

a) Make the normal and fitted values plots of residuals. Comment on why the normality and especially the homoscedasticity assumptions seem to be violated. Does the fitted values plot suggest the log transformation of GPA?

b) Fit the same model using log(GPA) as the response variable. Make the normal and fitted values plots of residuals. Are the normality and homoscedasticity assumptions satisfied?

4.5 (Research expenditures data) Refer to Exercise 3.13 on modeling research expenditures of 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 4.12 and 4.13 with each data point marked 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.

4.6 (Employee salaries: Checking normality and homoscedasticity) Refer to Exercise 3.15. Using the same predictors that were found significant in part (a) of that exercise fit two regressions, one using Salary as the response variable and the other using log(Salary) as the response variable.

a) Make normal plots for residuals from both regressions. Has the log transformation of Salary improved normality?

b) Make fitted values plots for both sets of residuals. Has the log transformation of Salary improved homoscedasticity?