STAT 3008
Exercise 1

Problems refer to the problem sets in the textbook: Applied Linear Regression, 3rd edition by Weisberg.

1. Problem 1.2.
   Mitchell data
   The data shown in Figure 1.12 give average soil temperature in degrees C at 20 cm depth in Mitchell, Nebraska, for 17 years beginning January 1976, plotted versus the month number.

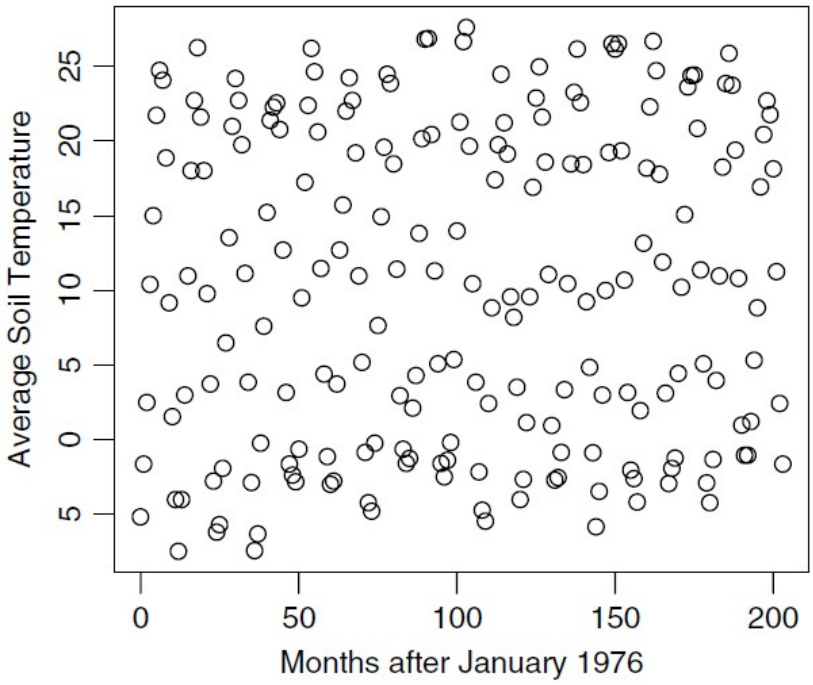

FIG. 1.12 Monthly soil temperature data.

1.2.1. Summarize the information in the graph about the dependence of soil temperature on month number.

1.2.2. The data used to draw Figure 1.12 are in the file Mitchell.txt. Redraw the graph, but this time make the length of the horizontal axis at least four times the length of the vertical axis. Repeat Problem 1.2.1.

2. Problem 1.3.
   United Nations
   The data in the file UN1.txt contains PPgdp, the 2001 gross national product per person in US dollars, and Fertility, the birth rate per 1000 females in the population in the year 2000. The data are for 193 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries; the third variable on the file called Locality gives the name of the locality. The data were collected from http://unstats.un.org/ unsd/demographic. In this problem, we will study the conditional distribution of Fertility given PPgdp.
   1.3.1. Identify the predictor and the response.

   1.3.2. Draw the scatterplot of Fertility on the vertical axis versus PPgdp on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be a plausible for a summary of this graph?

   1.3.3. Draw the scatterplot of log(Fertility) versus log(PPgdp), using logs to the base two. Does the simple linear regression model seem plausible for a summary of this graph?

3. For the Heights data, draw the loess smoothers with fraction $f = 0.2, 0.4, 0.6, 0.8$ on the same graph.

4. Problem 1.5.
   Water run-off in the Sierras
   Can Southern Californias water supply in future years be predicted from past data? One factor affecting water availability is stream run-off. If run-off could be predicted, engineers, planners and policy makers could do their jobs more efficiently. The data in the file water.txt contains 43 years worth of precipitation measurements taken at six sites in the Sierra Nevada mountains (labelled APMAM, APSAB, APSLAKE, OPBPC, OPRC, and OPSLAKE), and stream run-off volume at a site

near Bishop, California, labelled BSAAM. The data are from the UCLA Statistics WWW server.
In summarizing the information, give only

- Two pairs of variables that are not correlated.
- Two pairs of variables that are correlated.
- One pair of variables that do not have a constant variance function.
- One pair of variables that an outlier exists.
- One pair of variables that an influential point exists.

5. Let $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, show that

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i(y_i - \bar{y})$$