Chapter 1: Scatterplots and Regression

1 Introduction

Regression is a statistical technique for investigating and modeling the relationship between variables. It can be used to answer questions such as

- Does changing class size affect success of students?
- Do countries with higher per person income have lower birth rates than countries with lower income?
- Do changes in diet result in changes in cholesterol level?
- · · · · · ·

2 Scatterplots

Data consists of values (x_i, y_i), $i = 1, \ldots, n$, of (X, Y) observed on each of n units or cases. The goal of regression is to understand how the values of Y change as X is varied over its range of possible value. The scatterplot provides a graphical way to look at how Y changes as X is varied.

Inheritance of heights During the period 1893-1898, E.S. Pearson organized the collection of $n = 1375$

heights of mothers in the UK under the age of 65 and one of their adult daughters over the age of 18.

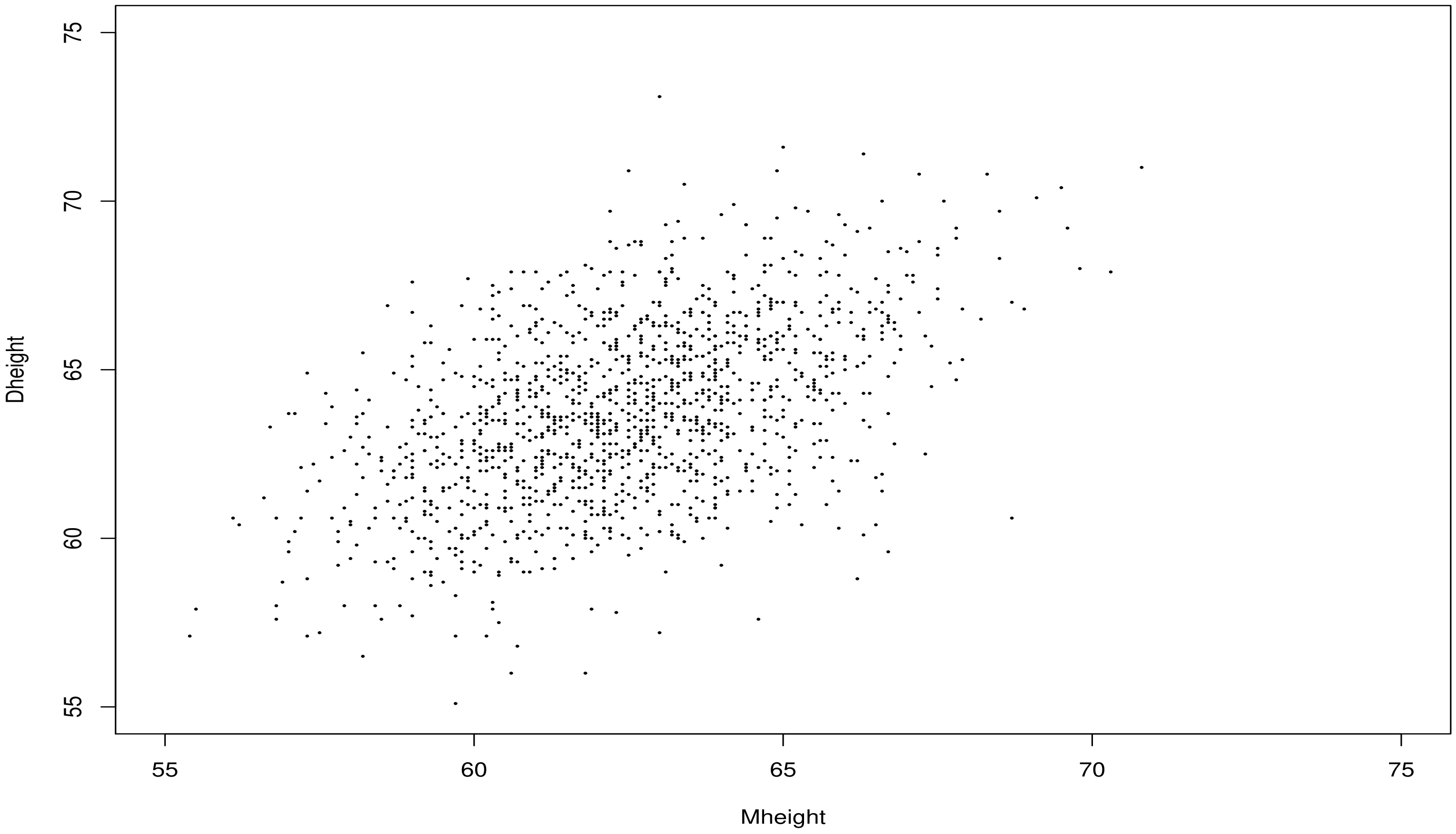

Figure 1: Scatterplot of mothers' and daughters' heights in the Pearson and Lee data.

Here are some important characteristics of Figure 1.

(1) The range of heights appear to be about the same for mothers and for daughters.

(2) Mothers’ heights and daughters’ heights are not independent.

(3) The scatter of points in the graph appears to be more or less elliptically shaped, with the axis of the ellipse tilted upward.

(4) Scatterplots are important for finding separated points, which are either points with values on the horizontal axis that are well separated from the other points or points with values on the vertical axis that, given the value on the horizontal axis, are either much too large or too small. These two types of separated points are called leverage points and outliers, respectively.

Forbes' data In an 1857 article, James D. Forbes discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water. The scatterplot of pressure versus temp is shown in Figure 1.3.

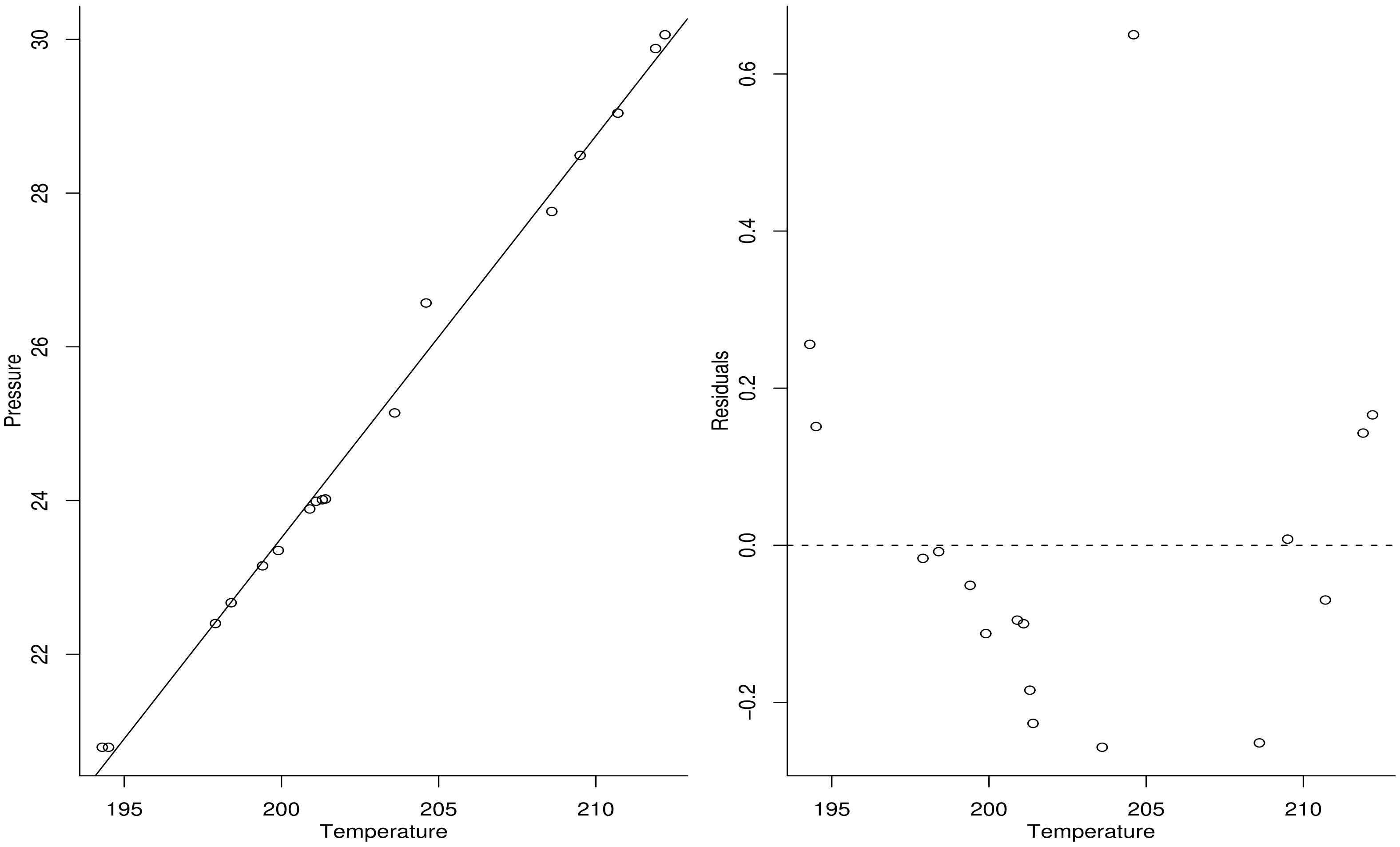

Figure 2: Forbes data

If the data is modeled by a straight line, the curvature in the residual plot is clearly visible. Forbes had a physical theory that suggested that $\log(Pressure)$ is linearly related to $Temp$. The residual plot in Figure 3(b) confirms that the derivations from the straight line are not systematic.

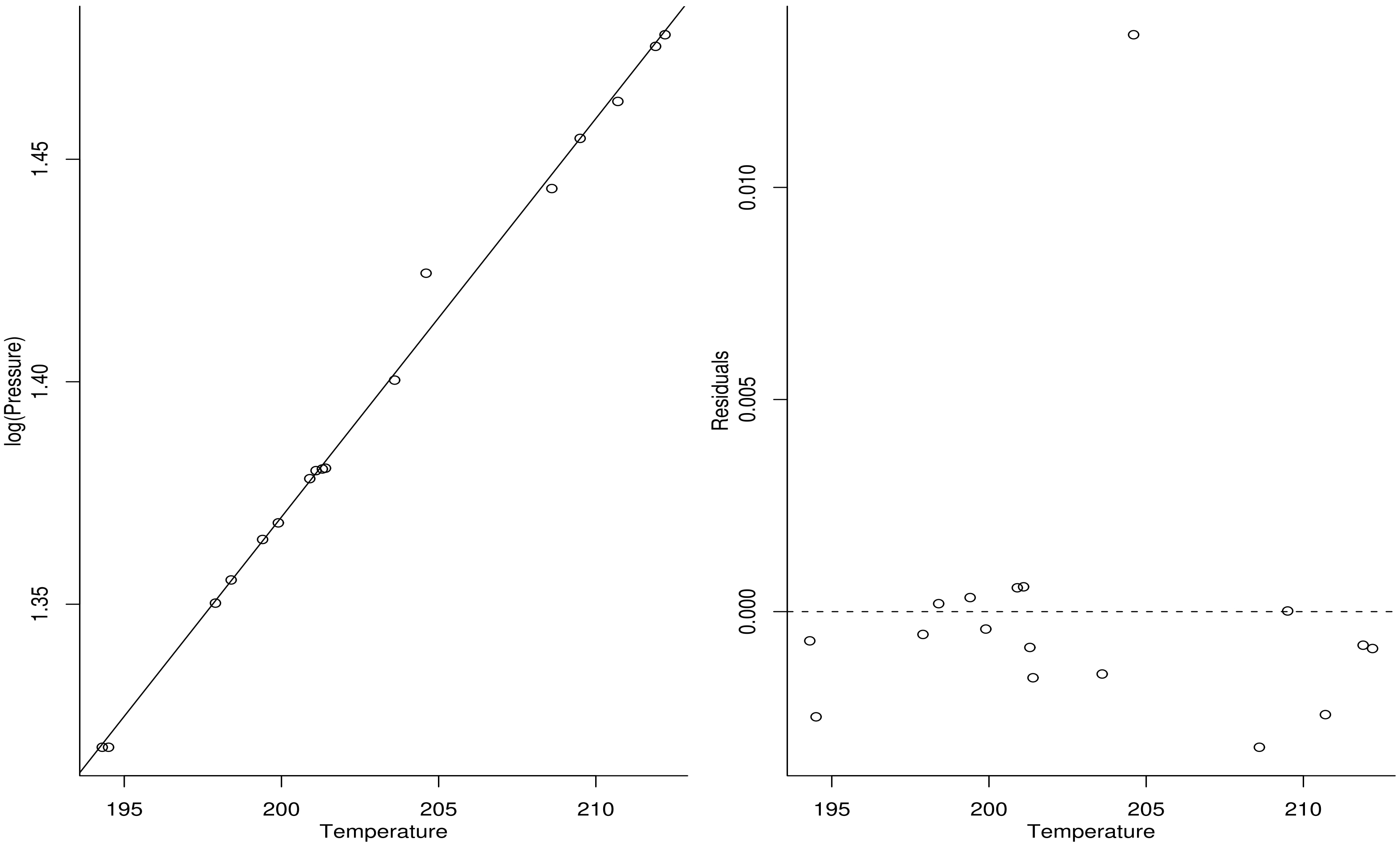

Figure 3: Forbes data

Length at age for smallmouth bass The smallmouth bass is a favorite game fish in inland lakes. One tool in the study of fish populations is to understand the growth pattern of fish such as the dependence of a measure of size like fish length on age of the fish.

Figure 1.5 displays the $Length$ at capture in mm versus Age at capture for $n = 439$ smallmouth bass measured in West Bearskin Lake in Northeastern Minnesota in 1991. Only fish of age seven or less are included in this graph. Fish scales have annular rings like trees, and these can be counted to determine the age of fish.

The predictor Age can only take on integer values, so we are really plotting seven distinct populations of fish. As might expected, length generally increases with age, but the longest fish at age-one fish exceeds the length of the shortest age-four fish, so knowing the age of a fish will not allow to predict its length exactly.

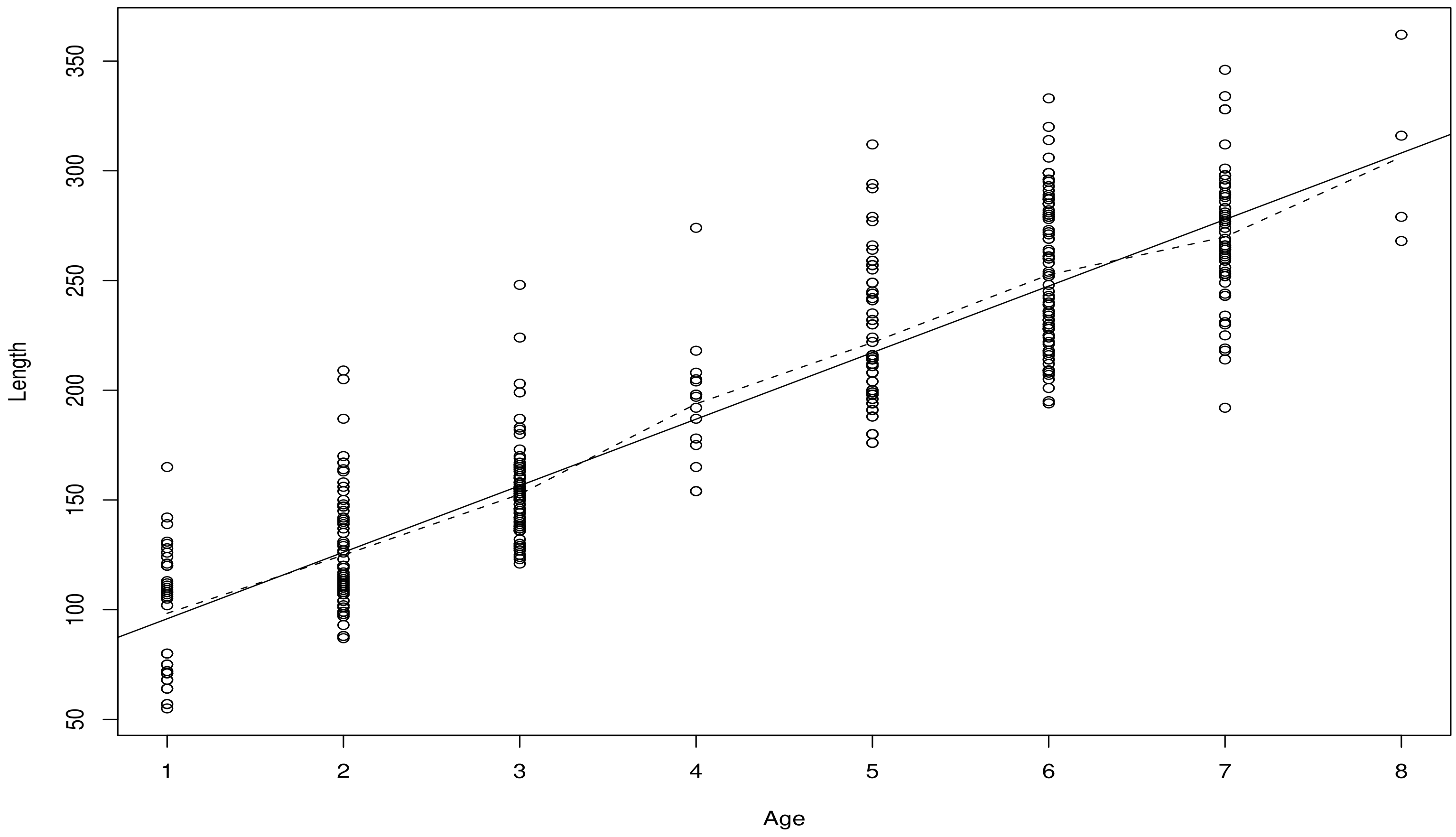

Figure 4: Length (mm) versus age for West Bearskin Lake smallmouth bass.

Predicting the weather Can early season snowfall from September 1 until December 31 predict snowfall in the remainder of the year, from January 1 to June 30? Figure 5 suggests that early winter snowfall and late winter snowfall may be completely unrelated, or uncorrelated. Interest in this regression problem will therefore be in testing the hypothesis that the two variables are uncorrelated versus the alternative that they are correlated.

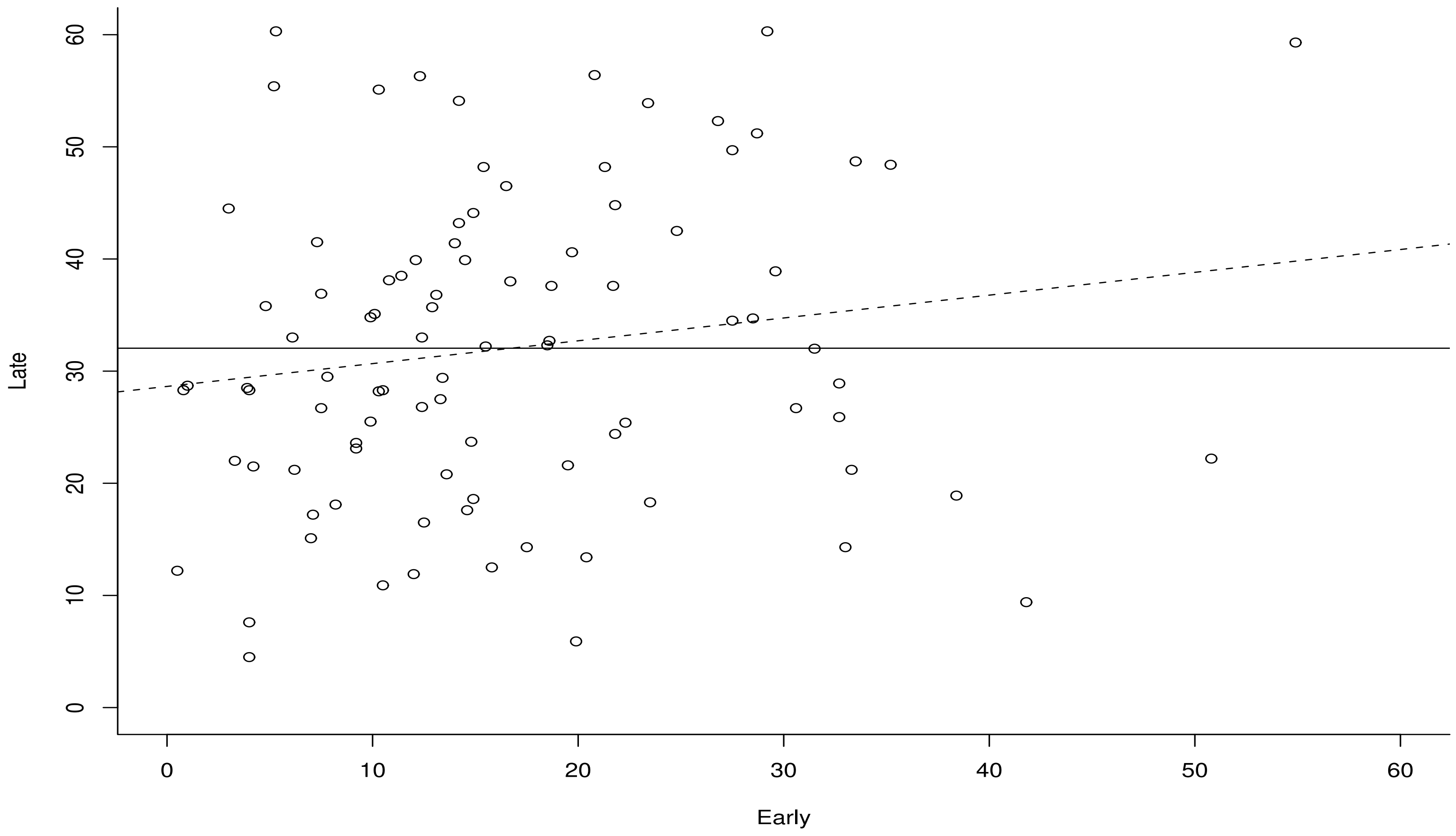

Figure 5: Plot of snowfall for 93 years from 1900 to 1992 in inches.

Turkey Growth Pens of turkeys were grown with an identical diet, except that each pen was supplemented with a dose of the amino acid methionine as a percentage of the total diet. The response is average weight gain in grams of all the turkeys in the pen. Weight gain seems to increase with increasing Dose, but the increase does not appear to be linear, meaning that a straight line does not seem to be a reasonable representation of the average dependence of the response on the predictor.

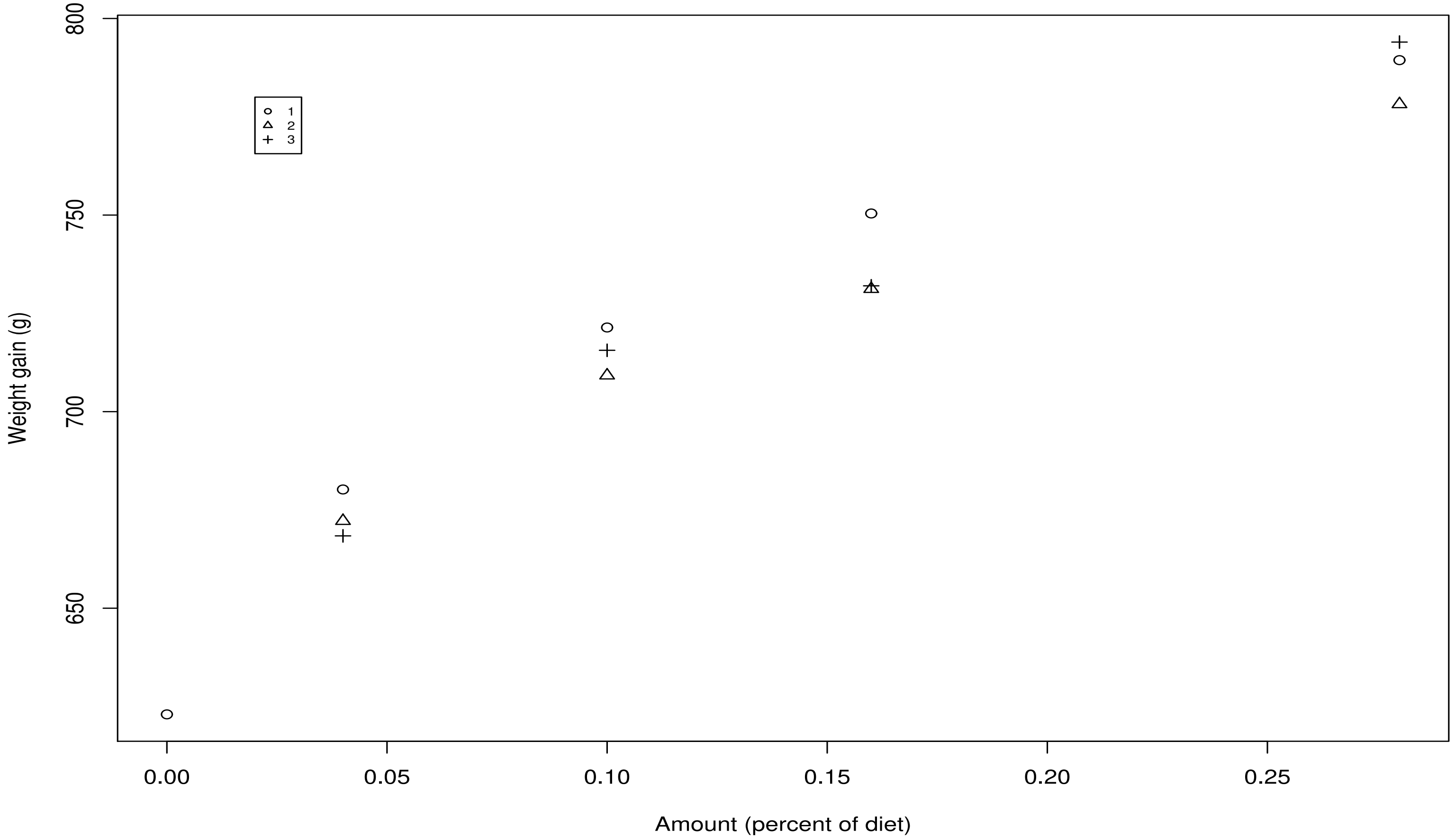

Figure 6: Weight gain versus dose of methionine for turkeys. The three symbols for the points refer to three different sources of methionine.

3 Mean Functions

Imagine a generic summary plot of Y versus X. The mean function is defined by

$$E(Y|X) = \text{a function that depends on the value of } x. \tag{1}$$

For example, in the heights data, we might believe that

$$E(Dheight|Mheight = x) = \beta_0 + \beta_1 x, \tag{2}$$

where the parameters β_0 and β_1 are called the intercept and the slope, respectively.

Note all summary graphs will have a straight-line mean function. In the turkey data and other growth models, a nonlinear mean function might be more appropriate, such as

$$E(Y|Dose = x) = \beta_0 + \beta_1[1 - \exp(-\beta_2 x)]. \tag{3}$$

This three-parameter mean function will be considered in Chapter 11.

4 Variance Function

The variance function is defined by Var$(Y|X = x)$; that is, the variance of the response distribution given that the predictor is fixed at $X = x$.

A frequent assumption in fitting linear regression models is that the variance function is the same for every value of x. This is usually written as

$$\text{Var}(Y|X = x) = \sigma^2. \tag{4}$$

5 Summary Graph

The scatterplots for the above examples are all typical. Examination of the summary graph is a first step in exploring the relationships these graphs portray. Anscombe (1973) provided the artificial data that consists of 11 pairs of points (x_i, y_i), to which the simple linear regression mean function with the same estimated slope, intercept, and other summary statistics, but the visual impression of each of the graph is very different.

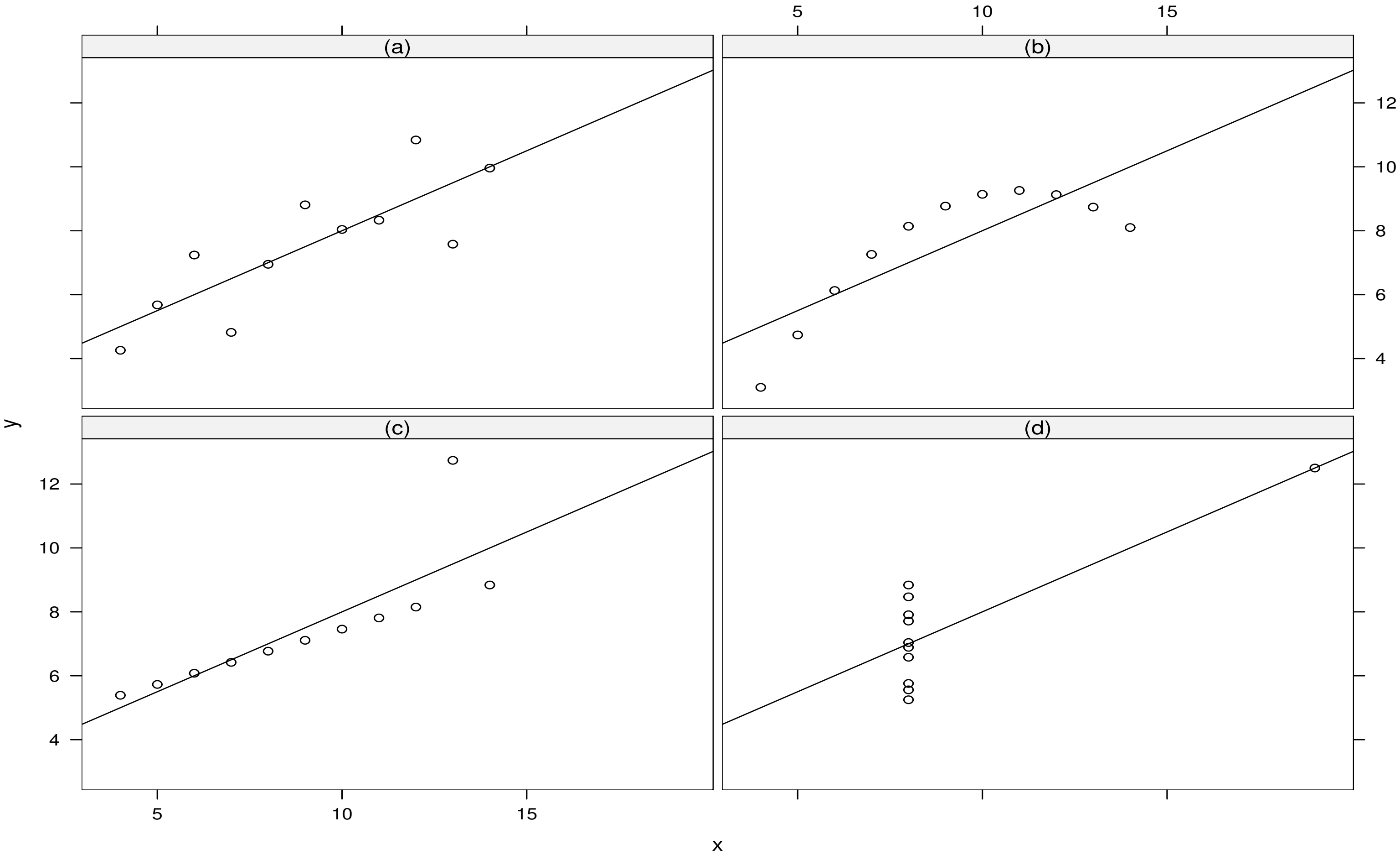

Figure 7: Four hypothetical data sets (from Anscombe, 1973).

6 Tools for looking at scatterplots

Because looking at scatterplots is so important to fitting regression models, we establish some common vocabulary for describing the information in them and some tools to help us extract the information they contain.

Size To extract all the available information from a scatterplot, we may need to interact with it by changing scales, by resizing, or by removing linear trends.

Transformations In some problems, either or both of X and Y can be replaced by transformations so the summary graph has desirable properties. Most of the time, we will use power transformations, replacing, for example, X by X^λ for some number λ.

Smoothers for the mean function Although many authors discuss nonparametric regression as an end in itself, we will generally use smoothers as plot enhancements to help us understand the information available in a scatterplot and to help calibrate the fit of a parametric mean function to a scatterplot.

The loess smooth estimates $E(Y|X = x)$ at the point x by fitting a straight line to a fraction of the points closet to x; we used the fraction of 0.2 in Figure 8 because the sample size is so large, but it is more usual to set the fraction to about 2/3.

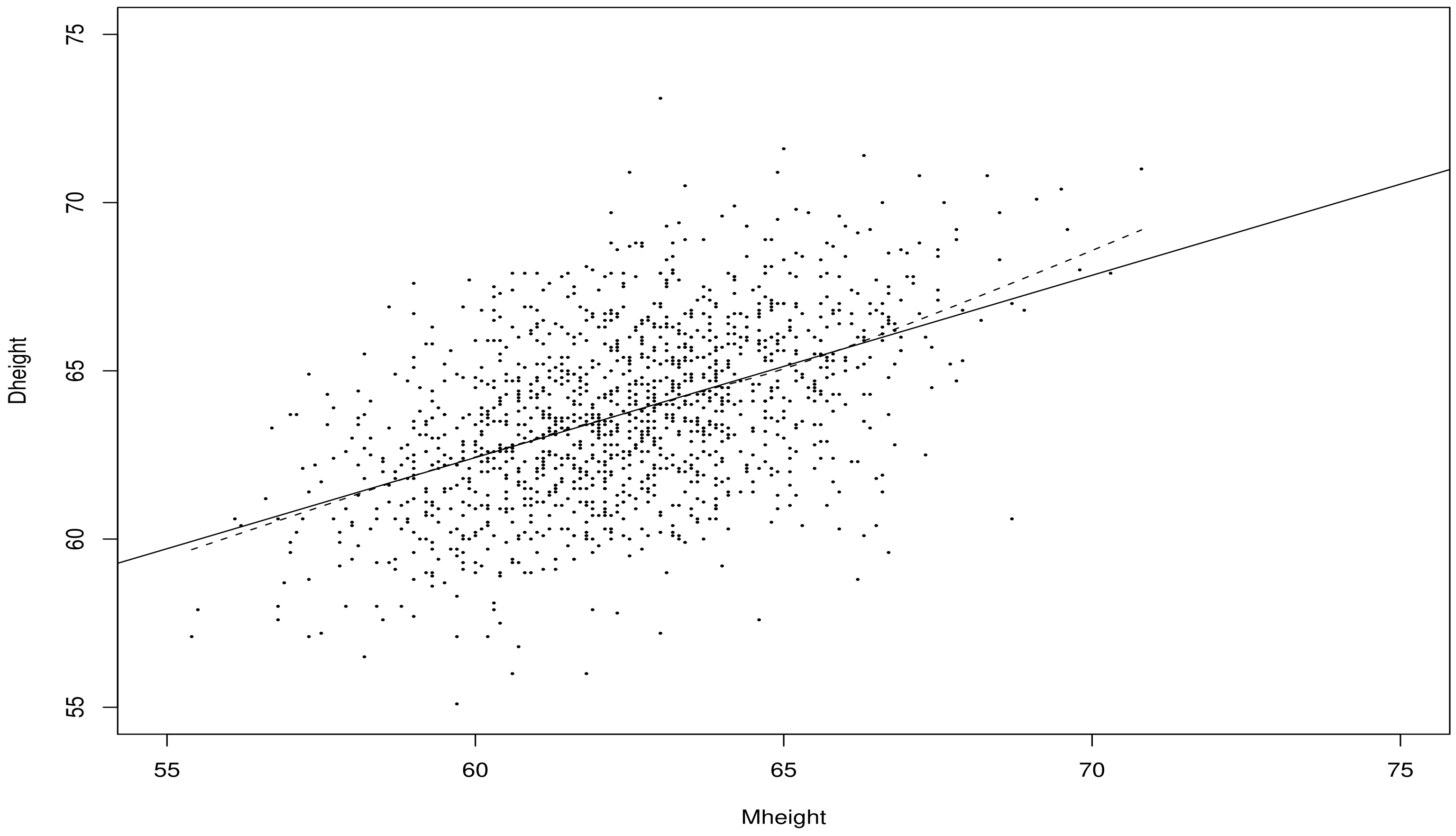

Figure 8: Heights data with the OLS line and a loess smooth with span=0.1.

7 Scatterplot Matrices

When there are many predictors, the scatterplot matrix can be used to look at the regression relationship between the response and the potential predictor.

Fuel Consumption The goal of this example is to understand how fuel consumption varies over the 50 Unites States. The variables considered in this example are as follows.

- Drivers: Number of licensed drivers in the state
- Fuel: Gasoline sold for road use, thousands of gallons
- Income: Per person personal income for the year 2000, in thousands of dollars
- Miles: Miles of Federal-aid highway miles in the state
- Pop: 2001 population age 16 and over
- Tax: Gasoline state tax rate, cents per gallon

- State: State name
- Fuel: 1000 × Fuel/Pop
- Dlic: 1000 × Drivers/Pop
- log(Miles): Base-two logarithm of Miles

The scatterplot matrix for the fuel data is shown in Figure 9. Each plot in a scatterplot matrix is relevant to a particular one-predictor regression of the variable on the vertical axis, given the variable on the horizontal axis.

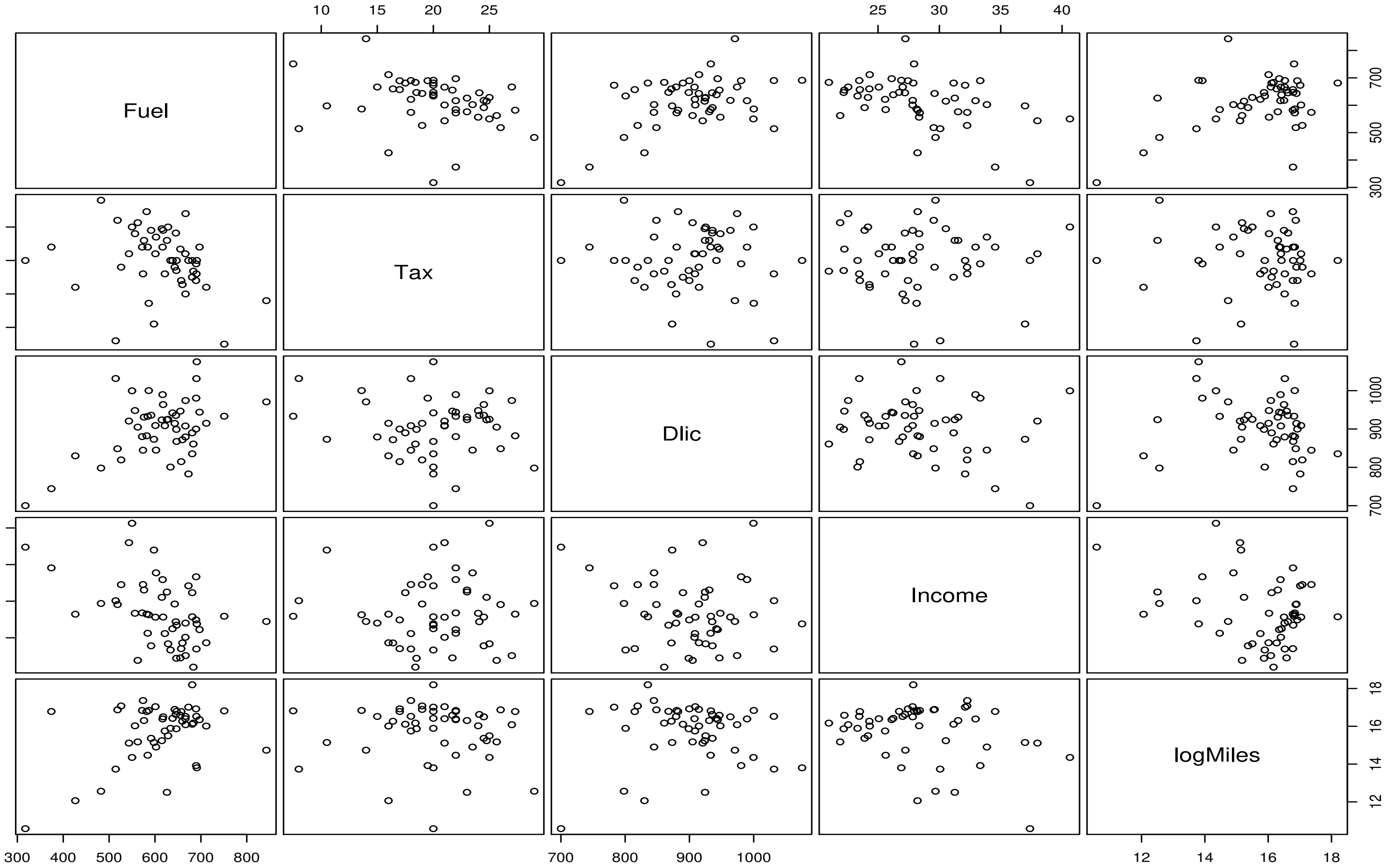

Figure 9: Scatterplot matrix for the fuel data.