Chapter 4 : Scatter Plots and Correlation

Data sometimes comes as pairs (x, y), or triples (x_1, x_2, x_3) or even in larger sets or dimensions.

In this chapter the text focuses on pairs, that is data of the form (x, y). In experiments or studies one observes $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

In some studies one wishes to understand some relationship between the variables x and y. In other studies one wishes to understand if and how one can use the variable x to predict the outcome or the distribution of the variable y. In the first case we typically use a tool called *correlation* and in the second we typically use a tool called *regression*. Chapter 4 introduces *correlation* and the next chapter (Chapter 5) introduces *regression*.

What is correlation? First we need a definition so we can compute it some pairs of data (pairs of numbers).

Consider the n pairs of data (x_i, y_i), $i = 1, 2, 3, \ldots, n$. This is short hand for data pairs $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$.

First we need to calculate the sample mean and variance for the x data, that is the sample mean and variance for $x_1, x_2, \ldots, x_n$. Call these $\bar{x}$ and s_x^2.

The s-squared notation s^2 is to help remind us this is an *average* of squared distances from the centre $\bar{x}$.

Similarly we need to calculate the sample mean and variance for the y data, that is the sample mean and variance for $y_1, y_2, \ldots, y_n$. Call these $\bar{y}$ and s_y^2.

The correlation r is given by

$$\begin{aligned} r &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\ &= \frac{1}{n-1} \left[\left(\frac{x_1 - \bar{x}}{s_x} \right) \left(\frac{y_1 - \bar{y}}{s_y} \right) + \left(\frac{x_2 - \bar{x}}{s_x} \right) \left(\frac{y_2 - \bar{y}}{s_y} \right) + \ldots + \left(\frac{x_n - \bar{x}}{s_x} \right) \left(\frac{y_n - \bar{y}}{s_y} \right) \right] \end{aligned}$$

For calculations we can factor out the common term in the denominator of each of the n terms, that is $s_x \cdot s_y$. Doing this we can calculate r by

$$\begin{aligned} r &= \frac{1}{(n-1) s_x s_y} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) \\ &= \frac{1}{(n-1) s_x s_y} \left[(x_1 - \bar{x}) (y_1 - \bar{y}) + (x_2 - \bar{x}) (y_2 - \bar{y}) + \ldots + (x_n - \bar{x}) (y_n - \bar{y}) \right] \end{aligned}$$

This second formula is easier to use.

Data

x_1, y_1	$-2, -4$
x_2, y_2	$0, 2$
x_3, y_3	$2, 5$

Calculate r = correlation

$\bar{x}$	0
$\bar{y}$	1
s_x^2	4
s_y^2	21
s_x	$\sqrt{4} = 2$
s_y	$\sqrt{21} = 4.58$
numerator r	1.964
r	$\frac{1.964}{2} = .982$

The line referring to the numerator of r means the first formula. The numerator is calculated by

$$\begin{aligned}
&\sum_{i=1}^{3}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)\\
&= \left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right)+\left(\frac{x_3-\bar{x}}{s_x}\right)\left(\frac{y_3-\bar{y}}{s_y}\right)\\
&= \left(\frac{-2-0}{\sqrt{4}}\right)\left(\frac{-4-1}{\sqrt{21}}\right)+\left(\frac{0-0}{\sqrt{4}}\right)\left(\frac{2-1}{\sqrt{21}}\right)+\left(\frac{2-0}{\sqrt{4}}\right)\left(\frac{5-1}{\sqrt{21}}\right)\\
&= \left(\frac{-2}{2}\right)\left(\frac{-5}{4.583}\right)+0+\left(\frac{2}{2}\right)\left(\frac{4}{4.583}\right)\\
&= 1.964
\end{aligned}$$

Thus the correlation is

$$r=\frac{1.964}{2}=.982$$

Example : Miles per gallon (US vehicle fuel rating measure)

A data set of the US EPA ratings for fuel consumption for 2004 model years is from the web site causeweb.org. It contains 428 vehicles, but some of the data is missing. There are 428 in this list.

Here we just consider weight and car mileage, measured as MPG = miles per (US) gallon. One US gallon is about 3.79 litres. There is also an imperial gallon (used in Great Britain and some other countries) that is 4.54 litres. One mile is a distance measurement, that is approximately 1.62 kilometres. The weights are measured in pounds, one pound equal to about .45 kg.

Figure 1 shows a scatter plot for all 428 vehicles, showing the vehicle weight on the x axis and the fuel rating MPG for city driving on the y axis. We see there is a general trend for MPG to decrease as the vehicle weight increases.

The (sample) correlation for these data pairs is -0.82.

This set of vehicles includes pickup trucks, 4, 6 and 8 cylinder cars.

To get a better picture we consider now just the 4 cylinder vehicles. Figure 2 shows the city MPG for these. The (sample) correlation is -0.56.

We further remove hybrids as they have a different level of fuel consumption, that is all engines with a normal or usual fuel source. Theses are shown in Figure 3. For these vehciles the (sample) correlation is -0.74.

Correlation measures how well one variable is related (in a linear way) to the other variable.

A positive correlation indicates that as one variable increase the other increases. A negative correlation indicates that when one variable increases the other tends to decrease.

The minimum correlation that can be is -1 and the maximum is +1. These happen only if the scatter plot data fall *perfectly* along a straight line.

Figure 4 shows the same data as Figure 3 but with an additional line drawn. This line has intercept 46.6 and slope -0.0076. Notice this line gives a similar description of the information in the scatter plot, in that for a given weight one can predict, but not perfectly, the fuel rating for a vehicle.

Note the slope -.0076 is small, but weights of cars differ by amounts of nearly 2000 pounds.

Where did this line come from? This idea is explored further when regression is introduced.

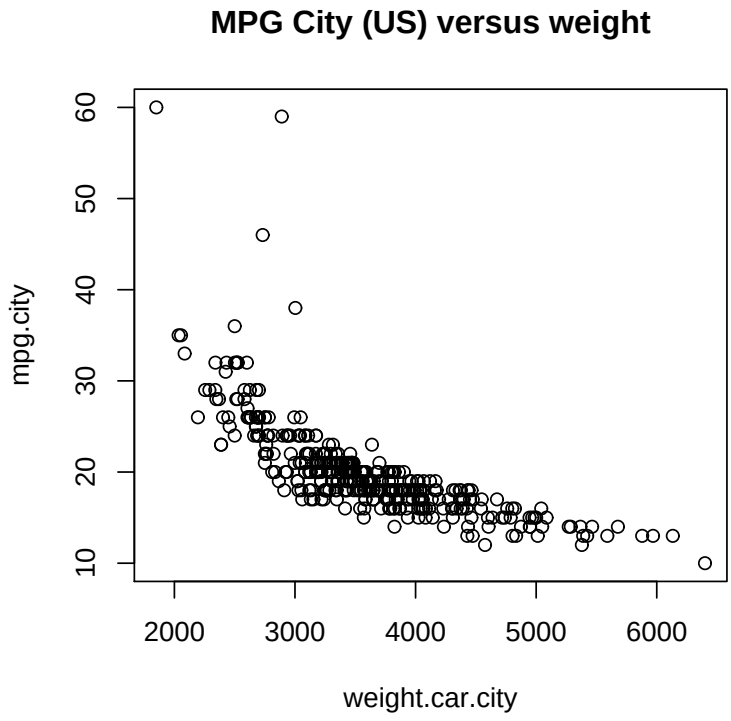

Figure 1: MPG City all Cars

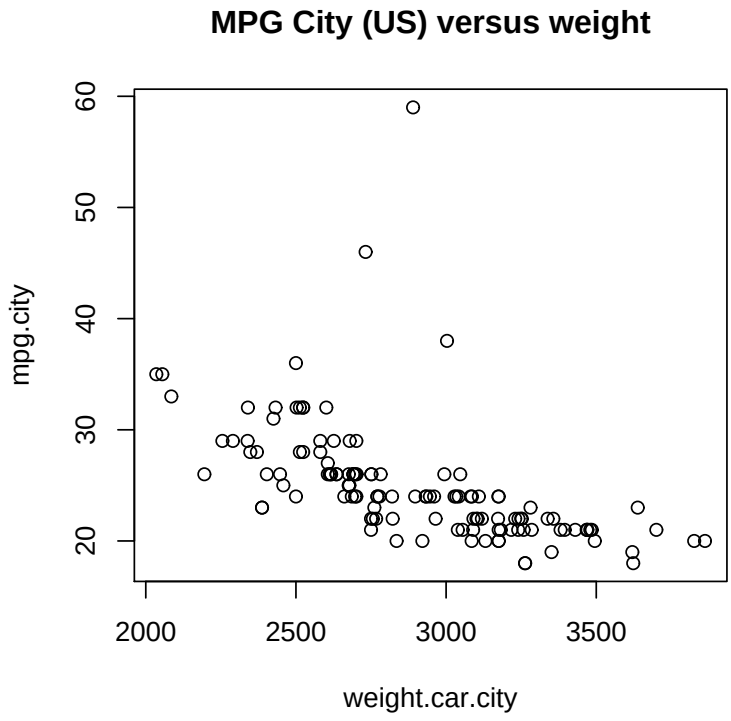

Figure 2: MPG City all 4 Cylinder Cars

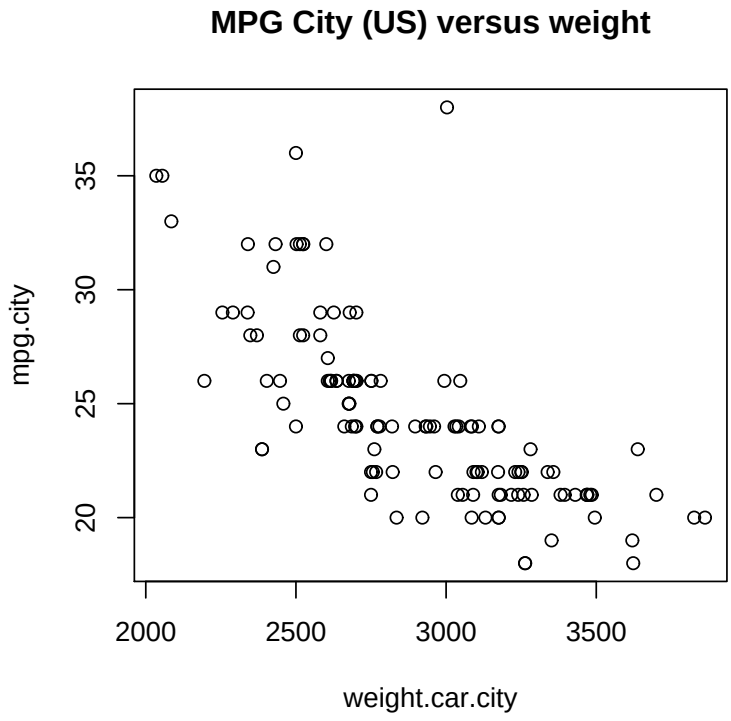

Figure 3: MPG City all 4 Cylinder Cars, Normal Power

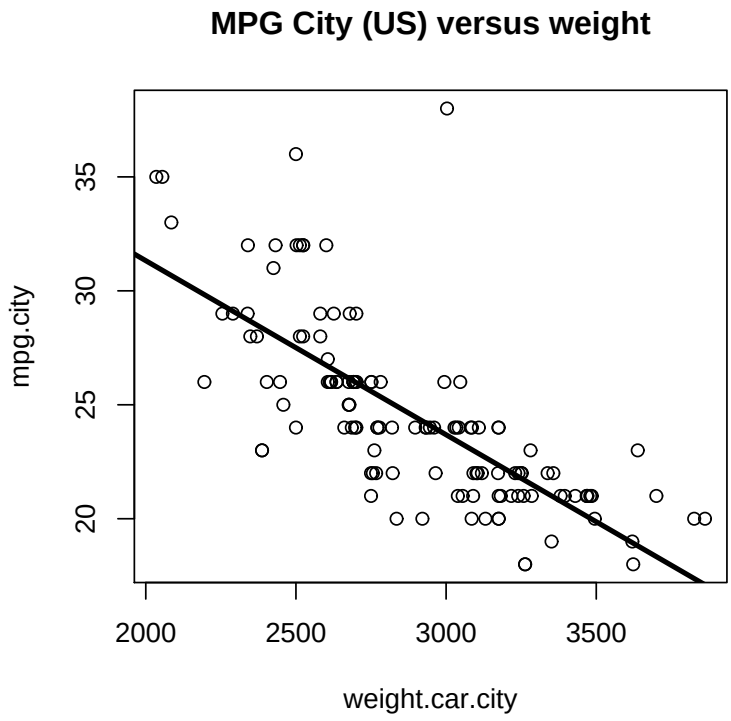

Figure 4: MPG City all 4 Cylinder Cars, Normal Power and Best Line : mpg = 46.6 - 0.0076*weight

If the $n-1$ in the denominator were instead n, then r is the average of the product of $x-\bar{x}$ with $y-\bar{y}$. If for a typical pair (x_i, y_i) it were the case that

- when x_i is bigger than the average or mean x then so is y_i bigger than the average or mean y
- when x_i is smaller than the average or mean x then so is y_i smaller than the average or mean y

then in both these cases

$$\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \text{positive} \cdot \text{positive} > 0$$

Similarly

If for a typical pair (x_i, y_i) it were the case that

- when x_i is bigger than the average or mean x then so is y_i smaller than the average or mean y
- when x_i is smaller than the average or mean x then so is y_i bigger than the average or mean y

then in in the first case

$$\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \text{positive} \cdot \text{negative} < 0$$

and in the second case

$$\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \text{negative} \cdot \text{positive} < 0$$

Notice this is what happens in our car mileage data example.

Some Properties (or facts) about Correlation

1. Positive correlation indicates a positive association between variables

   Negative correlation indicates a negative association between variables

2. Correlation is calculated for *standardized* variables, thus correlation does not change if we change in a linear way units of measurements. This will be the case for changes from feet to centimetres or metres, from miles to kilometres, from degrees Fahrenheit (used in USA) to degrees centigrade (used in Canada), from days to years

   It will make a difference if the units change is not linear, for example strength measures in foot-pounds or metres-grams versus a log scale, such as the Richter scale which measures earthquake energy or force on a logarithmic scale.

3. Correlation is always a number between -1 and $+1$. The closer correlation r is to the extremes (-1 or $+1$) the more closely the scatter plot is falling along a perfect straight line.

4. r is a number and does not have any units. Thus we **cannot interpret a correlation of** .6 **as *twice as correlated* as when the correlation is** .3.

Body Fat

There are several ways of measuring body fat, some complicated and some simple to perform. The simplest common one is the Adiposity index

$$\text{Adiposity index} = \frac{\text{Weight kg}}{(\text{Height metres})^2}$$

More complcated measurements are based on density, such as the Brozek method, the percent body fat using Brozek's equation

$$\text{Brozek} = \frac{457}{\text{Density}} - 414.2$$

How does these measure compare?

Data is collected on the body fat of 252 individuals of various ages, weights and heights. The body fat is measured by 3 methods, two based on density and one the Adiposity index. Here we only consider comparing the Brozek and Adiposity measurements. Figure 5 shows the scatter plot of these 252 pairs of data. Noitce how they cluster, indicating generally when one measurement increases so does the other.

The sample correlation is $r = 0.72$.

Also shown on this plot is a straight line with slope 1.55 and intercept -20.42. We can read it as

$$\text{Brozek} = -20.42 + 1.55 \cdot (\text{Adiposity index}) \ .$$

Thus for Adiposity index = 30 we would guess or predict a Brozek mea-

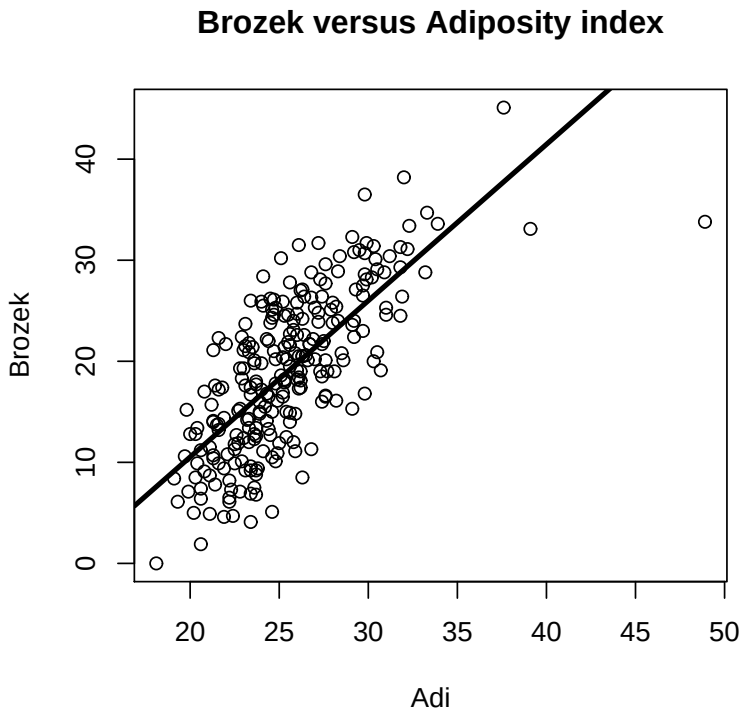

Figure 5: Brozek index versus Adiposity index (Body Fat Measurments)

surement of

$$-20.42 + 1.55 * 30 = 26$$

Notice that while this prediction might be reasonable for Adiposity from about 20 to 35, it is not a good predictor when the Adiposity index is bigger than 35 or so.

Diamond Prices in Singapore Dollars

This data set consist of $n = 48$ pairs of points. The first 9 are shown here. They are the carrots (measure of weight) and the retail price in Singapore dollars in the year (1992) the data was collected.

	carrots	price
1	0.17	355
2	0.16	328
3	0.17	350
4	0.18	325
5	0.25	642
6	0.16	342
7	0.15	322
8	0.19	485
9	0.21	483

The scatter plot of the 48 data points is shown in Figure 6. Here I have chosen to label the axis with something meaningful such as d.carrot and d.price for diamond carrot size and diamond price respectively. Other labels are equally good, such as carrots and price, or weight and price.

We can see from the plot that the sample correlation should be positive, and that it should be quite large, that is close to 1. Calculating this (excel, SAS, SPSS, R, or some other convenient program) we obtain r = .989. The best fitting straight line will be a good fit for this data.

Excel is used in small data sets in the business environment, SAS is

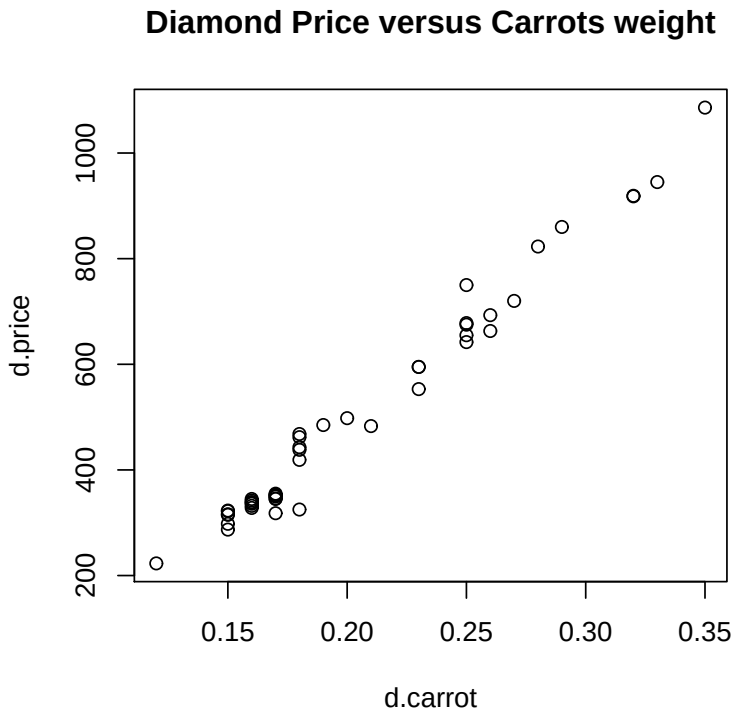

Figure 6: Brozek index versus Adiposity index (Body Fat Measurments)

used for many medical and pharmacutical environments, SPSS is used in the social sciences and R is being increasingly used several areas, such as statistical consulting work.