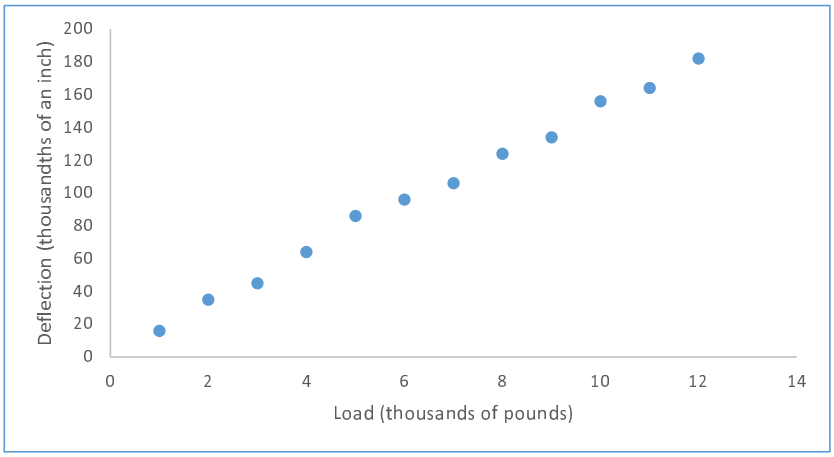
10 Linear Regression

Given a *bivariate* data set (i.e. a set of points in two variables), it is often desirable to express a concise relationship between the variables. The best way to start is to graph the data set using a *scatterplot* to visually assess the pattern of the points.

Consider, for example, the following table listing the deflection of a tensile ring at various loads [2, p. 338]. The x's are the load forces in thousands of pounds and the y values are the corresponding deflections in thousandths of an inch:

x	y
1	16
2	35
3	45
4	64
5	86
6	96
7	106
8	124
9	134
10	156
11	164
12	182



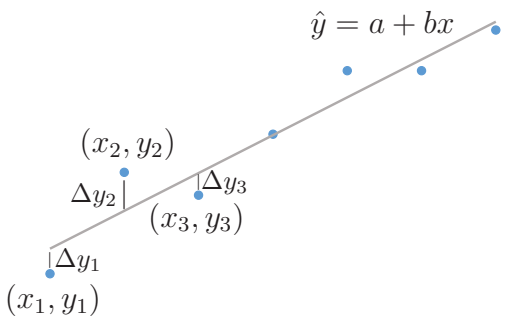
From the accompanying scatterplot, we can tell that this data set has a strong linear association. Bivariate data sets can also have non-linear correlation, but in this course we will only look at the linear case.

10.1 The Least Squares Regression Line

The *least squares regression line* is also commonly referred to as the *best-fit line*. It is denoted $\hat{y}$ and is the unique line that minimizes

$$\sum (\Delta y_i)^2,$$

where $\Delta y_i = y_i - \hat{y}_i$ is called the *residual* of the point (x_i, y_i).



If a scatterplot for a bivariate data set with n points reveals a linear association, we can find the equation of the least squares regression line as follows:

Step 1: Calculate the means $\bar{x}$ and $\bar{y}$

Step 2: Calculate

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \qquad \text{and} \qquad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

To find these two values by hand, it is useful to make a table:

x	y	xy	x^2
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$

Step 3: Calculate

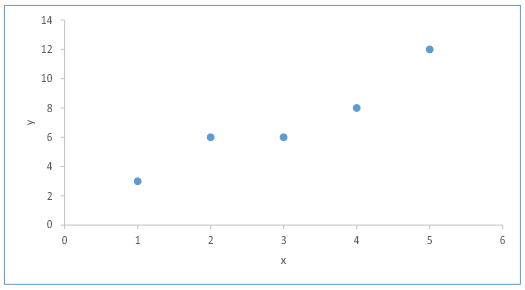
$$b = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad a = \overline{y} - b\overline{x}$$

Step 4: The equation is $\hat{y} = a + bx$

The SHARP EL-531X calculator has functions to calculate a and b for $\hat{y}$. Instructions are available in the Appendix.

Example 10.1. Consider the following data set:

x	y	xy	x^2
1	3		
2	6		
3	6		
4	8		
5	12		



(a) Find the equation of the least squares regression line.

(b) Graph the least squares regression line for $1 \leq x \leq 5$ on the given scatterplot.

(c) Find the residual for the point $(2, 6)$.

If a least squares regression line is to be used to predict values that are not in the given data set, it is important to do so only for values in the given ranges.

Example 10.2. For the data set in the previous example, use $\hat{y}$ to predict:

(a) the y-value for $x = 2.4$.

(b) the y-value for $x = 6$.

(c) the x-value for $y = 10$.

10.2 The Coefficients of Correlation and Determination

There are two useful coefficients that describe how well a linear regression fits a data set, but they should only be used in conjunction with scatterplots.

Definition The coefficient of correlation is a measure of the strength of the linear relationship of two variables, and is defined by

$$r = b\frac{s_x}{s_y},$$

where s_x and s_y are the standard deviations of the x-values and y-values, respectively, and b is the slope of $\hat{y}$. The values of r lie in the the range $-1 \leq r \leq 1$. If r is near 1, the variables are positively correlated, and if r is near -1, the variables are negatively correlated. For the values of r between -0.5 and 0.5, the linear correlation is poor.

The coefficient of determination is r^2 and indicates what percentage of the variation in y is accounted for by the best-fit line:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}.$$

The SHARP EL-531X calculator has a function to calculate r. Instructions are available in the Appendix.

An alternative way to find r is to use

$$r^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}},$$

where S_{xy} and S_{xx} were defined along with the instructions for finding the equation for $\hat{y}$, and

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}.$$

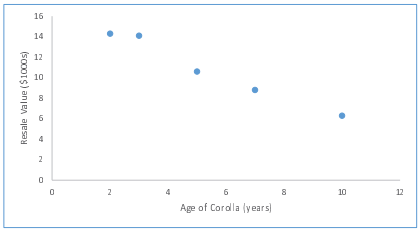
Demonstration: Anscombe's Quartet

Example 10.3. The following bivariate data set has $\hat{y} = 16.54 - 1.06x$ and a coefficient of determination of 0.9746:

x = age of Corolla (years)

y = resale value ($1000s)

x	y
2	14.3
3	14.1
5	10.6
7	8.8
10	6.3



(a) Is the linear association positive or negative?

(b) Find the correlation coefficient.

(c) What % of the variation in y is accounted for by the best-fit line?

(d) What resale value is predicted for a 4-year-old Corolla?

(e) Why should we not predict the resale value for a 1-year-old Corolla?

(f) What age corresponds to a resale value of $4,500?

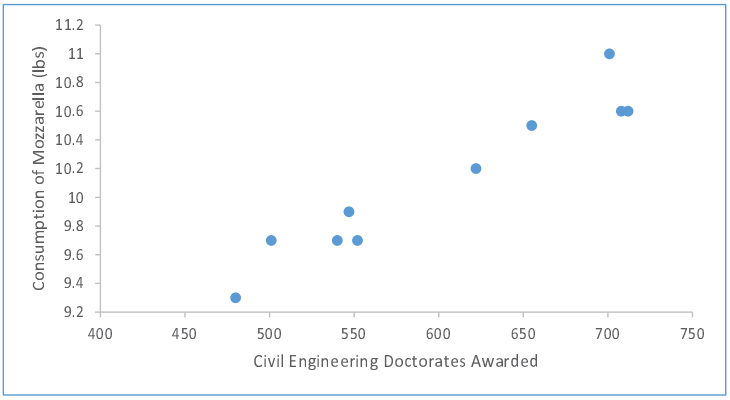
Correlation does NOT imply causation!

The following data set was taken from
`http://www.tylervigen.com/spurious-correlations`

$x =$ the number of civil engineering doctorates awarded in the US, and

$y =$ the per capita consumption of mozzeralla cheese in the US (in lbs):

year	x	y
2000	480	9.3
2001	501	9.7
2002	540	9.7
2003	552	9.7
2004	547	9.9
2005	622	10.2
2006	655	10.5
2007	701	11
2008	712	10.6
2009	708	10.6



Surprisingly, $r = 0.9586$. The best-fit line is $\hat{y} = 6.600 + 0.006x$, but this does NOT mean that an increase of 1 doctorate **causes** the mozzarella consumption to increase by 0.006 lbs per person (nor can we assume that mozzarella consumption causes an increase in Civil Engineering doctorates). The accurate interpretation of the slope is: *as the number of doctorates increases by 1, the mozzarella consumption increases on average by 0.006 lbs per person.*

Additional Notes

Additional Notes