Correlation and Simple Linear Regression

- Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other, or others.

- **Some Examples:**

  1. Waistline and Weight.

  2. SAT score and First year college GPA.

  3. Number of customers and Revenue.

  4. Family income and Family expenditures.

- **Functional Relation vs. Statistical Relation between two variables.**

  – A *functional relation* between two variables is expressed by a mathematical formula. If X is the *independent variable* and Y the *dependent variable*, a functional relation is of the form:

  $$Y = f(X).$$

  That is, given a particular value of X, we get only one corresponding value Y.

  1. For example, let x denote the number of printer cartridges that you order over the internet. Suppose each cartridge costs \$40 and there is a fixed shipping fee of \$10, determine the total cost y of ordering x cartridges.

  – A *statistical relation*, unlike a functional relation, is not a perfect one. If X is the *independent variable* and Y the *dependent variable*, a statistical relation is of the form:

  $$Y = f(x) + \epsilon.$$

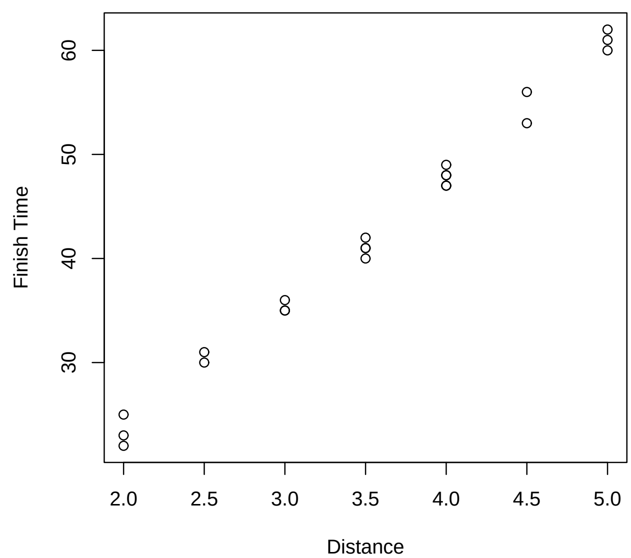
  In such cases, we call X an *explanatory variable* and Y a *response variable.*

  1. For example, let x denote the distance of a marathon and y the time that it will take a certain runner to finish it. If this runner can runner an average of 5 miles per hour, determine the time it will take for this runner to finish a 5-mile marathon.

  2. Consider the following 22 practice finish times of this runner.

	1	2	3	4	5	6	7	8	9	10	11
Distance (x)	2	2	3	3	2	2.5	2.5	3	3.5	3.5	4
Time (y)	25	22	35	36	23	30	31	35	41	40	49
	12	13	14	15	16	17	18	19	20	21	22
Distance (x)	4	4	4	4.5	4.5	5	5	5	3.5	3.5	4
Time (y)	47	48	48	56	53	62	60	61	42	41	47

- **Scatterplots.** A *scatterplot* (or *scatter diagram*) is a graph in which the paired (x, y) sample data are plotted with a horizontal $x-$axis and a vertical $y-$axis. Each individual (x, y) pair is plotted as a single point. Scatterplots are useful as they usually display the relationship between two quantitative variables.
  - Always plot the explanatory variable on the $x-$axis, while the response variable on the $y-$axis.
  - In examining a scatterplot, look for an overall pattern showing the **form**, **direction**, and **strength** of the relationship, and then for outliers or other deviations from this pattern.
    * Form - linear or not.
    * Direction - positive or negative association.
    * Strength - how close the points lie to the general pattern (usually a line).



- **Correlation.** A *correlation* exists between two variables when one of them is related to the other in some way.

- **Linear Correlation.** The *linear correlation coefficient* r measures the strength of the linear relationship between the paired $x-$ and $y-$quantitative values in a sample.

$$
\begin{aligned}
r &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} && (1)\\
&= \frac{n\sum x_iy_i-(\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}} && (2)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) && (3)\\
&= \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} && (4)
\end{aligned}
$$

where,

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = (n-1)s_x^2 \quad (5)$$

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = (n-1)s_y^2 \quad (6)$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) \quad (7)$$

- **Tree Circumference and Height.** Listed below are the circumferences (in feet) and the heights (in feet) of trees in Marshall, Minnesota (based on data from "Tree Measurements" by Stanley Rice, *American Biology Teacher*.

x (circ)	1.8	1.9	1.8	2.4	5.1	3.1	5.5
y (height)	21.0	33.5	24.6	40.7	73.2	24.9	40.4
x (circ)	5.1	8.3	13.7	5.3	4.9	3.7	3.8
y (height)	45.3	53.5	93.8	64.0	62.7	47.2	44.3

Compute for $s_x, s_y, \Sigma xy, SS_{xx}, SS_{yy}, SS_{xy}$, and r.