Correlation

- 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:

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

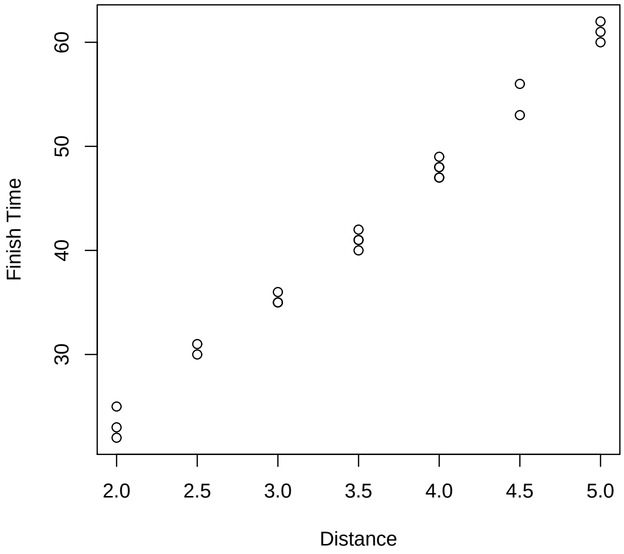
  In such cases, we call X an *explanatory variable* and Y a *response variable.*

  1. For example, let x denote the distance that a person plans to jog and y the time that it will take this person to finish it. Consider his 22 jogging distances and times from last month shown in the table below. If this person plans to jog for 5.5 miles tomorrow, predict how long it will take him to finish the run.

	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.

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

where,

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

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

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

- R commands

```
ss.xx=function(x){sum(x^2)-(sum(x))^2/length(x)}
ss.yy=function(y){sum(y^2)-(sum(y))^2/length(y)}       # Equivalent to ss.xx(y)
ss.xy=function(x,y){sum(x*y)-sum(x)*sum(y)/length(x)}

data.jog=read.csv("Jogging.csv",header=T)
attach(data.jog)
ssxx=ss.xx(Distance)          # 19.45455
ssyy=ss.xx(Time)              # 3025.091
ssxy=ss.xy(Distance,Time)     # 241.6364
r=ssxy/sqrt(ssxx*ssyy)        # 0.9960532

corr=function(x,y)
{
ss.xy=function(x,y){sum(x*y)-sum(x)*sum(y)/length(x)}
ssxy=ss.xy(x,y)
ssxx=ss.xy(x,x)
ssyy=ss.xy(y,y)
r=ssxy/sqrt(ssxx*ssyy)

list(corr=r,SSxx=ssxx,SSyy=ssyy,SSxy=ssxy)
}
corr(Distance,Time)
```

- **Practice.**

  1. **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	5.1	8.3	13.7	5.3	4.9	3.7	3.8
y (height)	21.0	33.5	24.6	40.7	73.2	24.9	40.4	45.3	53.5	93.8	64.0	62.7	47.2	44.3

  Use R to compute for $\Sigma xy, SS_{xx}, SS_{yy}, SS_{xy}, s_x, s_y$, and r.

  2. The table below displays data on age (in years) and price (in \$100)for a sample of 11 cars.

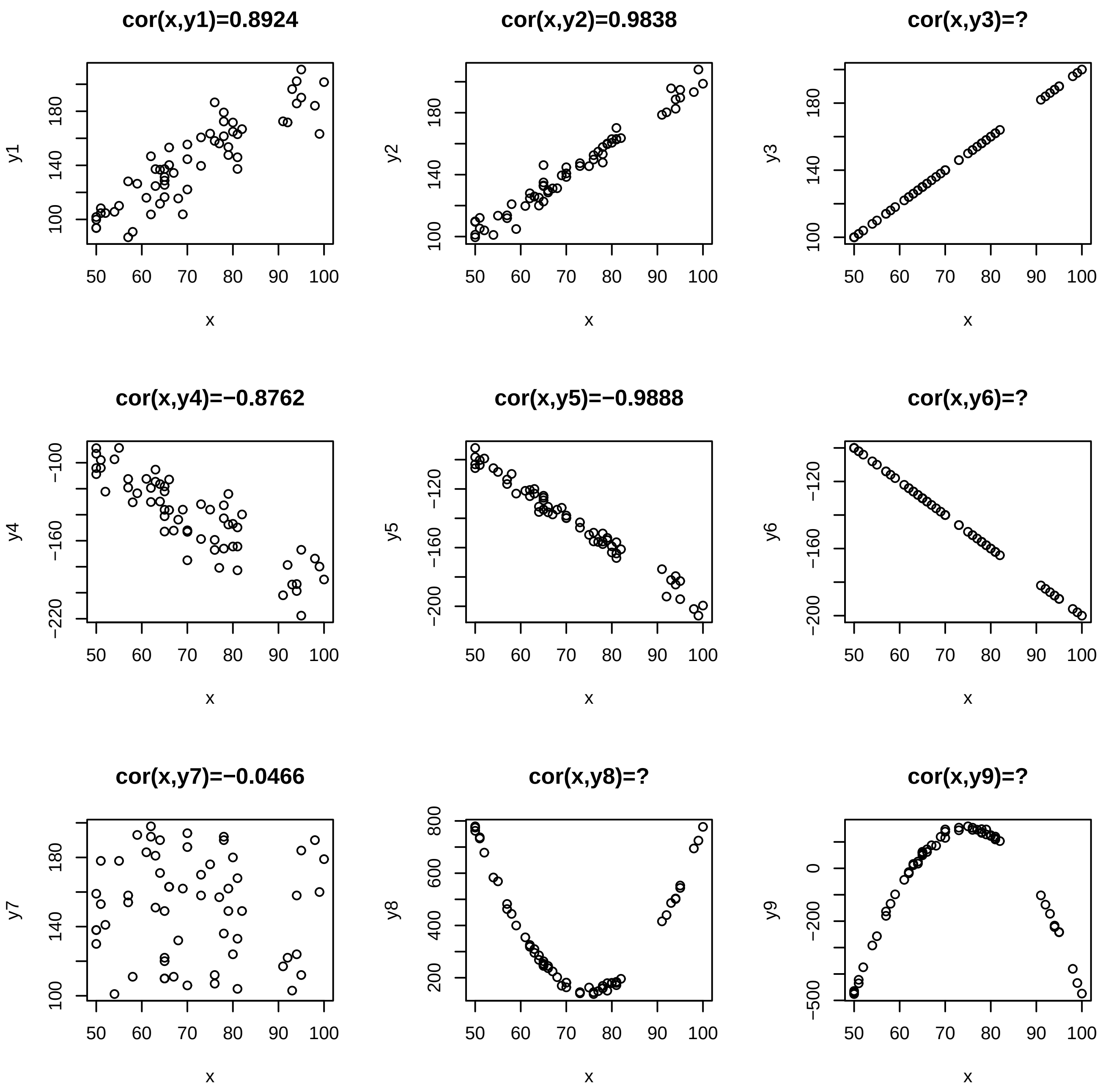
Age (x)	5	4	6	6	5	5	6	6	2	7	7
Price (y)	85	102	70	80	89	98	66	90	169	68	50

  a. Determine the values of SS_{xx}, SS_{yy}, and SS_{xy}.

  b. Determine the correlation coefficient r.

  c. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

- Sample plots with correlation values



- Some guidelines in interpreting r.

Value of $\lvert r\rvert$	Strength of linear relationship
If $\lvert r\rvert \geq .95$	Very Strong
If $.85 \leq \lvert r\rvert < .95$	Strong
If $.65 \leq \lvert r\rvert < .85$	Moderately to Strong
If $.45 \leq \lvert r\rvert < .65$	Moderate
If $.25 \leq \lvert r\rvert < .45$	Weak
If $\lvert r\rvert < .25$	Very weak/Close to none