Correlation Matrix

Module I (Lecture 3)

David Raj Micheal

August 2018

Contents

1 Correlation Matrix

Since the covariance between two random variables x_1 and x_2 depends on the scale of measurement of x_1 and x_2, it is difficult to compare covariances between different pairs of variables. For example, if we change a measurement from inches to centimeters, the covariance will change. To find a measure of linear relationship that is invariant to changes of scale, we can standardize the covariance by dividing by the standard deviations of the two variables. This standardized covariance is called a *correlation*.

The population correlation of two random variables x_1 and x_2 is defined as

$$\rho_{12} = \operatorname{cor}(x_1, x_2) = \frac{\operatorname{cov}(x_1, x_2)}{\operatorname{sd}(x_1)\operatorname{sd}(x_2)} = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \frac{E[(x_1-\mu_1)(x_2-\mu_2)]}{\sqrt{E[(x_1-\mu_1)^2]}\sqrt{E[(x_2-\mu_2)^2]}}$$

and the *sample correlation* is

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{E[(x_1-\overline{x_1})(x_2-\overline{x_2})]}{\sqrt{E[(x_1-\overline{x_1})^2]}\sqrt{E[(x_2-\overline{x_2})^2]}}.$$

Excercise Prove that $-1 \leq r_{ij} \leq 1$ for any two random variables x_i and x_j.

Similar to sample covariance matrix we can find sample correlation matrix using matrix algebra as follows:

Note that, the sample covariance matrix is obtained by

$$S = \frac{1}{n-1} Y^T Y$$

where

$$Y = \begin{pmatrix} x_{11}-\overline{x_1} & x_{12}-\overline{x_2} & \dots & x_{1p}-\overline{x_p} \\ x_{21}-\overline{x_1} & x_{22}-\overline{x_2} & \dots & x_{2p}-\overline{x_p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\overline{x_1} & x_{n2}-\overline{x_2} & \dots & x_{np}-\overline{x_p} \end{pmatrix}. \tag{1}$$

Sample correlation matrix is defined as

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{pmatrix},$$

where r_{ij} is the sample correlation between x_i and x_j. Further R can be written as

$$
\begin{aligned}
R &= \begin{pmatrix} \frac{s_{11}}{s_1 s_1} & \frac{s_{12}}{s_1 s_2} & \dots & \frac{s_{1p}}{s_1 s_p} \\ \frac{s_{21}}{s_2 s_1} & \frac{s_{22}}{s_2 s_2} & \dots & \frac{s_{2p}}{s_2 s_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{p1}}{s_p s_1} & \frac{s_{p2}}{s_p s_2} & \dots & \frac{s_{pp}}{s_p s_p} \end{pmatrix} \\
&= \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & s_p \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \dots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix} \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & s_p \end{pmatrix} \\
&= DSD
\end{aligned}
$$

where $D = \text{diag}(s_1, s_2, \dots, s_p)$ and S is the sample covariance matrix.

1.1 Geometrical meaning of correlation

The sample correlation r_{ij} is related to the cosine of the angle between two vectors. Before visualizing this relationship, let us revise some of the concepts from inner product.

What is the standard inner product of two vectors?

Let $x_i = (x_{1i}, x_{2i}, \dots, x_{ni})^T$ and $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$ be two vectors. Then

$$
\langle x_i, x_j \rangle = x_i^T x_j = \begin{pmatrix} x_{1i} & x_{2i} & \dots & x_{ni} \end{pmatrix} \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} = \sum_{k=1}^{n} x_{ki} x_{kj}
$$

and the norm of the vector x_i is defined as

$$
\|x_i\| = \sqrt{\langle x_i, x_i \rangle} = \sqrt{\sum_{k=1}^{n} (x_{ki})^2}.
$$

Suppose x_i and x_j are random variables then r_{ij} is given by

$$
r_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})}{\sqrt{\sum_{k=1}^{n} (x_{ki} - \overline{x_i})^2} \sqrt{\sum_{k=1}^{n} (x_{kj} - \overline{x_j})^2}} = \frac{\langle x_i - \overline{x_i}, x_j - \overline{x_j} \rangle}{\|x_i - \overline{x_i}\| \|x_j - \overline{x_j}\|} = \cos\theta,
$$

where θ is the angle between the vectors x_i and x_j.

The following code is been used to simulate the data with different correlations. Refer Figure 1 for the output.

```
n = 2000; m1 = 0; m2 = 0; s1 =1; s2 = 1;
plotcor = function(r){
  x = rnorm(n,m1,s1)
  y = s2*r*(x-m1)/s1+m2 + s2*rnorm(n,0,sqrt(1-r^2))
  plot(x,y, main = paste('Correlation =', r))
  abline(lm(y~x), col = 'red')
}
par(mfrow=c(3,3),oma=c(0,0,2,0))
for (i in c(-1,-.75,-.5,-0.25,0,0.25,0.5,.75,1)){
  plotcor(i)
}
```

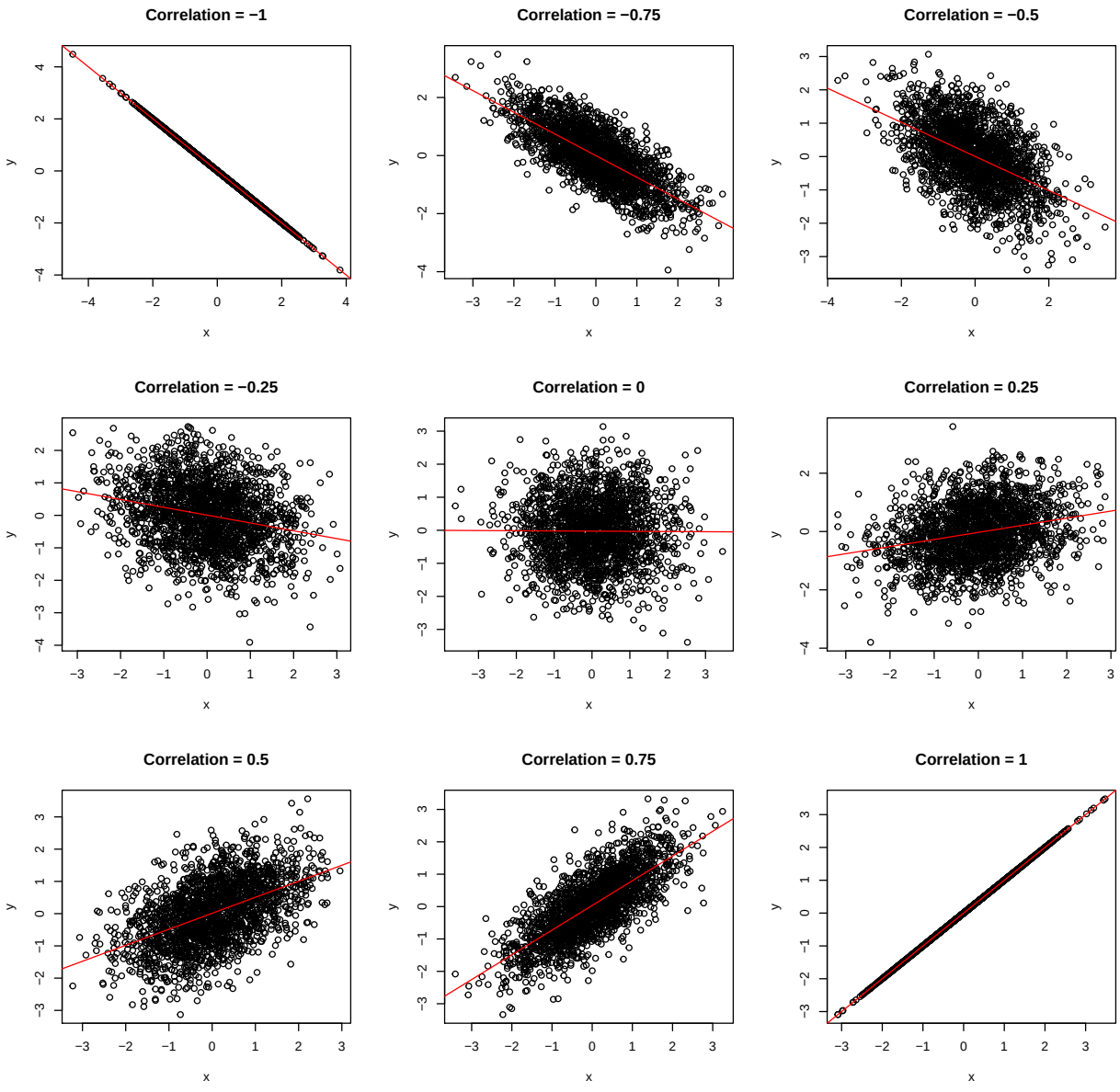

Figure 1: Plots for different correlation values