ST 437/537: Applied Multivariate and Longitudinal Data Analysis

Random Vectors and Multivariate Summary Statistics

Arnab Maity
NCSU Department of Statistics
SAS Hall 5240 919-515-1937 amaity[at]ncsu.edu

Review of univariate framework

Consider the `iris` data in `R`. For simplicity, let us only consider the `setosa` species and the `Sepal.length` variable. Consider the question

> *Estimate the mean sepal.length of the setosa flower.*

Such a statistical problem has four main components.

1. **Population:** A group of individuals/objects/items of interst.
   In our example, population consists of *all* setosa flowers.
2. **Parameter:** A summary of the population we want to know about.
   In our case, the parameter is the true mean of sepal length of all setosa flowers.
3. **Sample:** A subset of the population. Naturally, it is often impossible to observe data on the whole population due to time/resource constraints. Thus, we often collect data on a subset of the population.
   In our example, the sample consists of 50 setosa flowers for which data were collected.
4. **Statistic:** A summary computed from a sample. We use these to *estimate* the unknown parameter.
   In our case, we estimate the population mean by the *sample mean*.

The estimation framework relies on the concept of *random sample*. Specifically, we assume that

- The population mean (mean sepal length) is μ and the population variance is σ^2.
- We collect a random sample $X_1, X_2, \ldots, X_n$, where X_i denotes the `sepal length` of i-th flower.

Notice that $X_1, \ldots, X_n$ are random variables (that is, we define these before we see the actual numbers). We assume that $X_1, \ldots, X_n$ form a random sample.

Random sample

The random variables $X_1, \ldots, X_n$ are called a random sample of size n if $X_1, \ldots, X_n$ are *independent* and each X_i has the *same distribution*.

Thus we have

$$E(X_i) = \mu \text{ and } var(X_i) = \sigma^2 \text{ for all } i.$$

Since we want to know about the poulation mean μ, a natural way to estimate this is to use the sample mean

$$\bar{X} = (X_1 + \ldots + X_n)/n.$$

Specifically, we call $\bar{X}$ an *estimator* of μ.

Estimator

An estimator is a formula/rule that one can apply to *any* possible sample. It does not depend on the parameter.

Now we consider the actual sample (observed numeric data) at hand. We have 50 observations (values taken from the `iris` dataset):

$$x_1 = 5.1, x_2 = 4.9, \ldots, x_{50} = 5.$$

Thus the observed value of the sample mean for this sample is

$$\bar{x} = 5.006.$$

The specific value $\bar{x} = 5.006$ is called an **estimate** of μ.

Estimate

An estimate is a numeric value that is obtained by applying an estimator to a specific sample at hand.

Notice that we used lower case letters (e.g., x_i) to denote the observed data but upper case letters (e.g., X_i) to denote random varibles. We will use this convention throughout this course to differtiate between random variables and observed (numeric) values of the random variables in a particular sample.

Typically, along with the estimate, one also reports the *standard error* of the estimate.

Standard error

Standard error of an estimator is defined as

$$SE(Estimator) = \sqrt{var(Estimator)}.$$

In our case, the standard error of $\bar{X}$ is computed as

$$SE(\bar{X}) = \sqrt{var(\bar{X})} = \sqrt{\sigma^2/n}.$$

Notice that $SE(\bar{X})$ depends on σ^2, the unknown population variance. In practice, we estimate the population variance σ^2 by the observed *sample variance* s^2. Thus the estimated standard error is

$$\widehat{SE}(\bar{X}) = \sqrt{s^2/n}.$$

In R, we can compute the estimate and its standard error as follows.

```
# Get the species
species <- iris$Species

# Only take the setosa flowers
setosa <- iris[species == "setosa", 1:4]

# Extract only sepal.length (the first column)
SL <- setosa[, 1]

# sample size
n <- length(SL)

# Sample mean
xbar.SL <- mean(SL)
xbar.SL
```

```
## [1] 5.006
```

```
# Sample variance
s2.SL <- var(SL)

# Standard error of xbar
SE <- sqrt(s2.SL/n)
SE
```

```
## [1] 0.04984957
```

Random vector and sample mean

Let us now consider the multivariate problem:

> *Estimate the mean of all the four variables: Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width for the setosa flowers.*

In this case, our parameter of interest is the *mean vector*

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{SL} \\ \mu_{SW} \\ \mu_{PL} \\ \mu_{PW} \end{pmatrix},$$

where μ_{SL} = population mean of `sepal length`, μ_{SW} = population mean of `sepal width`, and so on. Thus the parameter is a 4×1 vector.

The data we observe each flower is also recorded as vectors. Specifically, we have a set of *random vectors*.

Random vector

The vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is called a random vector if each element X_i is a random variable.

Specifically, we have a set of random vectors

$$\boldsymbol{X}_1 = \begin{pmatrix} \text{sepal length of 1st flower} \\ \text{sepal width of 1st flower} \\ \text{petal length of 1st flower} \\ \text{petal width of 1st flower} \end{pmatrix} = \begin{pmatrix} SL_1 \\ SW_1 \\ PL_1 \\ PW_1 \end{pmatrix}$$

$$\vdots$$

$$\boldsymbol{X}_n = \begin{pmatrix} SL_n \\ SW_n \\ PL_n \\ PW_n \end{pmatrix}$$

Similar to the univariate case, the estimator of $\boldsymbol{\mu}$ is the *sample mean*.

Sample mean

Given a set of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, the sample mean is defined as

$$\bar{\boldsymbol{X}} = \frac{1}{n}(\boldsymbol{X}_1 + \ldots + \boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

The observed data (numeric data for our particular sample) are

$$\boldsymbol{x}_1 = \begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}, \ldots, \boldsymbol{x}_{50} = \begin{pmatrix} 5 \\ 3.3 \\ 1.4 \\ 0.2 \end{pmatrix}.$$

Thus the estimate of $\boldsymbol{\mu}$ is

$$\bar{\boldsymbol{x}} = \frac{1}{n}(\boldsymbol{x}_1 + \ldots + \boldsymbol{x}_n).$$

```
xbar <- colMeans(setosa)
xbar
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.006        3.428        1.462        0.246
```

How to quantify the variability in $\bar{\boldsymbol{X}}$? To understand this, we need to understand how to quantify the variability of a random vector.

Variance-covariance matrix

Suppose X_1 and X_2 are two scalar random variables. One way to measure the degree of (linear) relationship between X_1 and X_2 is to compute the *covariance* between them.

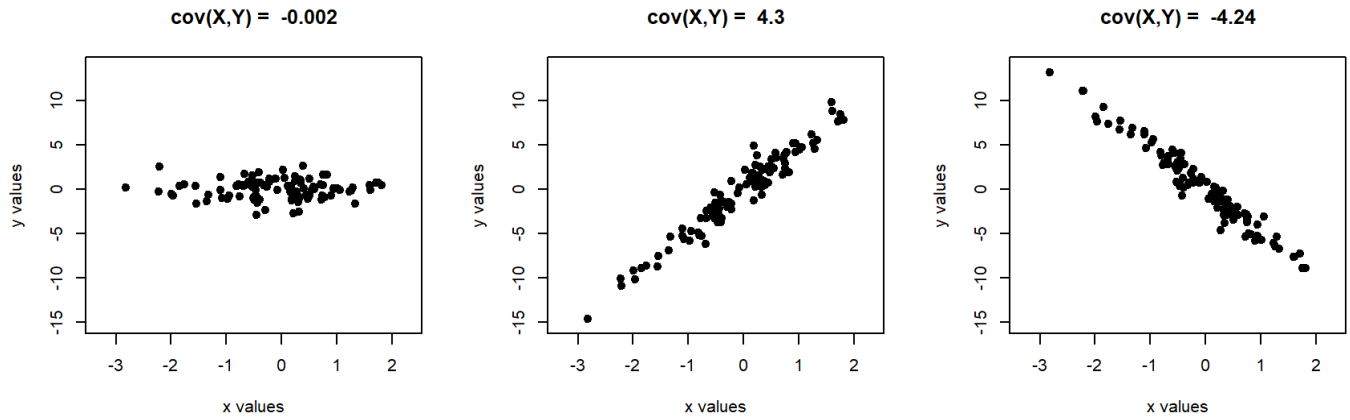
Covariance

We define the covariance between two random variables X_1 and X_2 as

$$cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)],$$

where $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$.

Thus $cov(X_1, X_2)$ takes positive values if larger values of X_1 pair with larger values of X_2, and takes negative if larger values of X_1 pair with smaller values of X_2. Zero or "small" values of covariance indicate that there is no linear relationship (i.e., slope is zero) between X_1 and X_2.



Notice that each if the random variable also has its own variance, that is, $var(X_1)$ and $V(X_2)$. Thus to get a complete picture of variability of X_1 and X_2, we need to look at all these quantities:

$$var(X_1), var(X_2), \text{ and } cov(X_1, X_2).$$

Thus, as the number of variables increases, the number of such quantities increases as well.

In the multivariate world, there is a nice way to summarize the variability of a set of random variables using matrices. To start, let us consider a random vector

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

The "variability" of $\boldsymbol{X}$ can be summarized by the 2×2 matrix

$$\boldsymbol{\Sigma} = cov(\boldsymbol{X}) = \begin{pmatrix} var(X_1) & cov(X_1, X_2) \\ cov(X_2, X_1) & var(X_2) \end{pmatrix}.$$

This matrix is called the variance-covariance matrix of $\boldsymbol{X}$.

Variance-covariance matrix

Suppose we have a $p \times 1$ random vector $\boldsymbol{X} = X_1, \ldots, X_p)^T)$. The variance-covariance matrix of $\boldsymbol{X}$ is defined as

$$\boldsymbol{\Sigma} = cov(\boldsymbol{X}) = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \ldots & cov(X_1, X_p) \\ cov(X_2, X_1) & var(X_2) & \ldots & cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_p, X_1) & cov(X_p, X_2) & \ldots & var(X_p) \end{pmatrix}.$$

Typically, the population variance-covariance matrix is unknown. We can estimate $\boldsymbol{\Sigma}$ by the *sample variance-covariance matrix*, denoted by $\boldsymbol{S}$.

Sample covariance matrix

Given a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, we compute the sample covariance $\boldsymbol{S}$ as

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T.$$

In `R`, we can compute $\boldsymbol{S}$ directly by using the formula above or using the `cov()` function. In our specific example, we can compute $\boldsymbol{S}$ as below (rounded to 4 decimal).

```
S <- cov(setosa)
round(S, 4)
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length       0.1242      0.0992       0.0164      0.0103
## Sepal.Width        0.0992      0.1437       0.0117      0.0093
## Petal.Length       0.0164      0.0117       0.0302      0.0061
## Petal.Width        0.0103      0.0093       0.0061      0.0111
```

Much like the univariate case, we can compute

$$cov(\bar{\boldsymbol{X}}) = \boldsymbol{\Sigma}/n,$$

and we can estimate this quantity by replacing $\boldsymbol{\Sigma}$ by its estimator $\boldsymbol{S}$, that is,

$$\widehat{cov(\bar{\boldsymbol{X}})} = \boldsymbol{S}/n.$$

In our example, we estimate $cov(\bar{\boldsymbol{X}})$ as follows.

```
# Sample size (number of setosa flowers)
n <- nrow(setosa)

# S/n, rounded to 5 digits
round(S/n, 5)
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length      0.00248     0.00198      0.00033     0.00021
## Sepal.Width       0.00198     0.00287      0.00023     0.00019
## Petal.Length      0.00033     0.00023      0.00060     0.00012
## Petal.Width       0.00021     0.00019      0.00012     0.00022
```

Linear combination of variables

General result: If $\boldsymbol{Y}$ is a random vector with mean vector $\boldsymbol{\mu}$ and variance covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$, and $\boldsymbol{A}$ is a matrix, then

$E(\boldsymbol{AY}) = \boldsymbol{A\mu}$ *and* $cov(\boldsymbol{AY}) = \boldsymbol{A\Sigma_Y A}^T$.

This result does *not* depend of the distribution of $\boldsymbol{Y}$.

Application: Often we want to compute differences between means of several variables. For example, suppose we want to know how different `sepal width`, `petal length` and `petal width` are from `sepal length` *on average*. Specifically, we want to estimate

$$\begin{pmatrix} \mu_{SW} - \mu_{SL} \\ \mu_{PL} - \mu_{SL} \\ \mu_{PW} - \mu_{SL} \end{pmatrix}.$$

These are typically called *contrasts*.

Recall that our original parameter is the mean vector

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{SL} \\ \mu_{SW} \\ \mu_{PL} \\ \mu_{PW} \end{pmatrix}.$$

Thus the vector of contrasts can be written as

$$\begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{SL} \\ \mu_{SW} \\ \mu_{PL} \\ \mu_{PW} \end{pmatrix} = \boldsymbol{A\mu}$$

Since $\bar{\boldsymbol{X}}$ is an estimator of $\boldsymbol{\mu}$, we can simply replace $\boldsymbol{\mu}$ in the quantity above by $\bar{\boldsymbol{X}}$, and say that

$\boldsymbol{A\bar{X}}$ *is an estimator of* $\boldsymbol{A\mu}$.

We can compute the variance-covariance matrix of this estimator as

$$cov(\boldsymbol{A\bar{X}}) = \boldsymbol{A}cov(\bar{\boldsymbol{X}})\boldsymbol{A}^T = \boldsymbol{A}(\boldsymbol{\Sigma}/n)\boldsymbol{A}^T.$$

Since $\boldsymbol{\Sigma}$ is unknown, we can replace $\boldsymbol{\Sigma}$ by $\boldsymbol{S}$:

$$\widehat{cov(\boldsymbol{A\bar{X}})} = \boldsymbol{A}(\boldsymbol{S}/n)\boldsymbol{A}^T.$$

In our example, we demonstrate these results as follows.

```
# Define the coefficient/contrast matrix A
A <- cbind(c(-1, -1, -1), c(1, 0, 0),
           c(0, 1, 0), c(0, 0, 1))
A
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   -1    1    0    0
## [2,]   -1    0    1    0
## [3,]   -1    0    0    1
```

```
# Estimate A\mu by A X-bar
A %*% xbar
```

```
##        [,1]
## [1,] -1.578
## [2,] -3.544
## [3,] -4.760
```

```
# Estimate the variance-covariance matrix
A %*% (S/n) %*% t(A)
```

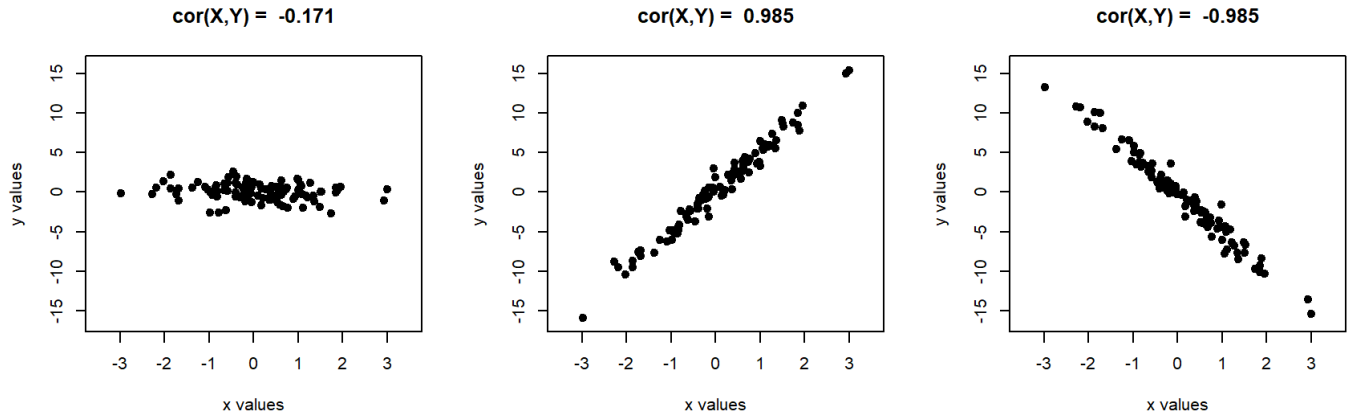
```
##              [,1]         [,2]        [,3]
## [1,] 0.0013901224 0.0004075102 0.000480000
## [2,] 0.0004075102 0.0024339592 0.002072653
## [3,] 0.0004800000 0.0020726531 0.002293878
```

Correlation

A disadvantage of covariance is that it is unbounded, and depends on the unit of measurement. A better measure of linear relationship between two random variables X_1 and X_2 is the correlation coefficient:

$$cor(X_1, X_2) = \frac{cov(X_1, X_2)}{\sqrt{var(X_1)}\sqrt{var(X_2)}}.$$

Correlation coefficient is bounded between -1 and 1. Large positive values indicate a strong positive relationship, and vice versa. Small values indicate absence of no linear relationship.



Now suppose we have a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. The correlation matrix of $\boldsymbol{X}$ is

$$cor(\boldsymbol{X}) = \begin{pmatrix} 1 & cor(X_1, X_2) & \ldots & cor(X_1, X_p) \\ cor(X_2, X_1) & 1 & \ldots & cor(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ cor(X_p, X_1) & cor(X_p, X_2) & \ldots & 1 \end{pmatrix}.$$

Note that the diagonal entries are 1 since $cor(X_i, X_i) = 1$.

Typically, $cor(\boldsymbol{X})$ is unknown and can be estimated using the sample by the *sample correlation* matrix $\boldsymbol{R}$.

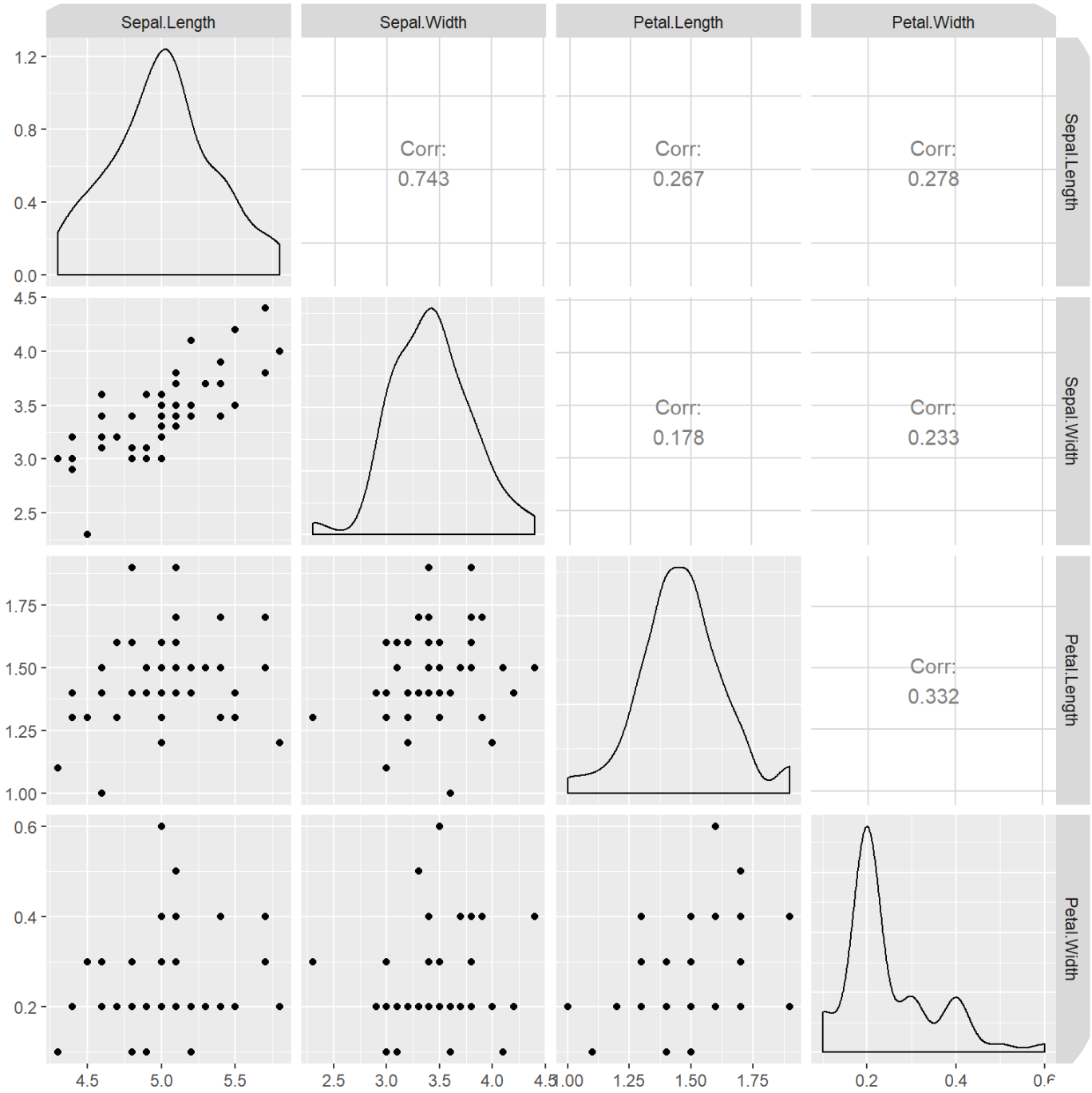
Given sample data, we can use the `cor()` function to compute $\boldsymbol{R}$.

```
# Sample correlation matrix
R <- cor(setosa)
R
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
## Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
## Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
## Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

Notice that `Sepal.Length` and `Sepal.Width` have high correlation compared to other pairs. We can visualize this phenomenon by using a “pairs-plot”.

```
library(ggplot2)
library(GGally)
ggpairs(setosa)
```



We will learn about various plotting techniques in a later chapter.

Main page: **ST 437/537: Applied Multivariate and Longitudinal Data Analysis (https://maityst537.wordpress.ncsu.edu/)**