1 Covariance and correlation

1.1 Covariance

Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, it's a measure of the degree to which two variables are linearly associated.

A covariance matrix is a square matrix that shows the covariance between many different variables. This can be a useful way to understand how different variables are related in a dataset.

1.1.1 Covariance between two variables

```
data <- data.frame(math = c(84, 82, 81, 89, 73, 94, 92, 70, 88, 95),
                   science = c(85, 82, 72, 77, 75, 89, 95, 84, 77, 94),
                   history = c(97, 94, 93, 95, 88, 82, 78, 84, 69, 78))

# create the covariance matrix for the two varibles "math" and "science"
cov(data[, c("math", "science")])
```

```
##             math  science
## math    72.17778 36.88889
## science 36.88889 62.66667
```

Explanation of the covariance matrix in the following section.

1.1.2 Covariance between multiple variables

```
# First, we'll create a data frame that contains the test scores of 10 different
# students for three subjects: math, science, and history.
data <- data.frame(math = c(84, 82, 81, 89, 73, 94, 92, 70, 88, 95),
                   science = c(85, 82, 72, 77, 75, 89, 95, 84, 77, 94),
                   history = c(97, 94, 93, 95, 88, 82, 78, 84, 69, 78))

# create the covariance matrix
cov(data)
```

```
##              math   science   history
## math     72.17778  36.88889 -27.15556
## science  36.88889  62.66667 -26.77778
## history -27.15556 -26.77778  83.95556
```

The numbers on the diagonal of the covariance matrix are the variances of the different variables in the data frame. Meaning that:

- The math scores have a variance of 72.177
- The science scores have a variance of 62.666
- The history scores have a variance of 83.955

The other values in the matrix represent the covariances between the various subjects. For example:

- The covariance between the math and science scores is 36.89
- The covariance between the math and history scores is -27.16
- The covariance between the science and history scores is -26.78

A positive number for covariance indicates that two variables tend to increase or decrease in tandem. For example, math and science have a positive covariance (36.89), which indicates that students who score high on math also tend to score high on science. Conversely, students who score low on math also tend to score low on science.

A negative number for covariance indicates that as one variable increases, a second variable tends to decrease. For example, math and history have a negative covariance (−27.16), which indicates that students who score high on math tend to score low on history. Conversely, students who score low on math tend to score high on history.

```
# if you want to make sure that you only calculate the correlation numerical
# variables you can use this function:
cor(data[, unlist(lapply(data, is.numeric))])
```

```
##               math    science    history
## math     1.0000000  0.5484986 -0.3488448
## science  0.5484986  1.0000000 -0.3691743
## history -0.3488448 -0.3691743  1.0000000
```

1.2 Correlation

One way to quantify the relationship between two variables is to use the Pearson correlation coefficient, which is a measure of the linear association between two variables. It always takes on a value between −1 and 1 where:

−1 indicates a perfectly negative linear correlation between two variables 0 indicates no linear correlation between two variables 1 indicates a perfectly positive linear correlation between two variables

1.2.1 Correlation between two variables

To calculate the correlation between two variables you can use the `cor(x, y)` function.

```
#create data frame
df <- data.frame(A = c(2, 3, 3, 5, 6, 9, 14, 15, 19, 21, 22, 23),
                 B = c(23, 24, 24, 23, 17, 28, 38, 34, 35, 39, 41, 43),
                 C = c(13, 14, 14, 14, 15, 17, 18, 19, 22, 20, 24, 26),
                 D = c(6, 6, 7, 8, 8, 8, 7, 6, 5, 3, 3, 2))
cor(df$A, df$B)
```

```
## [1] 0.9279869
```

This means that there is a strong positive linear correlation between the two variables, because it is almost 1 (1 means that there is a perfect positive linear correlation between two variables) 0.93

1.2.2 Correlation between multiple variables

You can calculate the correlation between a subset of the variables in your dataset e.g.:

```
cor(df[c("A", "B", "D")])
```

```
##            A          B          D
## A  1.0000000  0.9279869 -0.7915488
## B  0.9279869  1.0000000 -0.7917973
## D -0.7915488 -0.7917973  1.0000000
```

The way to interpret the output is as follows:

- The correlation between A and B is 0.9279869.
- The correlation between A and D is −0.7915488.
- The correlation between B and D is −0.7917973.

Meaning that A and D have a somewhat strong negative linear correlation. B and D also have some strong negative linear correlation and lastly A and B have a strong positive linear correlation.

You can also calculate the correlation all of the variables in your dataset e.g.:

```r
cor(df)
```

```
##            A          B          C          D
## A  1.0000000  0.9279869  0.9604329 -0.7915488
## B  0.9279869  1.0000000  0.8942139 -0.7917973
## C  0.9604329  0.8942139  1.0000000 -0.8063549
## D -0.7915488 -0.7917973 -0.8063549  1.0000000
```

```r
# if you want to make sure that you only calculate the correlation numerical
# variables you can use this function:
cor(df[, unlist(lapply(df, is.numeric))])
```

```
##            A          B          C          D
## A  1.0000000  0.9279869  0.9604329 -0.7915488
## B  0.9279869  1.0000000  0.8942139 -0.7917973
## C  0.9604329  0.8942139  1.0000000 -0.8063549
## D -0.7915488 -0.7917973 -0.8063549  1.0000000
```