What's the difference between the correlation and covariance matrix?

Francis Huang

January 19, 2017

Variance/Covariance

To start off, the sample variance formula is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

First of all, $x - \overline{x}$ is a deviation score (deviation from what? deviation from the mean). Summing the deviations will just get us zero so the deviations are squared and then added together. The numerator of this formula is then called the **sum of squared deviations** which is literally what it is. This is not yet what we refer to as the variance (s^2). We have to divide this by n - 1 which is the sample degrees of freedom.

If you have two variables, x and y, those two variables can covary. The formula is similar– instead of squaring the deviation scores, the product of the deviation scores of the two variables are used.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

The numerator is also called the sum of **cross products** (which is what it is). Then dividing this by n - 1 is the **covariance**. The covariance of a variable with itself is also the **variance** which makes sense (instead of the cross product, you are multiplying the deviance with itself or just squaring it).

$$cov(x,x) = s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})}{n-1}$$

That is pretty useful to know. However, the covariance though is not easy to interpret because it is dependent on the **scale** of your variables. For example, if you get the covariance of height and weight– one is measured in inches (or cm) and the other in pounds (or kg). Here's an example (not height or weight):

```
x <- c(12, 15, 20, 25, 30)
y <- c(2, 6, 8, 10, 12)
mean(x)
```

```
## [1] 20.4
```

```
mean(y)
```

```
## [1] 7.6
```

```
var(x)
```

```
## [1] 53.3
```

```
var(y)
```

```
## [1] 14.8
```

```r
cov(x, y) #gets one covariance at a time.
```

```
## [1] 27.2
```

We can put these two variables together in a data frame and estimate the covariance from there.

```r
df <- data.frame(x, y)
cov(df) #same result
```

```
##      x    y
## x 53.3 27.2
## y 27.2 14.8
```

Correlation

What then is the relationship with the correlation matrix? One way to think about it is that the covariance matrix is a bit hard to interpret (the covariances) because they are a mix of different units of measure. A way we get around that is standardizing the measures by converting them to z scores:

$$z - scores = \frac{(x_i - \overline{x})}{SD_x}$$

The scores then have a distribution with a M = 0 and SD = 1 (w/c also means a variance of 1). NOTE: how we can access variables in the data frame using the $ sign.

We can convert by using:

```r
zx <- ( (df$x) - mean(df$x) ) / sd(df$x)
zy <- ( (df$y) - mean(df$y) ) / sd(df$y)

zx
```

```
## [1] -1.15057698 -0.73965663 -0.05478938  0.63007787  1.31494512
```

```r
zy
```

```
## [1] -1.4556507 -0.4159002  0.1039750  0.6238503  1.1437255
```

NOTE: in R, a function to convert raw scores to z scores is the `scale` function.

```r
zx2 <- scale(df$x)
zy2 <- scale(df$y)
zx2
```

```
##             [,1]
## [1,] -1.15057698
## [2,] -0.73965663
## [3,] -0.05478938
## [4,]  0.63007787
## [5,]  1.31494512
## attr(,"scaled:center")
## [1] 20.4
## attr(,"scaled:scale")
## [1] 7.300685
```

```r
zy2
```

```
##            [,1]
## [1,] -1.4556507
```

```
## [2,] -0.4159002
## [3,]  0.1039750
## [4,]  0.6238503
## [5,]  1.1437255
## attr(,"scaled:center")
## [1] 7.6
## attr(,"scaled:scale")
## [1] 3.847077
```

If you want to scale the whole dataset:

```
zdf <- scale(df)
zdf
```

```
##                x          y
## [1,] -1.15057698 -1.4556507
## [2,] -0.73965663 -0.4159002
## [3,] -0.05478938  0.1039750
## [4,]  0.63007787  0.6238503
## [5,]  1.31494512  1.1437255
## attr(,"scaled:center")
##    x    y
## 20.4  7.6
## attr(,"scaled:scale")
##        x        y
## 7.300685 3.847077
```

NOTE: These variables now have a mean of 0 and sd of 1 (also a variance of 1).

A one unit change is a one standard deviation change. NOTE: this is how you interpret standardized beta coefficients in regression. These new measures are now 'unitless'.

If you get the covariance of the two standardized scores, that will be the correlation (or r),

```
cov(zx, zy)
```

```
## [1] 0.9684438
```

```
### You can compare if we just get compute the correlation using the raw scores
cov(zdf)
```

```
##           x         y
## x 1.0000000 0.9684438
## y 0.9684438 1.0000000
```

```
cor(df)
```

```
##           x         y
## x 1.0000000 0.9684438
## y 0.9684438 1.0000000
```

The result is the same. Can we convert a covariance matrix to a correlation matrix.

$$cor(x, y) = \frac{cov(x, y)}{sd(x) \times sd(y)}$$

You can take the variances from the covariance matrix (the diagonal) and then take the square root and those will be the standard deviations.

```r
#check
cov(df)
```

```
##      x    y
## x 53.3 27.2
## y 27.2 14.8
```

```r
sqrt(53.3) #see diagonal
```

```
## [1] 7.300685
```

```r
sd(df$x)
```

```
## [1] 7.300685
```

```r
sd(df$y)
```

```
## [1] 3.847077
```

So to convert the covariance of 27.2, we divide it by the product of sd(x) and sd(y).

```r
27.2 / (sd(df$x) * sd(df$y))
```

```
## [1] 0.9684438
```

Think about it: Can you then convert a correlation matrix to a covariance matrix if all you had is the correlation matrix?