Week 5 Notes: Covariance and Correlation

D. ODay

10/27/2021

Reading Notes

Packages needed

- mvtnorm

```
library('mvtnorm')
```

Covariance

We know that covariance measures the spread of a random variable. **Covariance** measures how two random variables vary together. Unlike variance, covariance can be negative or positive.
If two random variables are independent, their covariance is 0. If positive, they vary in teh same direction and if negative, they vary in opposite directions.

$$Cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)(E(Y)$$

This is basically the expected value of

Issues with Covariance

The biggest problem is with how arbitrary the units are. Say we measure a NFL player's vertical in inches and it's 12.4. If we converted this same measurement to centimeters the Covariance would change because of the unit change. This magnitude issue is not helpful for intuition. **The most informative part of this metric is the sign**.

Properties of Covariance

$$Cov(X,X) = E(X - E(X))^2 = Var(X)$$
$$Cov(X,Y) = Cov(Y,X)$$
$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

This is interesting since we know if X and Y are independent random variables, then E(XY) = E(X)E(Y) which also means the covariance is 0. If dependent, we can find E(X) and E(Y) like we usually do. We can use the 2D-Lotus calculation to find E(XY).

We know from earlier, that the variance of a constant is 0 since constants don't vary at all. Covariance is similar:

$$Cov(X,c) = 0$$

Remember, these case may not always be independent. We also have the following:

$$Cov(X, Y + Z) = Cov(X, Y + Cov(X, Z)$$

$$Cov(\sum_{i=1} X_i \sum_{j=1} Y_i) = \sum_{i,j} Cov(X_i Y_j)$$

In other words, to get the covariance of the sum of X's with the sum of Y's, you need the Covariance of every X with every Y (X_1 with Y_{10}, X_7 with Y_5, etc.).

Now we need the sum of Variances:

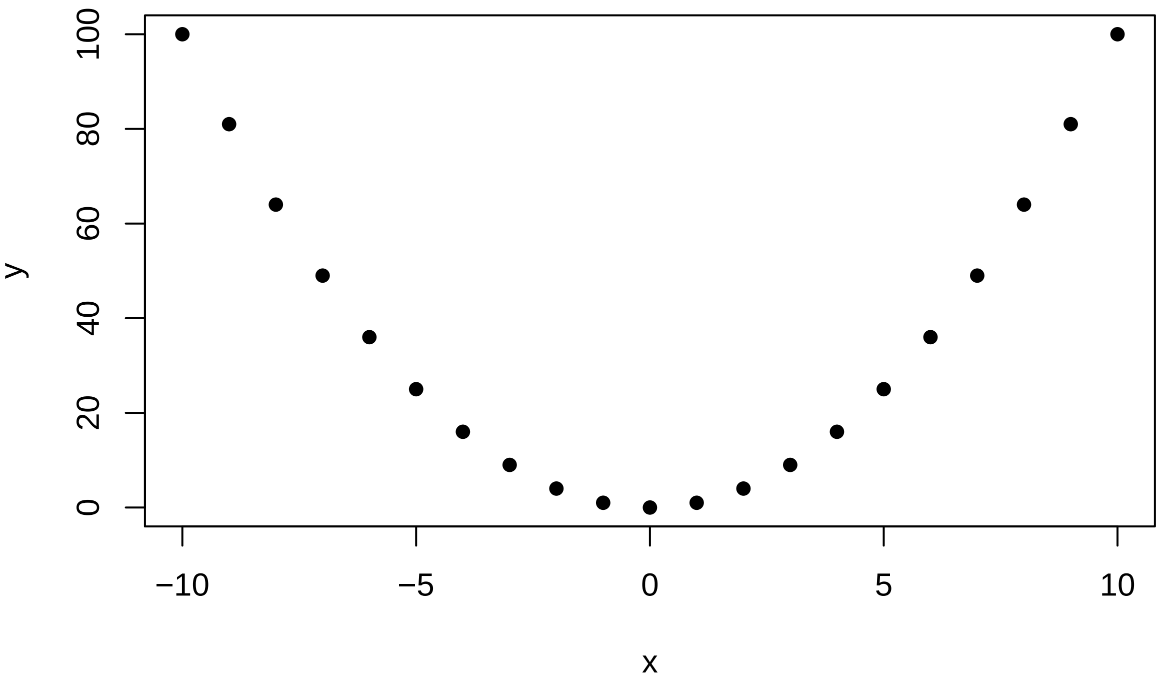
$$Var(X + Y) = Cov(X + Y, X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

In terms of stocks, lets gather some intuition on what $2Cov(X, Y)$ really means. If two stocks move together, they are viewed as riskier (if one does bad, the other has a great change of doing bad). This is an example of a positive covariance. This, accoriding to the formula above, will add a positive value to the variance. this covariance term can also be noted as the 'interactive' variance.

Also, remember **if two random variables are independent (covaraiance = 0)** then the sumb is just the sum of their variances. However, if two random variables have a covariance of 0, that does not necessarily imply that they are independent. This is because **covariance measure linear independence**. The two variables could have a quadratic relationship. Example shown below:

```r
#define x and y = x^2
x = -10:10
y = x^2

#plot x and y; clearly a relationship
plot(x, y, pch = 16)
```



```
cov(x,y)
```

```
## [1] 0
```

Correlation

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma(x)\sigma(y)}$$

Characteristics of Correlation

It is bounded between -1 and 1. Basically, this is saying the correlation grow in magnitude (away from 0) then the two variables has a stronger relationship.
independence implies a correlation of 0.
The reason we may prefer Correlation to Covariance is not only does it give us the direction of the relationship, but the strength of the relationship as well. The units doesn't matter in this situation. We can think of this as standardizing the Covariance.

Remember: correlation does not mean causation.

Transformations

We are interested in finding the expectation AND the distribution of teh transformation of a random variable. Remember, Y = X+4 is still considered a transformation of X. Thus, if this was the case, we would be curuious about the mean and distribution of Y. We can find the PDF of Y with the following:

$$f(y) = f(x)\frac{dx}{dy}$$

In other words, the PDF of Y is the PDF of X where $\frac{dx}{dy}$ derivative of X, in terms of Y. This makes sense since we're looking for the PDF of Y. Here are the steps for the above formula:

1. We write the pdf of X, in terms of Y. In other words, $2X = Y$ becomes $\frac{Y}{2}$.
2. Now that X is by itself, we derive X with respect to Y. Thus, $\frac{Y}{2}$ would be $\frac{1}{2}$.
3. We take what we found in part 2 and multiplt it by the PDF we got in part 1

Tip: Sometimes, finding $\frac{dx}{dy}$ can be difficult. If so, find $\frac{dy}{dx}$ and invert. That will five us $\frac{dx}{dy}$.

The same transformation formula applies to a vector of Random Variables as well. f(y) and f(x) would represent the joint PDF of all the random variables. The $\frac{dx}{dy}$ would become the Jacobian Matrix. This is the matrix of all possible partial derivatives.

Convolutions

Transformations are concerned with finding the distribution of a transformed random variable. With Convolutions, we are more interested in finding the distribution of a sum of random variable.

Let's say that we have two independent random variables X and Y, and we want ot find the distribution of Z, where Z is the sum of X and Y. We also know $X + Y =$ Z. We can find the PMF with the following:

$$P(Z = z) = \sum_{x} P(X = x)P(Y = z - x)$$

Thinking about this, we are fixing the X value to be x. We also know that $y = z - x$. When we sum up over all possible values of x, we get out PMF Z. The sames for continuous variables:

$$f(z) = \int_{-\infty}^{\infty} f_x(x) f_y(z - x) dx$$

MVN (Multivariate Normal Distribution)

Situation where the Y vector is a linear combination of a vector of random variables X and is Normally distributed. Sn example would be if we let $X \sim N(5, 2)$ and, independently, $Y \sim N(-1, 3)$. The vector (X, Y) would Multivariate Normal. The linear combo for this would be $c_1X + c_2Y$

If Random Variables in the vector are not normal, then the linear combination would not be Normal. We can also have a vector of Normal random variables but their linear combination is not. This can only happen if the random variables are dependent.

Bivariate Normal

If X and Y are bivariate normal, we write:

$$(X, Y), \sim BVN\Big(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \Big)$$

This simply means that $X \sim N(\alpha, \sigma_x^2)$ and $Y \sim N(\beta, \sigma_y^2)$. Because of the correlation forumla, we know that $Cor(X, Y)\sigma_x\sigma_y = Cov(X, Y)$. Thus, dimensions (1,2) and (2,1) are just the covariance. Techincally, all 4 entries are covaraince but remember Cov(X,X) is just Var(X).

The 2 x 1 matrix

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

is know as the **mean matrix** and the second matrix as the **Covariance matrix**. Generally, the covariance matrix is a symmetrical matrix. Using this matrix we get the Bivariate Normal Independence property.

Bivariate Normal Independenc

If X and Y are Bivariate Normal and $Cov(X, Y) = 0$ then X and Y are independent. Although we know that if Covariance is 0, that doesn't meant it's independent, in this case it does imply independence.

R and Mulivariate Normal

We can generate random variables from a Multivariate Normal distribution using `rmvnorm`. We need to specify the two matrices we discussed above: the *mean* matrix and the *Covariance* matrix using the `matrix` command.

```
#replicate
set.seed(110)

#define mean matrix; means of 2 and 1
mean.matrix = matrix(c(2, 1), nrow = 2, ncol = 1)

#define Covariance matrix (variances of 1, Covariance of 1/2)
cov.matrix =  matrix(c(1, 1/2, 1/2, 1), nrow = 2, ncol = 2)

#generate 4 data points for each of the two Normal r.v.'s
rmvnorm(4, mean = mean.matrix, sigma = cov.matrix)
```

```
##          [,1]     [,2]
## [1,] 2.640737 2.416906
## [2,] 3.009398 2.595495
## [3,] 2.558957 1.618070
## [4,] 2.341107 2.422288
```

The first column can represent X realizations while the second column represents Y realizations. Next we can use `dnorm` to find the density, evaluating the joint PDF, at point (1,1); in other words, when the first Normal random variable is at 1 and the second random variable is at 1.

```
#find the density at (1,1)
dmvnorm(c(1, 1), mean = mean.matrix, sigma = cov.matrix)
```

```
## [1] 0.0943539
```

Lecture Notes

Jointly Distributed Random Variables

Example: Insurance

An insurance agency services customers who have both a homeowner's policy and an automobile policy. For each type of poliyc, a deductible amount must be specified. for an automobile policy, the choices are \$100 or \$250 and for the homeowner's policy, the choices are \$0, \$100, or \$200.
Suppose an individual, let's say Bob, is selected at random from the agency's files. Let X be the deductible amount on the auto policy and let Y nbe the deductible amount on the homeowner's policy.
We want to understand the relationship between X and Y. We have the following **joint probability table**:

		y (home)		
		0	100	200
x (auto)	100	.20	.10	.20
	250	.05	.15	.30

We can calculate the following probabilities (remember these are also considered intersections):
$P(X = 100, Y = 0) = 0.2$
$P(X = 250, Y = 0) = 0.05$
This is the same as finding P(Y=0).
For any given two discrete random variavbles, X and Y, the **joint probability mass function** for X and Y is:

$$P(X, Y) = P(X = x, Y = y)$$

###Important Property X and Y are **independent random variables** if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all possible values of x and y.

Thus, in our insurance example, is X+Y independent. No, they are not (do example on own. To show if independent we must do for all pairs. To show if not, we only need to find one pair that doesn't work.).

Continuous

If X and Y are continuous random variables, the **joint probability density function** is:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dx dy$$

And to see if they are **independent** then $f(x, y) = f(x)f(y)$ for all possible values of x and y.

Example of Continuous

Suppose a room is lit with two light bulbs. Let X_1 be the lifetime of the first light bulb and X_2 be the lifetime of the second bulb. Suppose $X_1 \sim Exp(\lambda_1 = 1/2000)$ and $X_2 \sim Exp(\lambda_2 = 1/3000)$. If we assume the lifetimes of the light bulb are independent of each other, find the probability that the room is dark after 4000 hours.

So we know E(X_1) = 2000 hrs and E(X_2) = 3000 hrs. Because we know the lightbulbs function independently, $P(X_1 \leq 4000, X_2 \leq 4000) = P(X_1 \leq 4000)P(X_2 \leq 4000)$ which equals

$$(\int_0^{4000} \lambda_1 e^{-\lambda_1 x_1} dx_1) \cdot (\int_0^{4000} \lambda_2 e^{-\lambda_2 x_2} dx_2) \Rightarrow$$

$$-e^{-\lambda_1 x_1}|_0^{4000} \cdot -e^{-\lambda_2 x_2}|_0^{4000} \Rightarrow$$

$$(1 - e^{\frac{-4000}{2000}})(1 - e^{\frac{-4000}{3000}}) \Rightarrow (1 - e^{-2})(1 - e^{\frac{-4}{3}}) \approx 0.6368$$

```r
pexp(q=4000, rate = 1/2000) * pexp(q=4000, rate = 1/3000)
```

```
## [1] 0.6367416
```

Covariance and Correlation

We will be using the same example we using for Joint probability (The insurance agency

When X and Y ar enot independent, we can use the covariance to assess how strongly they are related to each other. This measures a ***linear relationship***.

Returning to our example before, we know $E(X) = 175$ and $E(Y) = 125$. We get the following result:

x	y	$x - \mu_x$	$y - \mu_y$	P(X=x, Y=y)
100	0	-75	-125	0.2
250	0	75	-125	0.05

This gives us Cov(X,Y) = 1875.

Correlation

Also, known as a *scaled* covariance. Because we know an independent X and Y is 0, then independent variables result in a correlation of 0.

What if $Y = aX + b$? We know the covariance will be $a\sigma_x^2$, the variance of Y is $a^2\sigma_x^2$. Thus the standard deviation is $\sigma_y = |a|\sigma_x$. Therforem=, the corelation would be:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{a\sigma_x^2}{\sigma_x |a| \sigma_x} = \begin{cases} 1 & if\ a > 0 \\ 0 & if\ a < 0 \end{cases}$$

Formula Summary

Covariance

Formula

$$Cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)(E(Y)$$

Properties of Covariance

$$Cov(X,X) = E(X - E(X))^2 = Var(X)$$
$$Cov(X,Y) = Cov(Y,X)$$
$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

If Independent the following equation (only the one following):

$$E(XY) = E(X)E(Y)$$
$$Cov(X,c) = 0$$
$$Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)$$
$$Cov(\sum_{i=1} X_i \sum_{j=1} Y_i) = \sum_{i,j} Cov(X_i Y_j)$$
$$Cov(aX, bY) = abCov(X,Y)$$

$$Cov(X+a, Y+b) = Cov(X,Y)$$

$$Var(X+Y) = Cov(X+Y, X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Correlation

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma(x)\sigma(y)}$$

If $Y = aX + b$:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{a\sigma_x^2}{\sigma_x |a| \sigma_x} = \begin{cases} 1 & if\ a > 0 \\ 0 & if\ a < 0 \end{cases}$$

Transformation

$$f(y) = f(x)\frac{dx}{dy}$$

Bivariate Normal

$$(X,Y), \sim BVN\Big(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \Big)$$

Jointly Distributed Random Variables

Discrete Joint Probability Mass Function

$$P(X,Y) = P(X=x, Y=y)$$

If Independent:

$$P(X=x, Y=y) = P(X=x)P(Y=y)$$

Continuous Joint Probability Density Function

probability density function} is:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)dxdy$$

Getting Marginal PDf from Joint PDF

Simply integrate or sum out the unwanted variable of its supprt which will leave a function with only the desired variable. The following example with $f(x,y) = xy$ where X and Y both run from 0 to 1

$$f(x) = \int_0^1 xy\ dy = \frac{xy^2}{2}\bigg|_0^1 = \frac{x}{2}$$