Lecture 2: Measure of Relations: Covariance and Correlation

Le Wang

U of Oklahoma

January 28, 2019

Motivation

It is important to test whether or not two variables are independent before we should even consider modeling the relationship between them.

But it is not enough, for example, we do not know

1. How large is the relationship?
2. How are they related? What is the direction of the relationship? Negative or Positive?

Roadmap

1. Measurement of Relationship (Magnitude and Direction): **Covariance** and **Correlation**
2. Implementation in R:
   2.1 `-cor()-`,`-cov()-`
   2.2 `-cor.test()-`
3. Applications:
   3.1 Stock Market Returns and Volatility
   3.2 Portfolio Diversification
   3.3 Optimal Forecast Combination
4. Shortcoming of Correlation: linear relationship!
5. Applications of Entropy in Financial Markets

How to Measure?

Consider the following problem:

(x_1, y_1)
(x_1, y_2)
(x_1, y_3)
$\vdots$
(x_2, y_1)
$\vdots$
$\vdots$

Table: Possible Issues

What will this measure look like?

We should measure **comovement**.

1. When one variable varies, the other variable also varies.
2. This number should tell us if they move together, either in the same direction or in the opposite direction
3. If the relationship is strong, the change for both variables should both be large (how do you define large?).

What will this measure look like?

The **first question** we need to answer is

What is the **change**? How can we tell the change from by looking at the following ordered pair (possible values)?

$$(x_1, y_1)$$

We need to find a common benchmark and consider every possible value as a change. The change is then considered as a deviation from this benchmark.

What will this measure look like?

(x_1, y_1)	$(x_1 - a, y_1 - b)$
(x_1, y_2)	$(x_1 - a, y_2 - b)$
(x_1, y_3)	$(x_1 - a, y_3 - b)$
$\vdots$	$\vdots$
(x_2, y_1)	$(x_2 - a, y_1 - b)$
$\vdots$	
$\vdots$	

Table: Deviation Form

What will this measure look like?

The **second question** we need to answer is:

How to measure the **direction**? How can we tell whether or not they are moving in the same or opposite direction?

What will this measure look like?

(x_1, y_1)	$(x_1 - \mu_x, y_1 - \mu_y)$	$(x_1 - \mu_x)(y_1 - \mu_y)$
(x_1, y_2)	$(x_1 - \mu_x, y_2 - \mu_y)$	$(x_1 - \mu_x)(y_2 - \mu_y)$
(x_1, y_3)	$(x_1 - \mu_x, y_3 - \mu_y)$	$(x_1 - \mu_x)(y_3 - \mu_y)$
$\vdots$	$\vdots$	$\vdots$
(x_2, y_1)	$(x_2 - \mu_x, y_1 - \mu_y)$	$(x_2 - \mu_x)(y_1 - \mu_y)$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$

Table: Deviation Form

What will this measure look like?

Measure of the relationship: **Covariance** (Average Deviations!)

$$\mathbb{E}[(x - \mu_x)(y - \mu_y)]$$

The **second moment** of the joint distribution!

How to estimate this? **Sample Covariance**

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

Question: Why would this work?

Covariance and Variance

A special case, consider two variables are identical, in other words, X and X. What will happen to their covariance?

$$\mathbb{E}[(x - \mu_x)(x - \mu_x)]$$

What will this measure look like?

Our **covariance** measure do not tell us the magnitude of the change or comovement. The problem is that it depends on the units of measurement.

What should we do?

What will this measure look like?

Correlation:

$$\rho_{xy} = \mathbb{E}\left[\frac{(x-\mu_x)}{\sigma_x} \cdot \frac{(y-\mu_y)}{\sigma_y}\right] = \frac{\mathbb{E}[(x-\mu_x)(y-\mu_y)]}{\sigma_x \cdot \sigma_y}$$

1. $-1 \leq \rho_{xy} \leq 1$
2. symmetric: $corr(x, y) = corr(y, x)$

Sample Correlation: r_{xy}

$$\frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2\right)\left(\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y})^2\right)}} = \frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}}$$

Properties of Covariance

1. $Cov(X, Y) = Cov(Y, X)$
2. $Cov(X, a) = 0$
3. $Cov(X, X) = Var(X)$
4. $Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$
5. $Cov(aX + bY, cZ) = acCov(X, Z) + bcCov(Y, Z)$

Implementation in R

Both **covariance** and **correlation** are important mathematical constructs that are ubquitous in finance. The scaling is the only that differs between the two. **Covariance** or **correlation matrix** is one of the cornerstones in modern finance.

Let's examine two index returns

1. **SPY**: S&P 500 stock market index
2. **VXX**: Equity Market Volatility tracks the S&P 500 VIX Short-Term Futures Index Total Return.

Implementation in R

Let's obtain the data first

```
library(quantmod)
# S&P 500 stock market index
# the date when VXX was available
getSymbols("SPY", from="2013-01-01")

## [1] "SPY"

# Equity Market Volatility
# track the S&P 500 VIX Short-Term
# Futures Index Total Return

getSymbols("VXX", from="2013-01-01")

## [1] "VXX"

# Generate daily returns

spy<-SPY$SPY.Adjusted
spy_returns<- na.omit(diff(log(spy)))
vxx<-VXX$VXX.Adjusted
vxx_returns <- na.omit(diff(log(vxx)))
```

Implementation in R

```
# Covariance
cov(spy_returns, vxx_returns)

##               VXX.Adjusted
## SPY.Adjusted -0.0002662572

# Correlation
cor(spy_returns, vxx_returns)

##              VXX.Adjusted
## SPY.Adjusted   -0.8408398

# DIY
cov(spy_returns, vxx_returns)/(sd(spy_returns)*sd(vxx_returns))

##              VXX.Adjusted
## SPY.Adjusted   -0.8408398
```

Question: Does this result make sense?

1. Positive or Negative relationship?
2. The VXX is an stock that tracks volatility.

Application 2: Portfolio Diversification

Property: $Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$

$$\rho_{xy} = \frac{\mathbb{E}[(x - \mu_x)(y - \mu_y)]}{\sigma_x \cdot \sigma_y} = \frac{Cov(X, Y)}{sd(X) \cdot sd(Y)}$$

$$Cov(X, Y) = sd(X) \cdot sd(Y) \cdot \rho_{xy}$$

Portfolio:

$$Z = w_1 \cdot X + w_2 \cdot Y$$

Variance of the Portfolio:

$$\begin{aligned} Var(Z) &= Var(w_1 \cdot X + w_2 \cdot Y + 0) \\ &= w_1^2 \cdot Var(X) + w_2^2 \cdot Var(Y) + 2w_1 w_2 Cov(X, Y) \\ &= w_1^2 \cdot Var(X) + w_2^2 \cdot Var(Y) + 2w_1 w_2 \cdot sd(X) \cdot sd(Y) \cdot \rho_{xy} \end{aligned}$$

Portfolio Diversification: Different Scenarios

Variance of the Portfolio

$$Var(Z) = w_1^2 \cdot Var(X) + w_2^2 \cdot Var(Y) + 2w_1 w_2 \cdot sd(X) \cdot sd(Y) \cdot \rho_{xy}$$

To illustrate the point, lets assume $Var(X) = Var(Y)$ and identical weights, $w_1 = w_2 = .5$.

1. **Perfect positive correlation $\rho_{xy} = 1$:**

   $$Var(Z) = 0.25 \cdot Var(X) + 0.25 \cdot Var(X) + 2 \cdot 0.25 \cdot 0.25 \cdot Var(X) \cdot 1 = Var(X)$$

2. **No correlation $\rho_{xy} = 0$:**

   $$Var(Z) = 0.25 \cdot Var(X) + 0.25 \cdot Var(X) + 2 \cdot 0.25 \cdot 0.25 \cdot Var(X) \cdot 0 = 0.5 \cdot Var(X)$$

3. **Perfect Negative correlation $\rho_{xy} = -1$:**

   $$Var(Z) = 0.25 \cdot Var(X) + 0.25 \cdot Var(X) + 2 \cdot 0.25 \cdot 0.25 \cdot Var(X) \cdot (-1) = 0$$

Portfolio Diversification: Different Scenarios

Lesson: If two assets could be found which have perfect negative correlation, then we combine these assets to create a risk free portfolio. These assets create a perfect hedge. This shows that diversification can be thought of as a partial hedge of risks.

Application 3: Optimal Forecast Combination

Suppose that you have two unbiased forecasts from two different models, f_1, f_2. Then the combination of these two are unbiased forecasts as well.

$$f_c = w_1 \cdot f_1 + w_2 \cdot f_2$$

$$\begin{aligned}
\mathbb{E}[f_c] &= \mathbb{E}[w_1 \cdot f_1 + w_2 \cdot f_2] \\
&= w_1 \cdot \mathbb{E}[f_1] + w_2 \cdot \mathbb{E}[f_2] \\
&= 0
\end{aligned}$$

Application 3: Optimal Forecast Combination

Regardless of w_1, w_2, the combined forecast is unbiased. But can we choose the weights that are optimal? In the sense that the forecast error variance is minimized.

$$\begin{aligned} f_c &= w_1 \cdot f_1 + w_2 \cdot f_2 \\ f_c - y &= w_1 \cdot (f_1 - y) + w_2 \cdot (f_2 - y) \\ e_c &= w_1 \cdot e_1 + w_2 \cdot e_2 \end{aligned}$$

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

implies that

$$Var(e_c) = w_1^2 \cdot Var(e_1) + (1 - w_1)^2 Var(e_2) + 2 \cdot w_1 \cdot (1 - w_1) Cov(e_1, e_2)$$

Application 3: Optimal Forecast Combination

$$
\begin{aligned}
\frac{d}{dw_1} Var(e_c) &= \frac{d}{dw_1}[w_1^2 \cdot Var(e_1) + (1-w_1)^2 Var(e_2) + 2 \cdot w_1 \cdot (1-w_1) Cov(e_1, \\
&= 2 \cdot w_1 \cdot Var(e_1) - 2 \cdot (1-w_1) Var(e_2) + 2(1 - 2 \cdot w_1) Cov(e_1, e_2) \\
&= 0
\end{aligned}
$$

implies

$$w_1 = \frac{Var(e_2) - Cov(e_1, e_2)}{Var(e_1) + Var(e_2) - 2Cov(e_1, e_2)}$$

If $Cov(e_1, e_2) = 0$, then it is nothing but the inverse share of the total variance.

Lesson: In practice, Bates and Ganger and a number of researchers recommend constructing the weights excluding the covariance terms since estimates of the covariance terms are often poor.

Distribution of Correlation

Question: How do we know that this correlation is real, but not due to some minor variations ("noise") in the data?

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

Test Statistic of Zero Correlation

(Tricky) Test Statistic (mathematically equilvalent statement)

$$r\sqrt{\frac{N-2}{1-r^2}} \sim t_{N-2}$$

or can be approximated by $N(0, \frac{1}{N-1})$.

Distribution of Correlation

```
# alternative can specify a two-tailed or one-tailed test
# "two.side", "less", or "greater"
cor.test(vxx_returns,spy_returns, alternative = "two.side")

## Warning in z + c(-1, 1) * sigma * qnorm((1 + conf.level)/2):
Recycling array of length 1 in array-vector arithmetic is deprecated.
##  Use c() or as.vector() instead.

##
##  Pearson's product-moment correlation
##
## data:  vxx_returns and spy_returns
## t = -60.643, df = 1524, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8549468 -0.8254899
## sample estimates:
##        cor
## -0.8408398
```

Closer look at correlation

Question: Should x be correlated with x^2?

```
set.seed(123456)
x<-rnorm(100)
x2<-x^2
cor.test(x,x2)

##
##  Pearson's product-moment correlation
##
## data:  x and x2
## t = -0.49281, df = 98, p-value = 0.6233
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2437572  0.1481453
## sample estimates:
##         cor
## -0.04971957
```

Correlation $\neq$ Independence

Correlation captures only **linear** relationship, but not the full dependence. If the relationship is at the higher order (moments), then it cannot detect it.

Examples in finance:

1. Conditional Volatility (second moment):
2. Other nonlinear time series:

Go back to the solution: **Entropy measure!**

Entropy and Predictability of Stock Market Returns

Maasoumi and Racine (2002): *We examine the predictability of stock market returns by employing a new metric entropy measure of dependence with several desirable properties. We compare our results with a number of traditional measures. The metric entropy is capable of detecting nonlinear dependence within the returns series, and is also capable of detecting nonlinear "affinity" between the returns and their predictions obtained from various models thereby serving as a measure of out-of-sample goodness-of-fit or model adequacy. Several models are investigated, including the linear and neural-network models as well as nonparametric and recursive unconditional mean models.*
dfdfdfdf

Entropy and Predictability of Stock Market Returns

$$S_\rho = \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}} \right)^2 dxdy$$

where $f_1 = f(x, y)$ is the joint density and $f_2 = g(x)h(y)$ is the product of the marginal densities.

Entropy and Predictability of Stock Market Returns

```r
library(quantmod)


getSymbols("PEP")

## [1] "PEP"

getSymbols("COKE")

## [1] "COKE"

pepsi<-PEP$PEP.Adjusted
pepsi_returns<- na.omit(diff(log(pepsi)))

coke<-COKE$COKE.Adjusted
coke_returns<- na.omit(diff(log(coke)))

# Correlation

cor(pepsi_returns, coke_returns)

##              COKE.Adjusted
## PEP.Adjusted     0.3426309

# Note that we strongly advocate the use of the integration (default) version o
```

Entropy and Stock Returns

```
library(np)


pepsi_returns<-as.vector(pepsi_returns)
coke_returns<-as.vector(coke_returns)

npdeptest(coke_returns,pepsi_returns,boot.num=30,method="summation")
```

Note that we strongly advocate the use of the integration (default) version of the statistic in applied settings but use the summation (i.e. moment) version below purely by way of demonstration as it is computationally faster.