Handout for Lecture 4

Covariance and Correlation

ECON 340: Economic Research Methods — Instructor: Div Bhagia

1. Let X be the average hours of sleep per day you got last week, and let Y be the average hours you exercised per day last week. You want to look at the relationship between these two variables over the last three weeks. Given values of X and Y fill in the following table.

Week	X_i	Y_i	$(X_i - \mu_X)^2$	$(Y_i - \mu_Y)^2$	$(X_i - \mu_X)(Y_i - \mu_Y)$
1	6	0.5	4	0.01	0.2
2	9	0.3	1	0.09	-0.3
3	9	1	1	0.16	0.4
Total	24	1.8	6	0.26	0.3

Note that $\mu_X = 24/3 = 8$ and $\mu_Y = 1.8/3 = 0.6$.

(a) Calculate the variance of X and Y.

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2 = \frac{6}{3} = 2$$

$$\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu_Y)^2 = \frac{0.26}{3} = 0.087$$

(b) Calculate the covariance and correlation between X and Y.

Covariance:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) = \frac{0.3}{3} = 0.1$$

Correlation:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0.1}{\sqrt{2}\sqrt{0.087}} = 0.24$$

(c) In class, we learned that covariance is positive when two variables move together, meaning that they increase or decrease together. Can you explain how the formula you used in (c) ensures that this is the case? Explain it to your peer.

While calculating the covariance, we add a positive number to the numerator in the formula whenever both variables are either above or below their respective averages. Conversely, we add a negative number to the numerator when one variable is above its average while the other is below. Therefore, a positive covariance indicates that, on average, the variables move together—either both being above or both being below their averages. While a negative covariance indicates that, on average, the variables move in opposite directions—one being above its average while the other is below.

Now say instead of recording the exercise in hours, you had recorded it in minutes. Then your data would look as below, where Z is the average minutes of exercise per day.

Week	X_i	Z_i	$(X_i - \mu_X)^2$	$(Z_i - \mu_Z)^2$	$(X_i - \mu_X)(Z_i - \mu_Z)$
1	6	30	4	36	12
2	9	18	1	324	-18
3	9	60	1	576	24
Total	26	108	6	936	18

(d) Do you think the covariance between sleep and exercise is going to be larger, smaller or the same now that exercise is measured in minutes instead of hours?

The covariance between sleep and exercise will be larger when exercise is measured in minutes instead of hours. This is because covariance is sensitive to the scale of the variables.

(e) Fill in the table above and calculate the covariance and correlation between X and Z.

Note that $\mu_Z = 108/3 = 36$ and $\sigma_Z^2 = 936/3 = 312$.

$$\text{Covariance: } \sigma_{XZ} = \frac{18}{3} = 6 \qquad \text{Correlation: } \rho_{XZ} = \frac{6}{\sqrt{312}\sqrt{2}} = 0.24$$

2. If a study finds a strong positive correlation between the number of houses and house prices across US cities, can we conclude that more housing supply leads to higher house prices? Why or why not? Discuss.

If we observe a strong positive correlation between the number of houses and house prices across US cities, it does not necessarily imply that an increase in housing supply causes higher house prices. In fact, it is possible that higher house prices lead to more housing supply, as builders try to maximize their profits by building in more profitable locations, resulting in a positive correlation between the stock of housing and prices. This is known as reverse causality.

It is also possible that external confounding factors are responsible for the observed positive correlation between housing stock and housing prices. For instance, if certain cities have more desirable amenities, more people may want to live there, leading to higher demand for housing. This, in turn, can result in higher prices and increased housing stock as builders construct more houses to meet the growing demand.