14.30 Introduction to Statistical Methods in Economics

Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 13

Konrad Menzel

March 31, 2009

1 Covariance

The covariance of X and Y is a measure of the strength of the relationship between the two random variables.

Definition 1 *For two random variables X and Y, the covariance is defined as*

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

First, by just applying the definitions, we get

Property 1

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

Property 2

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

Furthermore, we have the following result which is very useful to calculate covariances:

Property 3

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

This is a generalization of the analogous property of variances, and the proof uses exactly the same kind of argument. Let's do an example to see how this result is useful:

Example 1 *Suppose X and Y have joint p.d.f.*

$$f_{XY}(x,y) = \begin{cases} 8xy & \text{if } 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

What is the covariance $\mathrm{Cov}(X,Y)$? - Let's calculate the components which enter according to the right-hand side of the equation in property 7:

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)dxdy = \int_0^1\int_x^1 8x^2y^2 dydx \\
&= \int_0^1 8x^2\left(\int_x^1 y^2 dy\right)dx = \int_0^1 \frac{8}{3}x^2(1-x^3)dx \\
&= \frac{8}{3}\int_0^1 (x^2 - x^5)dx = \frac{8}{3}\left[\frac{x^3}{3} - \frac{x^6}{6}\right]_0^1 \\
&= \frac{8}{3}\left(\frac{1}{3} - \frac{1}{6}\right) = \frac{4}{9}
\end{aligned}$$

Also, by the same steps as above,

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^1 8x\left(\int_x^1 y^2 dy\right)dx = \int_0^1 \frac{8}{3}x(1-x^3)dx \\ &= \frac{8}{3}\int_0^1 (x-x^4)dx = \frac{8}{3}\left[\frac{x^2}{2}-\frac{x^5}{5}\right]_0^1 = \frac{8}{3}\cdot\frac{3}{10} = \frac{4}{5}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 8x^2\left(\int_x^1 y dy\right)dx = \int_0^1 \frac{8}{2}x^2(1-x^2)dx \\ &= 4\int_0^1 (x^2-x^4)dx = 4\left(\frac{1}{3}-\frac{1}{5}\right) = \frac{8}{15}\end{aligned}$$

Putting all pieces together, and applying property 7,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{4}{9} - \frac{8}{15}\cdot\frac{4}{5} = \frac{4\cdot 25 - 32\cdot 3}{225} = \frac{4}{225}$$

We already showed that for two *independent* random variables X and Y, the variance of the sum equals the sum of variances. Here's a generalization to random variables which are not necessariy independent:

Property 4

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

The idea behind the proof is to apply properties 3 and 7 to get

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}[X^2+2XY+Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \left(\mathbb{E}[X^2]-\mathbb{E}[X]^2\right) + \left(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2\right) + 2\left(\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]\right) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)\end{aligned}$$

Property 5 *For random variables X, Y, Z,*

$$\mathrm{Cov}(X, aY+bZ+c) = a\mathrm{Cov}(X,Y) + b\mathrm{Cov}(X,Z)$$

Property 6

$$\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X,Y)$$

Since by the last property, the covariance changes with the scale of X and Y, we would like to have a standardized measure which gives us the *strength* of the relationship between X and Y, and which is not affected by changing, say, the units of measurement of the two variables. The most frequently used measure of that kind is the correlation coefficient:

Definition 2 *The correlation coefficient of X and Y is given by*

$$\rho(XY) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

The correlation coefficient is normalized in a way such that

Property 7

$$-1 \leq \rho(X,Y) \leq 1$$

Broadly speaking, we distinguish three cases

- $\varrho(X,Y) > 0$: "X and Y are positively correlated"
- $\varrho(X,Y) = 0$: "X and Y are uncorrelated"
- $\varrho(X,Y) < 0$: "X and Y are negatively correlated"

Property 8

$$|\rho(X,Y)| = 1 \Leftrightarrow Y = aX + b$$

for some constants $a \neq 0$ and b.

I.e. the absolute value of the correlation coefficient equals 1 if there is a deterministic *linear* relationship between the two random variables. In that case we say that X and Y are *perfectly correlated*.

Remark 1 *A very important principle of data analysis is that the statistical relationship between two random variables does not necessarily correspond to mechanical or causal statements which we would actually want to make based on the data. E.g. we typically observe in data sets that the time people spend working out in the gym is positively correlated with health, but this does not necessarily mean that sports* causes *health to improve. But it could as well be that some people are so unhealthy that they wouldn't even think about going to the gym.*
A more abstract way to see why correlation of X and Y *and* causation of Y by X *are inherently different concepts, notice that the covariance is* symmetric *in X and Y, so we can change their roles. For causality however, we think of a specific* direction *of the relationship, i.e. we could have $X \rightarrow Y$ or "X causes/affects Y" but simultaneously "Y doesn't cause/affect X", so we can't change the roles of X and Y. Therefore:*

CORRELATION DOES NOT EQUAL CAUSATION!

(Much) more about that in your Econometrics class.

1.1 Preview: Regression

Say, we are interested in the relationship between a worker's income Y and her education as measured with years of schooling X (for simplicity let's just assume that both are continuous random variables). Then we can always rewrite the relationship between X and Y as

$$Y = \alpha + \beta X + U$$

where U is a random variable with $\mathbb{E}[U] = 0$ and $\text{Cov}(X,U) = 0$ (in the regression context it will be called the *residual*).
One way of determining the value of the parameters (α, β) is to solve

$$(\alpha, \beta) = \arg\min_{\alpha,\beta} \mathbb{E}[(Y - X\beta - \alpha)^2]$$

Taking first-order conditions with respect to β (notice that expectations are linear in their argument, so we can pass the derivative through the integral), we get

$$\begin{aligned} 0 &= \frac{d}{d\beta}\mathbb{E}[(Y - X\beta - \alpha)^2] \\ &= \frac{d}{d\beta}\int (y - x\beta - \alpha)^2 f_{XY}(x,y)dydx \\ &= \int \frac{d}{d\beta}\left[(y - x\beta - \alpha)^2\right] f_{XY}(x,y)dydx \\ &= \mathbb{E}\left[\frac{d}{d\beta}[(Y - X\beta - \alpha)^2]\right] \\ &= \mathbb{E}\left[2X(Y - X\beta - \alpha)\right] \end{aligned}$$

Similarly, taking first-order conditions with respect to α,

$$0 = \frac{d}{d\alpha}\mathbb{E}[(Y - X\beta - \alpha)^2] = \mathbb{E}\left[\frac{d}{d\alpha}[(Y - X\beta - \alpha)^2]\right] = \mathbb{E}[2(Y - X\beta - \alpha)]$$

Solving the last expression for α,

$$\alpha = \mathbb{E}[Y] - \mathbb{E}[X]\beta$$

Plugging this into the first-order condition for β, we get

$$0 = \mathbb{E}\left[X(Y - X\beta - (\mathbb{E}[Y] - \mathbb{E}[X]\beta))\right] = \mathbb{E}[XY] - \mathbb{E}[X^2]\beta - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]^2\beta$$

so that we can now solve for the parameter

$$\beta = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2] - \mathbb{E}[X]^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

We can now verify that indeed $\mathbb{E}[U] = \mathbb{E}[Y - X\beta - \alpha] = 0$ and $\mathrm{Cov}(X,U) = \mathrm{Cov}(X, Y - X\beta - \alpha) = 0$ (in fact the first follows directly from the first-order condition for α, and the second from the first-order condition for β).
Then the "regression fit" $\alpha + \beta X$ is the part of Y which is correlated with (or "explained" by) X, and U is the part of Y which is uncorrelated with X. The parameters α and β are usually called *regression parameters* or *least-squares coefficients*. Linear regression is the "workhorse" of much of econometrics, and you'll see this in many variations over the course of 14.32 and other econometrics classes.

2 Conditional Expectations

Example 2 *Each year, a firm's R&D department produces X innovations according to some random process, where $\mathbb{E}[X] = 2$ and $\mathrm{Var}(X) = 2$. Each invention is a commercial success with probability $p = 0.2$ (assume independence). The number of commercial successes in a given year are denoted by S. Since we know that the mean of $S \sim B(x,p) = xp$, conditional on $X = x$ innovations in a given year, xp of them should be successful on average.*

The *conditional* expectation of X given Y is the expectation of X taken over the conditional p.d.f.:

Definition 3

$$\mathbb{E}[Y|X] = \begin{cases} \sum_y y f_{Y|X}(y|X) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|X)dy & \text{if } Y \text{ is continuous} \end{cases}$$

Note that since $f_{Y|X}(y|X)$ carries the *random variable* X as its argument, the conditional expectation is also a random variable. However, we can also define the conditional expectation of Y given a particular value of X,

$$\mathbb{E}[Y|X=x] = \begin{cases} \sum_y y f_{Y|X}(y|x) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy & \text{if } Y \text{ is continuous} \end{cases}$$

which is just a number for any given value of x as long as the conditional density is defined.
Since the calculation goes exactly like before, only that we now integrate over the *conditional* distribution, won't do a numerical example (for the problem set, just apply definition). Instead let's discuss more qualitative examples to illustrate the difference between conditional and unconditional examples:

Example 3 ***(The Market for "Lemons")*** *The following is a simplified version of a famous model for the market for used cars by the economist George Akerlof. Suppose that there are three types X of used cars: cars in an excellent state ("melons"), average-quality cars ("average" not in a strict, statistical, sense), and cars in a poor condition ("lemons"). Each type of car is equally frequent, i.e.*

$$P(\text{"}lemon\text{"}) = P(\text{"}average\text{"}) = P(\text{"}melon\text{"}) = \frac{1}{3}$$

The seller and a buyer have the following (dollar) valuations Y_S and Y_B, respectively, for each type of cars:

Type	*Seller*	*Buyer*
"Lemon"	*5,000$*	*6,000$*
"Average"	*6,000$*	*10,000$*
"Melon"	*10,000$*	*11,000$*

The first thing to notice is that for every type of car, the buyer's valuation is higher than the seller's, so for each type of car, trade should take place at a price between the buyer's and the seller's valuations. However, for used cars, quality is typically not evident at first sight, so if neither the seller nor the buyer know the type X of a car in question, their expected valuations are, by the law of iterated expectations

$$\begin{aligned} \mathbb{E}[Y_S] &= \mathbb{E}[Y_S|\text{"}lemon\text{"}]P(\text{"}lemon\text{"}) + \mathbb{E}[Y_S|\text{"}average\text{"}]P(\text{"}average\text{"}) + \mathbb{E}[Y_S|\text{"}melon\text{"}]P(\text{"}melon\text{"}) \\ &= \frac{1}{3}(5,000 + 6,000 + 10,000) = 7,000 \\ \mathbb{E}[Y_S] &= \mathbb{E}[Y_B|\text{"}lemon\text{"}]P(\text{"}lemon\text{"}) + \mathbb{E}[Y_B|\text{"}average\text{"}]P(\text{"}average\text{"}) + \mathbb{E}[Y_B|\text{"}melon\text{"}]P(\text{"}melon\text{"}) \\ &= \frac{1}{3}(6,000 + 10,000 + 11,000) = 9,000 \end{aligned}$$

so trade should still take place.
But in a more realistic setting, the seller of the used car knows more about its quality than the buyer (e.g. history of repairs, accidents etc.) and states a price at which he is willing to sell the car. If the seller can perfectly distinguish the three types of cars, whereas the buyer can't, the buyer should form expectations conditional *on the seller willing to sell at the quoted price.*
If the seller states a price less than $6,000$ dollars, the buyer knows for sure that the car is a "lemon" because otherwise the seller would demand at least $6,000$, i.e.

$$\mathbb{E}[Y_B|Y_S < 6000] = \mathbb{E}[Y_B|\text{"}lemon\text{"}] = 6000$$

and trade would take place. However, if the car was in fact a "melon", the seller would demand at least $10,000$ dollars, whereas the buyer would pay at most

$$\mathbb{E}[Y_B|Y_S \leq 10,000] = \mathbb{E}[Y_B] = 9,000 < 10,000$$

so that the seller won't be able to sell the high-quality car at a reasonable price.
The reason why the market for "melons" breaks down is that in this model, the seller can't credibly assure the buyer that the car in question is not of lower quality, so that the buyer factors the possibility of getting the bad deal into his calculation.

An important relationship between conditional and unconditional expectation is the Law of Iterated Expectations (a close "cousin" of the Law of Total Probability which we saw earlier in the lecture):

Proposition 1 ***(Law of Iterated Expectations)***

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y]$$

PROOF: Let $g(x) = \mathbb{E}[Y|X = x]$, which is a function of x. We can now calculate the expectation

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x] f_X(x) dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y \frac{f_{XY}(x,y)}{f_X(x)} dy \right) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x,y) dy dx \\
&= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{XY}(x,y) dx \right) dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) dy = \mathbb{E}[Y]
\end{aligned}$$