6.2 Correlation

Throughout the past chapters, we often made the assumption that two random variables are independent in various exercises and methods. In reality, most random variables are not actually independent. In this section we give some measures to quantify how “related” two random variables are.

The example in the Linear Regression chapter began with the observation that GPAs are positively correlated with future salaries. That is, we assumed that as college GPA increased, future salary also increased. Qualitatively, this was enough to motivate the problem of regression. However, there were other predictors that contributed to the future salary, some of which were also positively correlated to the projected salary. The fact that some variables contributed “more positively” than others was manifested in the size of the weights that were attached to the variables in the equation $Y = \mathbf{X} \cdot \mathbf{w} + \varepsilon$. If one X_i were more predictive of Y than another, then its corresponding weight was larger. In the following section we examine the covariance of two random variables, which is another attempt to quantify the relationship between random variables.

6.2.1 Covariance

Suppose we have two random variables X and Y, not necessarily independent, and we want to quantify their relationship with a number. This number should satisfy two basic requirements.

(a) The number should be positive when X and Y increase/decrease together.

(b) It should be negative when one of X or Y decreases while the other increases.

Consider the following random variable.

$$(X - EX)(Y - EY).$$

Now consider the possible realizations of these random variables. We denote the possible values using lowercase x and y. The collection of these pairs is the sample space Ω. We can think of the outcomes of sampling an X and a Y as pairs $(x, y) \in \Omega$. Suppose the probability distribution governing X and Y on Ω assigns most of the probability mass on the pairs (x, y) such that $x > EX$ and $y > EY$. In this case, the random variable $(X - EX)(Y - EY)$ is likely to be positive most of the time. Similarly, if more mass were placed on pairs (x, y) such that $x < EX$ and $y < EY$, the product $(X - EX)(Y - EY)$ would be a negative number times a negative number, which means it would still be positive most of the time. Hence the product $(X - EX)(Y - EY)$ being positive is indicative of X and Y being mutually more positive or mutually more negative.

By similar reasoning, the product $(X-EX)(Y-EY)$ is more often negative if the distribution assigns more mass to pairs (x, y) that have $x < EX$ and $y > EY$, or that satisfy $x > EX$ and $y < EY$. In either case, the product $(X-EX)(Y-EY)$ will be a product of a positive and negative number, which is negative.

We are almost done. Remember at the beginning of this discussion we were searching for a number to summarize a relationship between X and Y that satisfied the requirements (a) and (b). But $(X - EX)(Y - EY)$ is a random variable, (that is, a function mapping Ω to $\mathbb{R}$) not a number. To get a number, we take the expectation. Finally we arrive at the definition of covariance.

Definition 6.2. *The* ***covariance*** *of two random variables X and Y, written $Cov(X, Y)$, is defined*

$$Cov(X, Y) = E[(X - EX)(Y - EY)].$$

This definition may look similar to the definition for variance of a random variable X, except we replace one of the terms in the product with the difference $Y - EY$. Similar to Proposition 2.11 (c), there is another useful form of the covariance.

Proposition 6.3. *Let X and Y be two random variables with means EX and EY respectively. Then*

$$Cov(X, Y) = E[XY] - E[X]E[Y].$$

Proof. By the definition of covariance, we can foil the product inside the expectation to get

$$\begin{aligned}
\text{Cov}(X, Y) &= E[XY - XEY - YEX + EXEY] \\
&= E[XY] - E[XEY] - E[YEX] + E[EXEY] && \text{(linearity of } E\text{)} \\
&= E[XY] - EYEX - EXEY + EXEY && \text{(linearity of } E\text{)} \\
&= E[XY] - EXEY.
\end{aligned}$$

□

6.2.2 The Correlation Coefficient

The covariance quantity we just defined satisfies conditions (a) and (b), but can become arbitrarily large depending on the distribution of X and Y. Thus comparing covariances between different pairs of random variables can be tricky. To combat this, we normalize the quantity to be between -1 and 1. The normalized quantity is called the correlation, defined below.

Definition 6.4. *The* ***correlation coefficient*** *between two random variables X and Y with standard deviations σ_x and σ_y, is denoted ρ and is defined*

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}.$$

Exercise 6.5. Verify that for given random variables X and Y, the correlation $\rho_{x,y}$ lies between -1 and 1.

Heuristic. Given a random variable X, the first question we ask is,

What is the random variable *most positively* correlated with X?

The random variable that correlates most positively with X should increase *exactly* with X and decrease *exactly* with X. The only random variable that accomplishes this feat is X itself. This implies that the correlation coefficient between X and any random variable Y is less than that between X and itself. That is,

$$\rho_{x,y} \leq \rho_{x,x} = \frac{\text{Cov}(X,X)}{\sigma_x \sigma_x} = \frac{\text{Var}(X)}{\text{Var}(X)} = 1.$$

By now you've probably guessed the second question we need to ask, that is,

What is the random variable *least positively* correlated with X?

In other words, we are looking for a random variable with which the correlation between X and this random variable is the most negative it can be. This random variable should increase *exactly* as X decreases, and it should also decrease *exactly* as X increases. The candidate that comes to mind is $-X$. This would imply that the correlation coefficient between X and any random variable Y is greater than that between X and $-X$.

This implies that

$$\rho_{x,y} \geq \rho_{x,-x} = \frac{\text{Cov}(X,-X)}{\sigma_x \sigma_{-x}}.$$

By Proposition 6.3, the expression on the right becomes

$$= \frac{E[X(-X)] - E[X]E[-X]}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(-X)}} = \frac{-(E[X^2] - (EX)^2)}{\text{Var}(X)} = \frac{-\text{Var}(X)}{\text{Var}(X)} = -1.$$

Hence, we conclude that $-1 = \rho_{x,-x} \leq \rho_{x,y} \leq \rho_{x,x} = 1$. This completes the proof. □

6.2.3 Interpretation of Correlation

The correlation coefficient between two random variables X and Y can be understood by plotting samples of X and Y in the plane. Suppose we sample from the distribution on X and Y and get

$$\text{Sample} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}.$$

There are three possibilities.

Case 1: $\rho_{x,y} > 0$. We said that this corresponds to X and Y increasing mutually or decreasing mutually. If this is the case, then if we took many samples and plotted the observations, the best fit line would have a positive slope. In the extreme case if $\rho_{xy} = 1$, the samples (X_i, Y_i) would all fall perfectly on a line with slope 1.

Case 2: $\rho_{x,y} = 0$. This typically corresponds to X and Y having no observable relationship. However, this does not necessarily mean that X and Y have no relationship whatsoever. It just means that the measure we are using the quantify their relative spread (the correlation) doesn't capture the underlying relationship. We'll see an example of this later. In terms of the plot, the samples (X_i, Y_i) might look scattered on the $\mathbb{R}^2$ plane with no apparent pattern.

Case 3: $\rho_{x,y} < 0$. We said that this case corresponds to one of X or Y decreasing while the other increases. If this were the case, then the best fit line is likely to have a negative slope. In the extreme case when $\rho_{xy} = -1$, all samples fall perfectly on a line with slope -1.

6.2.4 Independence vs Zero Correlation

There is a commonly misunderstood distinction between the following two statements.

1. "X and Y are independent random variables."
2. "The correlation coefficient between X and Y is 0."

The following statement is always true.

Proposition 6.6. *If X and Y are independent random variables, then $\rho_{x,y} = 0$.*

The converse is not. That is, $\rho_{x,y} = 0$ does not *necessarily* imply that X and Y are independent.

In "Case 2" of the previous section, we hinted that even though $\rho_{x,y} = 0$ corresponded to X and Y having no observable relationship, there could still be some underlying relationship between the random variables, i.e. X and Y are still not independent. First let's prove Proposition 6.6

Proof. Suppose X and Y are independent. Then functions of X and Y are independent. In particular, the functions

$$\begin{aligned} f(X) &\doteq X - EX \\ g(Y) &\doteq Y - EY \end{aligned}$$

are independent. By the definition of correlation,

$$\begin{aligned} \rho_{xy} &= \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} \\ &= \frac{E[(X-EX)(Y-EY)]}{\sigma_x \sigma_y} \\ &= \frac{E[f(X)g(Y)]}{\sigma_x \sigma_y} \\ &= \frac{E[f(X)]E[g(Y)]}{\sigma_x \sigma_y} && \text{(independence of } f(X) \text{ and } g(Y)) \\ &= \frac{0 \cdot 0}{\sigma_x \sigma_y} && (E[f(X)] = E(X - EX) = 0) \\ &= 0 \end{aligned}$$

Hence, if X and Y are independent, then $\rho_{x,y} = 0$. □

Now let's see an example where the converse does not hold. That is, an example of two random variables X and Y such that $\rho_{x,y} = 0$, but X and Y are *not* independent.

Example 6.7. Suppose X is a discrete random variable taking on values in the set $\{-1, 0, 1\}$, each with probability $\frac{1}{3}$. Now consider the random variable $|X|$. These two random variables are clearly not independent, since once we know the value of X, we know the value of $|X|$. However, we can show that X and $|X|$ are uncorrelated. By the definition of correlation and Proposition 6.3,

$$\rho_{x,|x|} = \frac{E(X \cdot |X|) - EX \cdot E|X|}{\sigma_x \sigma_{|x|}} \tag{3}$$

Let's compute the numerator. By looking at the distribution of X, we can see that the product $X \cdot |X|$ can only take on three possible values. If $X = 0$, then $|X| = 0$ so $X \cdot |X| = 0$. If $X = -1$, then $|X| = 1$ and $X \cdot |X| = -1$. Finally if $X = 1$, then $|X| = 1$ and $X \cdot |X| = 1$. Each of these cases occur with probability $\frac{1}{3}$. Hence,

$$X \cdot |X| \sim \text{Uniform}\{-1, 0, 1\}$$

It follows that the expectation of $X \cdot |X|$ is

$$E(X \cdot |X|) = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot (0) + \frac{1}{3} \cdot (1) = 0.$$

Also by the definition of expectation,

$$E[X] = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot (0) + \frac{1}{3} \cdot (1) = 0.$$

Plugging these values into the numerator in expression (3), we get $\rho_{x,|x|} = 0$. Thus, the two random variables X and $|X|$ are not always equal, they are not independent, and yet they have correlation 0. It is important to keep in mind that zero correlation *does not* necessarily imply independence.