Math 40 Probability and Statistical Inference

Winter 2021

Yoonsang Lee (yoonsang.lee@dartmouth.edu)

Lecture 9-2: Correlation and linear regression (3.4)

Correlation and linear regression (3.4)

In doing data science, we are interested in knowing about a random variable using information of another random variable. In doing so, the correlation between them plays an important role. In this lecture, we focus on the correlation and its applications in linear regression.
For two random variables, the covariance is defined as

$$cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

Properties of the covariance

- $cov(X, X) = Var(X)$
- $cov(X, Y) = cov(Y, X)$
- $cov(X + a, Y) = cov(X, Y + b) = cov(X, Y)$
- $cov(X, Y + Z) = cov(X, Y) + cov(X, Z)$
- $cov(aX, Y) = a \times cov(X, Y), cov(X, bY) = b \times cov(X, Y)$

Correlation and linear regression (3.4)

In doing data science, we are interested in knowing about a random variable using information of another random variable. In doing so, the correlation between them plays an important role. In this lecture, we focus on the correlation and its applications in linear regression.

For two random variables, the covariance is defined as

$$cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

Properties of the covariance

- $Var(X + Y) = Var(X) + 2cov(X, Y) + Var(Y)$
- If X and Y do not correlate, then $Var(X + Y) = Var(X) + Var(Y)$
- $cov(X, Y) = E(XY) - \mu_X\mu_Y = E(X(Y - \mu_Y)) = E((X - \mu_X)Y)$
- $cov(X, Y) = 0$ if X and Y are independent.

Correlation and linear regression (3.4)

The correlation coefficient is scaling invariant, which is defined as below

$$\rho = cor(X, Y) = \frac{cov(X, Y)}{std(X)std(Y)}$$

Note Zero correlation does not imply independence. See example 3.43.

Correlation and linear regression (3.4)

Example 3.45 Coefficients of determination and correlation. On page 174, we defined the coefficient of determination

$$\rho^2(x) = \frac{Var(E(Y|X))}{Var(Y)}$$

We also have the correlation coefficient

$$\rho = \frac{cov(X, Y)}{std(X)std(Y)}$$

We consider $X \sim \mathcal{N}(0, 1)$ and $Y = 1 - X^2$. The coefficient of determination is given by

$$\rho^2(x) = 1$$

while the correlation coefficient $\rho = 0$. In this case, X and Y are linearly uncorrelated, while they are nonlinearly dependent.

Correlation and linear regression (3.4)

In section 3.3, we were discussing the conditional mean as the minimizer of $E((Y-c)^2)$.
Let's do a specific calculation. We assume that c takes the form of $a+bX$ where a and b are unknown constants. By taking the derivatives with respect to a and b, and solve for a and b, we find that

$$b = \frac{E((X-\mu_X)(Y-\mu_Y)}{E((X-\mu_X)^2)} = \rho\frac{\sigma_y}{\sigma_x}$$

and

$$a = \mu_Y - b\mu_X$$

$y = a+bX$ is called the least squares linear regression between Y and X.

Correlation and linear regression (3.4)

Difference between the conditional expectation and the least squares linear regression.

If you need to predict Y using X, what would you choose as a predictor, $E(Y|X)$ or $Y = a + bX$?

Correlation and linear regression (3.4)

Difference between the conditional expectation and the least squares linear regression.
If you need to predict Y using X, what would you choose as a predictor, $E(Y|X)$ or $Y = a + bX$?
$Y = a + bX$ is a special case of the conditional expectation (by assuming a linear relation between X and Y).

Correlation and linear regression (3.4)

Example 3.48 Variance decomposition and linear coefficient of determination.
On p.174, Theorem 3.29 shows that

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

If $E(Y|X) = a + bX$, then

$$Var(Y) = E[(Y - E(Y|X))^2] + E[(\mu_Y - a - bX)^2]$$

$$= E[(y - a - bX)^2] + E[(b\mu_X - bX)^2]$$

$$= \sigma^2 + \rho^2\sigma_Y^2$$

where $\sigma^2 = E[(y - a - bX)^2]$.

Correlation and linear regression (3.4)

Several remarks on **Example 3.48**

- Exercise 3.4.10 can be solved using Example 3.48.
- If there is a linear relation between X and Y, the coefficient of determination is the square of the correlation coefficient.