1 Conditional expectation given an event

Let Y be a random variable, and A an event. We assume we simulate n values $\{y_1, y_2, \ldots, y_n\}$ independently from the distribution of Y. The expectation $\mathsf{E}(Y)$ can be numerically estimated as the average of the simulated values (this is called a Monte Carlo estimate):

$$\mathsf{E}(Y) = \frac{y_1 + y_2 + \ldots + y_n}{n}.$$

The conditional expectation $\mathsf{E}(Y \mid A)$ is numerically approximated by considering only those values among $\{y_1, y_2, \ldots, y_n\}$ where A occurred, and taking their average:

$$\mathsf{E}(Y \mid A) = \frac{\sum_{\{j:\text{A occurred for } y_j\}} y_j}{\#\{j : \text{A occurred for } y_j\}}.$$

Definition (Conditional expectation given an event)**.** Let A be an event with positive probability, $\mathsf{P}(A) > 0$. If Y is a discrete random variable, then the *conditional expectation of Y given A* is

$$\mathsf{E}(Y \mid A) = \sum_y y \mathsf{P}(Y = y \mid A),$$

where the sum is over the support of Y. If Y is a continuous random variable with PDF f, then

$$\mathsf{E}(Y \mid A) = \int_{-\infty}^{\infty} y f(y \mid A)\, dy,$$

where the conditional PDF $f(y \mid A)$ is defined as the derivative of the conditional CDF

$$F(y \mid A) = \mathsf{P}(Y \leq y \mid A),$$

and can also be computed by a hybrid version of Bayes' rule:

$$f(y \mid A) = \frac{\mathsf{P}(A \mid Y = y) f(y)}{\mathsf{P}(A)}.$$

Theorem (Law of total expectation)**.** *Let $A_1, A_2, \ldots, A_n$ be a partition of a sample space, with $\mathsf{P}(A_i) > 0$ for all $i = 1, 2, \ldots, n$. Let Y be a random variable on this sample space. Then*

$$\mathsf{E}(Y) = \sum_{i=1}^{n} \mathsf{E}(Y \mid A_i)\mathsf{P}(A_i).$$

The law of total probability (LOTP) is a particular case of the law of total expectation. Let B be an event, and let $Y = \mathsf{I}_B$ be its indicator. The law of total expectation says:

$$\mathsf{E}(\mathsf{I}_B) = \sum_{i=1}^{n} \mathsf{E}(I_B \mid A_i)\mathsf{P}(A_i).$$

But, by the fundamental bridge, we have

$$\mathsf{E}(\mathsf{I}_B) = \mathsf{P}(B), \quad \mathsf{E}(I_B \mid A_i) = \mathsf{P}(B \mid A_i).$$

Thus we obtain LOTP:

$$\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i).$$

Example: Geometric expectation

Let $X \sim \mathsf{Geom}(p)$. Then X represents the number of failures before the first successful trial in a sequence of independent Bernoulli trials, each with the same probability of success $p \in (0, 1)$. We denote $q = 1 - p$. We calculate $\mathsf{E}(X)$ by conditioning on the outcome of the first trial: if this outcome is a success, $X = 0$. Otherwise, if this outcome is a failure, we are back where we started by memorylessness (this property comes from the independence of the Bernoulli trials). Thus

$$\mathsf{E}(X) = \mathsf{E}(X \mid \text{outcome of first trial is success}) \cdot p + \mathsf{E}(X \mid \text{outcome of first trial is failure}) \cdot q = 0 \cdot p + (1 + \mathsf{E}(X))q.$$

We solve the equation $\mathsf{E}(X) = (1 + \mathsf{E}(X))q$, and obtain $\mathsf{E}(X) = q/p$.

2 Conditional expectation given a random variable

The key to understanding the conditional expectation of a random variable Y given another random variable X, denoted by $\mathsf{E}(Y \mid X)$, is first to understand the conditional expectation $\mathsf{E}(Y \mid X = x)$ of Y given the event $X = x$. If Y is discrete, we have:

$$\mathsf{E}(Y \mid X = x) = \sum_y y \mathsf{P}(Y = y \mid X = x).$$

If Y is continuous, we have:

$$\mathsf{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy.$$

Definition (Conditional expectation given a random variable)**.** Let $g(x) = \mathsf{E}(Y \mid X = x)$. The *conditional expectation of Y given X*, denoted by $\mathsf{E}(Y \mid X)$, is defined to be the random variable $g(X)$. This random variable can have an expectation $\mathsf{E}(\mathsf{E}(Y \mid X))$, a variance $\mathsf{Var}(\mathsf{E}(Y \mid X))$, and also higher order moments.

Example:

Suppose we have a stick of length 1, and break the stick at a random point X chosen uniformly at random. Given that $X = x$, we then choose another breakpoint Y uniformly on the interval $[0, x]$. Find the conditional expectation $\mathsf{E}(Y \mid X)$, its mean $\mathsf{E}(\mathsf{E}(Y \mid X))$, and its variance $\mathsf{Var}(\mathsf{E}(Y \mid X))$.

Solution: We have $X \sim \mathsf{Unif}(0, 1)$ and $Y \mid X = x \sim \mathsf{Unif}(0, x)$. Then

$$g(x) = \mathsf{E}(Y \mid X = x) = \frac{x}{2}.$$

Thus the conditional mean of Y given X is

$$\mathsf{E}(Y \mid X) = g(X) = \frac{X}{2}.$$

Since $X \sim \mathsf{Unif}(0, 1)$, we have $\frac{X}{2} \sim \mathsf{Unif}\left(0, \frac{1}{2}\right)$. Thus the mean and the variance of the conditional expectation $\mathsf{E}(Y \mid X)$ are:

$$\mathsf{E}(\mathsf{E}(Y \mid X)) = \mathsf{E}\left(\frac{X}{2}\right) = \frac{1}{4}, \quad \mathsf{Var}(\mathsf{E}(Y \mid X)) = \mathsf{Var}\left(\frac{X}{2}\right) = \frac{1}{48}.$$

3 Properties of conditional expectation

Theorem (Dropping what is independent). *If X and Y are independent, then $\mathsf{E}(Y \mid X) = \mathsf{E}(Y)$. That is, the random variable $\mathsf{E}(Y \mid X)$ is the constant $\mathsf{E}(Y)$.*

Proof. Independence implies $\mathsf{E}(Y \mid X = x) = \mathsf{E}(Y)$ for all x. □

Theorem (Taking out what is known). *For any function $h(\cdot)$, we have*

$$\mathsf{E}(h(X)Y \mid X) = h(X)\mathsf{E}(Y \mid X).$$

Note that the above equality means that the random variable $g_1(X) = \mathsf{E}(h(X)Y \mid X)$ is equal with the random variable $g_2(X) = h(X)\mathsf{E}(Y \mid X)$.

Proof. We know that, for any constant $c \in \mathbb{R}$, we have

$$\mathsf{E}(cX) = c\mathsf{E}(X).$$

Once we know the value x of X, the function $h(X)$ becomes a constant $h(x)$. □

Theorem (Linearity). *For any random variables $Y_1, Y_2, \ldots, Y_n$ and X, we have*

$$\mathsf{E}(Y_1 + Y_2 + \ldots + Y_n \mid X) = \mathsf{E}(Y_1 \mid X) + \mathsf{E}(Y_2 \mid X) + \ldots + \mathsf{E}(Y_n \mid X).$$

Example:
Let $Y_1, \ldots, Y_n$ be i.i.d., and $S_n = Y_1 + \ldots + Y_n$. Find $\mathsf{E}(X_1 \mid S_n)$.

Solution: By symmetry,

$$\mathsf{E}(Y_1 \mid S_n) = \mathsf{E}(Y_2 \mid S_n) = \ldots = \mathsf{E}(Y_n \mid S_n).$$

By linearity and by taking out what is known,

$$\mathsf{E}(Y_1 \mid S_n) + \mathsf{E}(Y_2 \mid S_n) + \ldots + \mathsf{E}(Y_n \mid S_n) = \mathsf{E}(S_n \mid S_n) = S_n.$$

Thus

$$\mathsf{E}(Y_1 \mid S_n) = \frac{S_n}{n}.$$

Theorem (Adam's law: connecting conditional expectation with unconditional expectation). *For any random variables X and Y, we have*

$$\mathsf{E}(\mathsf{E}(Y \mid X)) = \mathsf{E}(Y).$$

Theorem (Adam's law with extra conditioning). *For any random variables X, Y and Z, we have*

$$\mathsf{E}(\mathsf{E}(Y \mid X, Z) \mid Z) = \mathsf{E}(Y \mid Z).$$

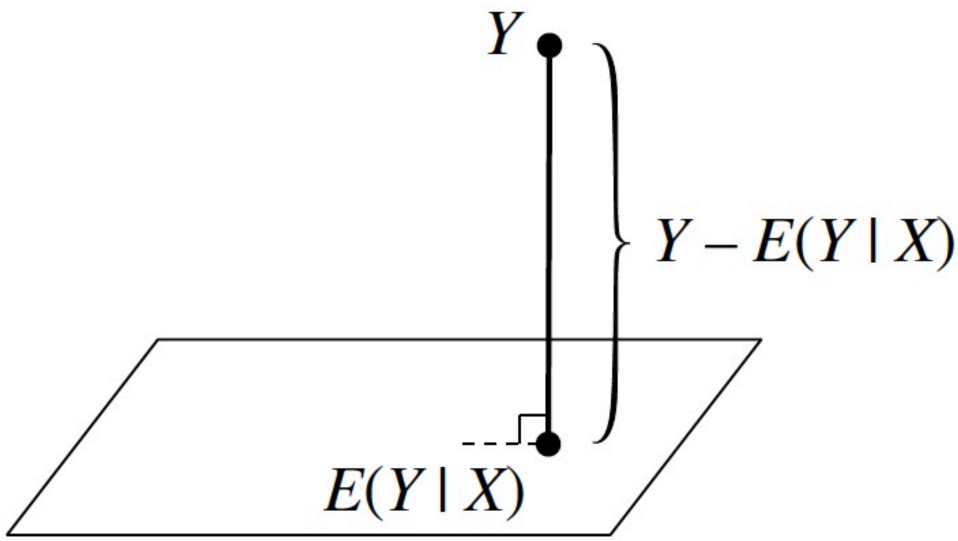

Figure 1: In this figure, we consider the vector space of all the random variables with zero mean and finite variance on a certain probability space. Each random variable corresponds with a point (vector) in this space. The subspace of random variables of the form $h(X)$ is represented as a plane. The conditional expectation $\mathsf{E}(Y \mid X)$ belongs to this plane. To obtain $\mathsf{E}(Y \mid X)$, we project Y onto this plane and obtain $\mathsf{E}(Y \mid X)$ as the function of X that is closest to Y. The line from Y to $\mathsf{E}(Y \mid X)$ is orthogonal (perpendicular) to the plane since any other route from Y to $\mathsf{E}(Y \mid X)$ would be longer. Then the residual $Y - \mathsf{E}(Y \mid X)$ is orthogonal to $h(X)$ for all functions $h(\cdot)$, and $\mathsf{E}(Y \mid X)$ is the function of X and best predicts Y, where "best" means that the mean squared error $\mathsf{E}\left((Y - h(X))^2\right)$ is minimized by choosing $h(X) = \mathsf{E}(Y \mid X)$.

Theorem (Projection interpretation)**.** *For any function $h(\cdot)$, the random variable $Y - \mathsf{E}(Y \mid X)$ is uncorrelated with $h(X)$. Since*

$$\mathsf{E}(Y - \mathsf{E}(Y \mid X)) = \mathsf{E}(Y) - \mathsf{E}(\mathsf{E}(Y \mid X)) = \mathsf{E}(Y) - \mathsf{E}(Y) = 0,$$

this is equivalent with

$$\mathsf{E}((Y - \mathsf{E}(Y \mid X))h(X)) = 0.$$

The geometric interpretation of this result is given in Figure 1.

Proof.

$$\mathsf{E}((Y-\mathsf{E}(Y \mid X))h(X)) = \mathsf{E}(Yh(X))-\mathsf{E}(\mathsf{E}(Y \mid X)h(X)) = h(X)\mathsf{E}(Y)-h(X)\mathsf{E}(\mathsf{E}(Y \mid X)) = h(X)\mathsf{E}(Y)-h(X)\mathsf{E}(Y) = 0.$$

□

Example: Linear regression

In its simplest form, the linear regression model uses a single explanatory variable X to predict a response variable Y. It assumes that the conditional expectation of Y given X is linear in X:

$$\mathsf{E}(Y \mid X) = a + bX.$$

(a) Show that an equivalent way to express this is to write:

$$Y = a + bX + \epsilon,$$

where ϵ is a random variable (called the error) with $\mathsf{E}(\epsilon \mid X) = 0$.

(b) Express the constants a and b in terms of $\mathsf{E}(X)$, $\mathsf{E}(Y)$, $\mathsf{Cov}(X, Y)$, and $\mathsf{Var}(X)$.

Solution: (a) Let $Y = a + bX + \epsilon$ with $\mathsf{E}(\epsilon \mid X) = 0$. Then

$$\mathsf{E}(Y \mid X) = \mathsf{E}(a \mid X) + \mathsf{E}(bX \mid X) + \mathsf{E}(\epsilon \mid X) = a + b\mathsf{E}(X \mid X) = a + bX.$$

Conversely, assume that $\mathsf{E}(Y \mid X) = a + bX$, and define $\epsilon = Y - (a + bX)$. Then

$$\mathsf{E}(\epsilon \mid X) = \mathsf{E}(Y \mid X) - \mathsf{E}(a + bX \mid X) = \mathsf{E}(Y \mid X) - (a + bX) = 0.$$

(b) By Adam's law, the unconditional mean of Y is

$$\mathsf{E}(Y) = \mathsf{E}(\mathsf{E}(Y \mid X)) = \mathsf{E}(a + bX)) = a + b\mathsf{E}(X).$$

The unconditional mean of the error ϵ is

$$\mathsf{E}(\epsilon) = \mathsf{E}(\mathsf{E}(\epsilon \mid X)) = \mathsf{E}(0) = 0.$$

X and ϵ are uncorrelated because:

$$\mathsf{E}(\epsilon X) = \mathsf{E}(\mathsf{E}(\epsilon X \mid X)) = \mathsf{E}(X\mathsf{E}(\epsilon \mid X)) = \mathsf{E}(X \cdot 0) = 0.$$

Thus

$$\mathsf{Cov}(X, Y) = \mathsf{Cov}(X, a + bX + \epsilon) = \mathsf{Cov}(X, a) + b\mathsf{Cov}(X, X) + \mathsf{Cov}(X, \epsilon) = b\mathsf{Var}(X).$$

Thus $b = \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(X)}$, and

$$a = \mathsf{E}(Y) - b\mathsf{E}(X) = \mathsf{E}(Y) - \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(X)}\mathsf{E}(X)$$

4 Conditional variance

Definition (Conditional variance)**.** The *conditional variance of Y given X* is

$$\mathsf{Var}(Y \mid X) = \mathsf{E}\left([Y - \mathsf{E}(Y \mid X)]^2 \mid X\right).$$

This is equivalent to

$$\mathsf{Var}(Y \mid X) = \mathsf{E}(Y^2 \mid X) - (\mathsf{E}(Y \mid X))^2.$$

The conditional variance $\mathsf{Var}(Y \mid X)$ is a random variable, and it is a function of X.

Example:
Let $Z \sim \mathsf{N}(0,1)$ and $Y = Z^2$. Find $\mathsf{E}(Y \mid Z)$, $\mathsf{Var}(Y \mid Z)$, $\mathsf{E}(Z \mid Y)$, and $\mathsf{Var}(Z \mid Y)$.

Solution: We apply the taking out what is known theorem with $h(z) = z^2$, and obtain:

$$\mathsf{E}(Y \mid Z) = \mathsf{E}(Z^2 \mid Z) = \mathsf{E}(h(Z) \cdot 1 \mid Z) = h(Z) \cdot \mathsf{E}(1 \mid Z) = h(Z) = Z^2.$$

We can see this directly: conditional on $Z = z$, $Y = Z^2 = z^2$ a constant. Since the expectation of a constant is the constant itself, and the variance of a constant is 0, we have

$$\mathsf{E}(Y \mid Z) = Z^2, \quad \mathsf{Var}(Y \mid Z) = 0.$$

On the other hand, if $Y = Z^2 = y$, then $Z \in \{-\sqrt{t}, \sqrt{t}\}$. From the Bayes' rule we have:

$$\mathsf{P}(Z = -\sqrt{t} \mid Y = t) \propto \phi(-\sqrt{t}), \quad \mathsf{P}(Z = \sqrt{t} \mid Y = t) \propto \phi(\sqrt{t}),$$

where $\phi(\cdot)$ is the standard Normal PDF. By the symmetry of the standard Normal, we have $\phi(-\sqrt{t}) = \phi(\sqrt{t})$. Hence

$$\mathsf{P}(Z = -\sqrt{t} \mid Y = t) = \mathsf{P}(Z = \sqrt{t} \mid Y = t) = \frac{1}{2}.$$

Thus

$$\mathsf{E}(Z \mid Y = t) = -\sqrt{t}\mathsf{P}(Z = -\sqrt{t} \mid Y = t) + \sqrt{t}\mathsf{P}(Z = \sqrt{t} \mid Y = t) = 0,$$

which implies $\mathsf{E}(Z \mid Y) = 0$. Moreover

$$\mathsf{E}(Z^2 \mid Y = t) = t\mathsf{P}(Z = -\sqrt{t} \mid Y = t) + t\mathsf{P}(Z = \sqrt{t} \mid Y = t) = t,$$

which implies $\mathsf{E}(Z^2 \mid Y) = Y$. By definition:

$$\mathsf{Var}(Z \mid Y) = \mathsf{E}(Z^2 \mid Y) - (\mathsf{E}(Z \mid Y))^2 = Y.$$

Theorem (Eve's law: connecting conditional variance to unconditional variance)**.** *For any random variables X and Y, we have*

$$\mathsf{Var}(Y) = \mathsf{E}\left(\mathsf{Var}(Y \mid X)\right) + \mathsf{Var}\left(\mathsf{E}(Y \mid X)\right).$$

This relation is known as the law of total variance, or as the variance decomposition formula.

Proof. Let $g(X) = \mathsf{E}(Y \mid X)$. By Adam's law, $\mathsf{E}(g(x)) = \mathsf{E}(Y)$. Then

$$\mathsf{E}\left(\mathsf{Var}(Y \mid X)\right) = \mathsf{E}\left(\mathsf{E}(Y^2 \mid X) - g(X)^2\right) = \mathsf{E}\left(Y^2\right) - \mathsf{E}\left(g(X)^2\right),$$

and

$$\mathsf{Var}\left(\mathsf{E}(Y \mid X)\right) = \mathsf{E}\left(g(X)^2\right) - (\mathsf{E}g(X))^2 = \mathsf{E}\left(g(X)^2\right) - (\mathsf{E}Y)^2.$$

Thus

$$\mathsf{E}\left(\mathsf{Var}(Y \mid X)\right) + \mathsf{Var}\left(\mathsf{E}(Y \mid X)\right) = \mathsf{E}\left(Y^2\right) - (\mathsf{E}Y)^2 = \mathsf{Var}(Y).$$

$\square$

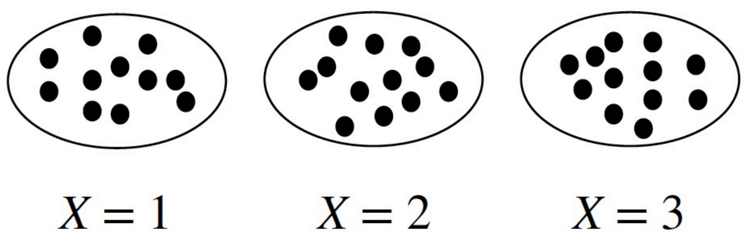

Figure 2: Illustration of Eve's law. Consider a population where each person has a value of X (e.g., age) and a value of Y (height). Assume that we divide this population into subpopulations or groups: here, we have three age groups associated with $X = 1$, $X = 2$ and $X = 3$. There are two sources contributing to the variation in people's heights in the overall population: (1) *within-group variation* $\mathsf{E}\left(\mathsf{Var}(Y \mid X)\right)$ represents the average amount of variation in Y (height) within (conditional on) each age group; and (2) *between-group variation* $\mathsf{Var}\left(\mathsf{E}(Y \mid X)\right)$ (here, the variance of the group means $\mathsf{E}(Y \mid X = 1)$, $\mathsf{E}(Y \mid X = 2)$ and $\mathsf{E}(Y \mid X = 3)$) represents the variance of average heights across age groups. Eve's law says that total variance is the sum of within-group and between-group variation.