6.7 Adam's law and other properties of conditional expectation

Unit 6: Joint Distributions and Conditional Expectation

Adapted from Blitzstein-Hwang Chapters 7 and 9.

Conditional expectation has some very useful properties, that often allow us to solve problems without having to go all the way back to the definition.

- **Dropping what's independent:** If X and Y are independent, then $E(Y|X) = E(Y)$.
- **Taking out what's known:** For any function h, $E(h(X)Y|X) = h(X)E(Y|X)$.
- **Linearity:** $E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X)$, and $E(cY|X) = cE(Y|X)$ for c a constant (the latter is a special case of taking out what's known).
- **Adam's law:** $E(E(Y|X)) = E(Y)$.
- **Projection interpretation:** The r.v. $Y - E(Y|X)$, which is called the *residual* from using X to predict Y, is uncorrelated with $h(X)$ for any function h.

Let's discuss each property individually.

> THEOREM 6.7.1 (DROPPING WHAT'S INDEPENDENT).
>
> If X and Y are independent, then $E(Y|X) = E(Y)$.

This is true because independence implies $E(Y|X = x) = E(Y)$ for all x, hence $E(Y|X) = E(Y)$. Intuitively, if X provides no information about Y, then our best guess for Y, even if we get to know X, is still the unconditional mean $E(Y)$. However, the converse is false: a counterexample is given in Example 6.7.4 below.

> THEOREM 6.7.2 (TAKING OUT WHAT'S KNOWN).
>
> For any function h,

$$E(h(X)Y|X) = h(X)E(Y|X).$$

Intuitively, when we take expectations given X, we are treating X as if it has crystallized into a known constant. Then any function of X, say $h(X)$, also acts like a known constant while we are conditioning on X.

Theorem 6.7.3 (Linearity).

$$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X).$$

This result is the conditional version of the unconditional fact that $E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$, and is true since conditional probabilities *are* probabilities.

Example 6.7.4.

Let $X_1, \ldots, X_n$ be i.i.d., and $S_n = X_1 + \cdots + X_n$. Find $E(X_1|S_n)$.

Solution

By symmetry,

$$E(X_1|S_n) = E(X_2|S_n) = \cdots = E(X_n|S_n),$$

and by linearity,

$$E(X_1|S_n) + \cdots + E(X_n|S_n) = E(S_n|S_n) = S_n.$$

Therefore,

$$E(X_1|S_n) = S_n/n = \bar{X}_n,$$

the sample mean of the X_j's. This is an intuitive result: if we have 2 i.i.d. r.v.s X_1, X_2 and learn that $X_1 + X_2 = 10$, it makes sense to guess that X_1 is 5 (accounting for half of the total). Similarly, if we have n i.i.d. r.v.s and get to know their sum, our best guess for any one of them is the sample mean.

The next theorem connects conditional expectation to unconditional expectation. It goes by many names, including the law of total expectation, iterated expectations, and the tower property. We call it *Adam's law* because it is used so frequently that it deserves a pithy name, and since it is often used in conjunction with another law we'll encounter soon, which has a complementary name.

Theorem 6.7.5 (Adam's law).

For any r.v.s X and Y,

$$E(E(Y|X)) = E(Y).$$

Proof

We present the proof in the case where X and Y are both discrete (the proofs for other cases are analogous). Let $E(Y|X) = g(X)$. We proceed by applying LOTUS, expanding the definition of $g(x)$ to get a double sum, and then swapping the order of summation:

$$\begin{aligned} E(g(X)) &= \sum_x g(x)P(X = x) \\ &= \sum_x \left(\sum_y yP(Y = y|X = x) \right) P(X = x) \\ &= \sum_x \sum_y yP(X = x)P(Y = y|X = x) \\ &= \sum_y y \sum_x P(X = x, Y = y) \\ &= \sum_y yP(Y = y) = E(Y). \end{aligned}$$

Adam's law is a more compact, more general version of the law of total expectation (Theorem 6.5.4). For X discrete, the statements

$$E(Y) = \sum_x E(Y|X = x)P(X = x)$$

and

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

mean the same thing, since if we let $E(Y|X = x) = g(x)$, then

$$E(E(Y|X)) = E(g(X)) = \sum_x g(x)P(X = x) = \sum_x E(Y|X = x)P(X = x).$$

But the Adam's law expression is shorter and applies just as well in the continuous case.

Armed with Adam's law, we have a powerful strategy for finding an expectation $E(Y)$, by conditioning on an r.v. X that we wish we knew. First obtain $E(Y|X)$ by treating X as known, and then take the expectation of $E(Y|X)$. As mentioned earlier, we can think of $E(Y|X)$ as a prediction for Y based on X. This is an extremely common problem in statistics: predict or estimate the future observations or unknown parameters based on data. The projection interpretation of conditional expectation implies that $E(Y|X)$ is the *best predictor* of Y based on X, in the sense that it is the function of X with the lowest *mean squared error* (expected squared difference between Y and the prediction of Y).

Example 6.7.6 (Linear regression).

An extremely widely used method for data analysis in statistics is *linear regression*. In its most basic form, the linear regression model uses a single explanatory variable X to predict a response variable Y, and it assumes that the conditional expectation of Y is *linear* in X:

$$E(Y|X) = a + bX.$$

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

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

where ϵ is an r.v. (called the *error*) with $E(\epsilon|X) = 0$.

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

Solution

(a) Let $Y = a + bX + \epsilon$, with $E(\epsilon|X) = 0$. Then by linearity,

$$E(Y|X) = E(a|X) + E(bX|X) + E(\epsilon|X) = a + bX.$$

Conversely, suppose that $E(Y|X) = a + bX$, and define

$$\epsilon = Y - (a + bX).$$

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

$$E(\epsilon|X) = E(Y|X) - E(a + bX|X) = E(Y|X) - (a + bX) = 0.$$

(b) First, by Adam's law, taking the expectation of both sides gives

$$E(Y) = a + bE(X).$$

Note that ϵ has mean 0 and X and ϵ are uncorrelated, since

$$E(\epsilon) = E(E(\epsilon|X)) = E(0) = 0$$

and

$$E(\epsilon X) = E(E(\epsilon X|X)) = E(XE(\epsilon|X)) = E(0) = 0.$$

Taking the covariance with X of both sides in $Y = a + bX + \epsilon$, we have

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

Thus,

$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)},$$

$$a = E(Y) - bE(X) = E(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \cdot E(X).$$