A Conditional expectation

A.1 Review of conditional densities, expectations

We start with the continuous case. This is sections 6.6 and 6.8 in the book. Let X, Y be continuous random variables. We defined the conditional density of X given Y to be

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Then

$$P(a \le X \le b|Y = y) = \int_a^b f_{X,Y}(x|y)\,dx$$

Conditioning on $Y = y$ is conditioning on an event with probability zero. This is not defined, so we make sense of the left side above by a limiting procedure:

$$P(a \le X \le b|Y = y) = \lim_{\epsilon \to 0^+} P(a \le X \le b||Y - y| < \epsilon)$$

We then define the conditional expectation of X given $Y = y$ to be

$$E[X|Y = y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\,dx$$

We have the following continuous analog of the partition theorem.

$$E[Y] = \int_{-\infty}^{\infty} E[Y|X = x]\, f_X(x)\,dx$$

Now we review the discrete case. This was section 2.5 in the book. In some sense it is simpler than the continuous case. Everything comes down to the very first definition involving conditioning. For events A and B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

assuming that $P(B) > 0$. If X is a discrete RV, the conditional density of X given the event B is

$$f(x|B) = P(X = x|B) = \frac{P(X = x, B)}{P(B)}$$

and the conditional expectation of X given B is

$$E[X|B] = \sum_x x\, f(x|B)$$

The partition theorem says that if B_n is a partition of the sample space then

$$E[X] = \sum_n E[X|B_n]\, P(B_n)$$

Now suppose that X and Y are discrete RV's. If y is in the range of Y then $Y = y$ is a event with nonzero probability, so we can use it as the B in the above. So $f(x|Y = y)$ is defined. We can change the notation to make it look like the continuous case and write $f(x|Y = y)$ as $f_{X|Y}(x|y)$. Of course it is given by

$$f_{X|Y}(x|y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

This looks identical to the formula in the continuous case, but it is really a different formula. In the above $f_{X,Y}$ and f_Y are pmf's; in the continuous case they are pdf's. With this notation we have

$$E[X|Y = y] = \sum_x x\, f_{X|Y}(x|y)$$

and the partition theorem is

$$E[X] = \sum_y E[X|Y = y]\, P(Y = y)$$

A.2 Conditional expectation as a Random Variable

Conditional expectations such as $E[X|Y = 2]$ or $E[X|Y = 5]$ are numbers. If we consider $E[X|Y = y]$, it is a number that depends on y. So it is a function of y. In this section we will study a new object $E[X|Y]$ that is a random variable. We start with an example.

Example: Roll a die until we get a 6. Let Y be the total number of rolls and X the number of 1's we get. We compute $E[X|Y = y]$. The event $Y = y$ means that there were $y - 1$ rolls that were not a 6 and then the yth roll was a six. So given this event, X has a binomial distribution with $n = y - 1$ trials and probability of success $p = 1/5$. So

$$E[X|Y = y] = np = \frac{1}{5}(y - 1)$$

Now consider the following process. We do the experiment and get an outcome ω. (In this example, ω would be a string of $1, 2, 3, 4, 5$'s ending with a 6.) Then we compute $y = Y(W)$. (In this example y would just be the number of rolls.) Then we compute $E[X|Y = y]$. This process gives a function

$$\omega \to E[X|Y = y]$$

So this is a random variable. It is usually written as $E[X|Y]$. In our example ω is mapped to $(y-1)/5$ where $y = Y(\omega)$. So ω is mapped to $(Y(\omega)-1)/5$. So the random variable $E[X|Y]$ is just $(Y-1)/5$. Note that $E[X|Y]$ is a function of Y. This will be true in general.

We try another conditional expectation in the same example: $E[X^2|Y]$. Again, given $Y = y$, X has a binomial distribution with $n = y-1$ trials and $p = 1/5$. The variance of such a random variable is $np(1-p) = (y-1)4/25$. So

$$E[X^2|Y=y] - (E[X|Y=y])^2 = (y-1)\frac{4}{25}$$

Using what we found before,

$$E[X^2|Y=y] - (\frac{1}{5}(y-1))^2 = (y-1)\frac{4}{25}$$

And so

$$E[X^2|Y=y] = \frac{1}{25}(y-1)^2 + \frac{4}{25}(y-1)$$

Thus

$$E[X^2|Y] = \frac{1}{25}(Y-1)^2 + \frac{4}{25}(Y-1) = \frac{1}{25}(Y^2+2Y-3)$$

Once again, $E[X^2|Y]$ is a function of Y.

Intuition: $E[X|Y]$ is the function of Y that bests approximates X. This is a vague statement since we have not said what "best" means. We consider two extreme cases. First suppose that X is itself a function of Y, e.g., Y^2 or e^Y. Then the function of Y that best approximates X is X itself. (Whatever best means, you can't do any better than this.) The other extreme case is when X and Y are independent. In this case, knowing Y tells us nothing about X. So we might expect that $E[X|Y]$ will not depend on Y. Indeed, we have

A.3 Properties of conditional expectation

Before we list all the properties of $E[X|Y]$, we need to consider conditioning on more that one random variable. Let X, Y, Z be discrete random variables. Then $E[X|Y=y, Z=z]$ makes sense. We can think of it as a function of the random outcome ω:

$$\omega \to E[X|Y = Y(\omega), Z = Z(\omega)]$$

So it is a random variable. We denote it by $E[X|Y,Z]$. In the continuous case we need to define $E[X|Y=y, Z=z]$ by a limiting process. The result is a function of y and z that we can once again interpret as a random variable.

Theorem 1 *Let X, Y, Z be random variables, $a, b \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Assuming all the following expectations exist, we have*
(i) $E[a|Y] = a$
(ii) $E[aX + bZ|Y] = aE[X|Y] + bE[Z|Y]$
(iii) $E[X|Y] \geq 0$ *if* $X \geq 0$.
(iv) $E[X|Y] = E[X]$ *if* X *and* Y *are independent.*
(v) $E[E[X|Y]] = E[X]$
(vi) $E[Xg(Y)|Y] = g(Y)E[X|Y]$. *In particular,* $E[g(Y)|Y] = g(Y)$.
(vii) $E[X|Y, g(Y)] = E[X|Y]$
(viii) $E[E[X|Y, Z]|Y] = E[X|Y]$

Partial proofs: The first three are not hard to prove, and we leave them to the reader.

Consider (iv). We prove the continuous case and leave the discrete case to the reader. If X and Y are independent then

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

So

$$E[X|Y = y] = \int x\, f_{X|Y}(x|y)\, dx = \int x\, f_X(x)\, dx = E[X]$$

Consider (v). Suppose that the random variables are discrete. We need to compute the expected value of the random variable $E[X|Y]$. It is a function of Y and it takes on the value $E[X|Y = y]$ when $Y = y$. So by the law of the unconscious whatever,

$$E[E[X|Y]] = \sum_y E[X|Y = y]\, P(Y = y)$$

By the partition theorem this is equal to $E[X]$. So in the discrete case, (iv) is really the partition theorem in disguise. In the continuous case it is too.

Consider (vi). We must compute $E[Xg(Y)|Y = y]$. Given that $Y = y$, the possible values of $Xg(Y)$ are $xg(y)$ where x varies over the range of X. The probability of the value $xg(y)$ given that $Y = y$ is just $P(X = x|Y = y)$. So

$$\begin{aligned} E[Xg(Y)|Y = y] &= \sum_x xg(y)P(X = x|Y = y) \\ &= g(y) \sum_x x\, P(X = x|Y = y) = g(y)E[X|Y = y] \end{aligned}$$

This proves (vi).

Consider (viii). Again, we only consider the discrete case. We need to compute $E[E[X|Y;Z]|Y=y]$. $E[X|Y;Z]$ is a random variable. Given that $Y=y$, its possible values are $E[X|Y=y;Z=z]$ where z varies over the range of Z. Given that $Y=y$, the probability that $E[X|Y;Z]=E[X|Y=y;Z=z]$ is just $P(Z=z|Y=y)$. Hence,

$$\begin{aligned}
E[E[X|Y;Z]|Y=y] &= \sum_z E[X|Y=y,Z=z]P(Z=z|Y=y) \\
&= \sum_z \sum_x x\, P(X=x|Y=y,Z=z)P(Z=z|Y=y) \\
&= \sum_{z,x} x \frac{P(X=x,Y=y,Z=z)}{P(Y=y,Z=z)} \frac{P(Z=z,Y=y)}{P(Y=y)} \\
&= \sum_{z,x} x \frac{P(X=x,Y=y,Z=z)}{P(Y=y)} \\
&= \sum_x x \frac{P(X=x,Y=y)}{P(Y=y)} \\
&= \sum_x x\, P(X=x|Y=y) \\
&= E[X|Y=y]
\end{aligned} \tag{1}$$

Example: Let X and Y be independent; each is uniformly distributed on $[0,1]$. Let $Z=X+Y$. Find $E[Z|X], E[X|Z], E[XZ|X], E[XZ|Z]$.

We start with the easy ones.

$$E[Z|X]=E[X+Y|X]=E[X|X]+E[Y|X]=X+E[Y]=X+\frac{1}{2}$$

where we have used the independence of X and Y and properties (iv) and the special case of (vi). Using property (vi), $E[XZ|X]=XE[Z|X]=X(X+\frac{1}{2})$.

Now we do the hard one: $E[X|Z]$. We need the joint pdf of X and Z. So we do a change of variables. Let $W=X$, $Z=X+Y$. This is a linear transformation, so the Jacobian will be a constant. We postpone computing it. We need to find the image of the square $0\le x,y\le 1$ under this transformation. Look at the boundaries. Since it is a linear transformation, the four edges of the square will be mapped to line segments. To find them we can just compute where the four corners of the square are mapped.

$$\begin{aligned}
(x,y)=(0,0) &\to (w,z)=(0,0) \\
(x,y)=(1,0) &\to (w,z)=(1,1) \\
(x,y)=(0,1) &\to (w,z)=(0,1) \\
(x,y)=(1,1) &\to (w,z)=(1,2)
\end{aligned}$$

So the image of the square is the parallelogram with vertices $(0,0),(1,1),(0,1)$ and $(1,2)$. The joint density of W and Z will be uniform on this region. Let A denote the interior of the parallelogram. Since it has area 1, we conclude

$$f_{W,Z}(w,z) = 1((w,z) \in A)$$

Note that we avoided computing the Jacobian. Now we can figure out what $f_{X|Z}(x|z)$ is. We must consider two cases. First suppose $0 \le z \le 1$. Given $Z = z$, X is uniformly distributed between 0 and z. So $E[X|Z = z] = z/2$. In the other case $1 \le z \le 2$. Then X is uniformly distributed between $z-1$ and 1. So $E[X|Z = z] = (z-1+1)/2 = z/2$. So in both cases $E[X|Z = z] = z/2$. Thus $E[X|Z] = Z/2$. Finally, we have $E[XZ|Z] = ZE[X|Z] = Z^2/2$.

We get a small check on our answers using property (v). We found $E[Z|X] = X + 1/2$. So its mean is $E[X] + 1/2 = 1/2 + 1/2 = 1$. Property (v) says $E[E[Z|X]] = E[Z] = E[X] + E[Y] = 1/2 + 1/2 = 1$.

We also found $E[X|Z] = Z/2$. Thus the mean of this random variable is $E[Z]/2 = 1/2$. And property (v) says it should be $E[X] = 1/2$.

Example (random sums): Let X_n be an i.i.d. sequence with mean μ and variance σ^2. Let N be a RV that is independent of all the X_n and takes on the values $1, 2, 3, \cdots$. Let

$$S_N = \sum_{j=1}^{N} X_j$$

Note that the number of terms in the sum is random. We will find $E[S_N|N]$ and $E[S_N^2|N]$ and use them to compute the mean and variance of S_N.

Given that $N = n$, S_N is a sum with a fixed number of terms:

$$\sum_{j=1}^{n} X_j$$

So $E[S_N|N = n] = n\mu$. Thus $E[S_N|N] = N\mu$. Since the X_j are independent, their variances add and so

$$E[S_N^2|N = n] - (E[S_N|N = n])^2 = var[S_N|N = n] = n\sigma^2$$

So

$$E[S_N^2|N] = N\sigma^2 + (E[S_N|N])^2 = N\sigma^2 + N^2\mu^2$$

Now using property (v), we have

$$E[S_N] = E[E[S_N|N]] = E[N\mu] = \mu E[N]$$

$$E[S_N^2] = E[E[S_N^2|N]] = E[N\sigma^2 + N^2\mu^2] = \sigma^2 E[N] + \mu^2 E[N^2]$$

and so the variance of S_N is

$$var(S_N) = E[S_N^2] - (E[S_N])^2 = \sigma^2 E[N] + \mu^2 E[N^2] - \mu^2 E[N]^2 = \sigma^2 E[N] + \mu^2 var(N)$$

A.4 Conditional expectation as a "best approximation

We have said that $E[X|Y]$ is the function of Y that best approximates X. In this section we make this precise. We will assume we have discrete random variables. The main result of this section is true in the continuous case as well.

We need to make precise what it means to "be a function of Y." In the discrete case we can simply say that X is a function of Y if there is a function $g : \mathbb{R} \to \mathbb{R}$ such that $X = g(Y)$. (In the continuous case we need some condition that g be nice.) We start with a characterization of functions of Y.

Proposition 1 *Let X, Y be discrete RV's. Let y_n be the possible value of Y and $B_n = \{Y = y_n\}$. Then there is a function $g : \mathbb{R} \to \mathbb{R}$ such that $X = g(Y)$ if and only if X is constant on every event B_n.*

Proof: The direction $\Rightarrow$ is easy.

For the other direction, suppose X is constant on the B_n. Then we can write it as

$$X = \sum_n x_n 1_{B_n}$$

where 1_{B_n} is the random variable that is 1 on B_n and 0 on its complement. Define $g(y_n) = x_n$. For y not in the range of Y we define $g(y) = 0$. (It doesn't matter what we define it to be here.) Then $Y = g(X)$.

Theorem 2 *For any function $h : \mathbb{R} \to \mathbb{R}$,*

$$E[(X - E[X|Y])^2] \leq E[(X - h(Y))^2]$$

and we have equality if and only if $h(Y) = E[X|Y]$.

Proof: Let y_n be the possible values of Y and $B_n = \{Y = y_n\}$. Then $h(Y)$ is of the form

$$h(Y) = \sum_n x_n 1_{B_n}$$

for some x_n. (The number x_n is just the value of $h(Y)$ on B_n.) So

$$E[(X - h(Y))^2] = E[(X - \sum_n x_n 1_{B_n})^2]$$

We think of this as a function of all the x_n and try to find its mininum. It is quadratic in each x_n with a positive coef of x_n^2. So the mininum will occur at the critical point given by setting all the partial derivatives with respect to the x_n equal to 0.

$$\frac{\partial}{\partial x_j} E[(X - \sum_n x_n 1_{B_n})^2] = -2E[(X - \sum_n x_n 1_{B_n}) 1_{B_j}] = x_j P(B_j) - E[X 1_{B_j}]$$

It is easy to check that

$$E[X|Y = y_j] = \frac{E[X1_{B_j}]}{P(B_j)}$$

which completes the proof.