14.1: Conditional Expectation, Regression

Conditional expectation, given a random vector, plays a fundamental role in much of modern probability theory. Various types of "conditioning" characterize some of the more important random sequences and processes. The notion of conditional independence is expressed in terms of conditional expectation. Conditional independence plays an essential role in the theory of Markov processes and in much of decision theory.

We first consider an elementary form of conditional expectation with respect to an event. Then we consider two highly intuitive special cases of conditional expectation, given a random variable. In examining these, we identify a fundamental property which provides the basis for a very general extension. We discover that conditional expectation is a random quantity. The basic property for conditional expectation and properties of ordinary expectation are used to obtain four fundamental properties which imply the "expectationlike" character of conditional expectation. An extension of the fundamental property leads directly to the solution of the regression problem which, in turn, gives an alternate interpretation of conditional expectation.

Conditioning by an event

If a conditioning event C occurs, we modify the original probabilities by introducing the conditional probability measure $P(\cdot|C)$. In making the change form

$$P(A) \text{ to } P(A|C) = \frac{P(AC)}{P(C)}$$

we effectively do two things:

- We limit the possible outcomes to event C
- We "normalize" the probability mass by taking $P(C)$ as the new unit

It seems reasonable to make a corresponding modification of mathematical expectation when the occurrence of event C is known. The expectation $E[X]$ is the probability weighted average of the values taken on by X. Two possibilities for making the modification are suggested.

- We could replace the prior probability measure $P(\cdot)$ with the conditional probability measure $P(\cdot|C)$ and take the weighted average with respect to these new weights.
- We could continue to use the prior probability measure $P(\cdot)$ and modify the averaging process as follows:
  - Consider the values $P(\omega)$ for only those $\omega \in C$. This may be done by using the random variable $I_C X$ which has value $X(\omega)$ for $\omega \in C$ and zero elsewhere. The expectation $E[I_C X]$ is the probability weighted *sum* of those values taken on in C.
  - The weighted average is obtained by dividing by $P(C)$.

These two approaches are equivalent. For a simple random variable $X = sum_{k=1}^{n} t_k I_{A_k}$ in canonical form

$$E[I_C X]/P(C) = \sum_{k=1}^{n} E[t_k I_C I_{A_k}]/P(C) = \sum_{k=1}^{n} t_k P(C A_k)/P(C) = \sum_{k=1}^{n} t_k P(A_k|C)$$

The final sum is expectation with respect to the conditional probability measure. Arguments using basic theorems on expectation and the approximation of general random variables by simple random variables allow an extension to a general random variable X. The notion of a conditional distribution, given C, and taking weighted averages with respect to the conditional probability is intuitive and natural in this case. However, this point of view is limited. In order to display a natural relationship with more the general concept of conditioning with repspect to a random vector, we adopt the following

Definition

The conditional expectation of X, given event C with positive probability, is the quantity

$$E[X|C] = \frac{E[I_C X]}{P(C)} = \frac{E[I_C X]}{E[I_C]}$$

Remark. The product form $E[X|C]P(C) = E[I_C X]$ is often useful.

Example 14.1.1 A numerical example

Suppose $X \sim$ exponential (λ) and $C = \{1/\lambda \le X \le 2/\lambda\}$. Now $I_C = I_M(X)$ where $M = [1/\lambda, 2/\lambda]$.

$$P(C) = P(X \ge 1/\lambda) - P(X > 2/\lambda) = e^{-1}e^{-2} \quad \text{and}$$

$$E[I_C X] = \int I_M(t) t\lambda e^{-\lambda t}\, dt = \int_{1/\lambda}^{2/\lambda} t\lambda e^{-\lambda t}\, dt = \frac{1}{\lambda}(2e^{-1} - 3e^{-2})$$

Thus

$$E[X|C] = \frac{2e^{-1} - 3e^{-2}}{\lambda(e^{-1} - e^{-2})} \approx \frac{1.418}{\lambda}$$

Conditioning by a random vector—discrete case

Suppose $X = \sum_{i=1}^{n} t_i I_{A_i}$ and $Y = \sum_{j=1}^{m} u_j I_{B_j}$ in canonical form. We suppose $P(A_i) = P(X = t_i) > 0$ and $P(B_j) = P(Y = u_j) > 0$, for each permissible i, j. Now

$$P(Y = u_j | X = t_i) = \frac{P(X = t_i, Y = u_j)}{P(X = t_i)}$$

We take the expectation relative to the conditional probability $P(\cdot|X = t_i)$ to get

$$E[g(Y)|X = t_i] = \sum_{j=1}^{m} g(u_j) P(Y = u_j | X = t_i) = e(t_i)$$

Since we have a value for each t_i in the range of X, the function $e(\cdot)$ is defined on the range of X. Now consider any reasonable set M on the real line and determine the expectation

$$\begin{aligned} E[I_M(X)g(Y)] &= \sum_{i=1}^{n} \sum_{j=1}^{m} I_M(t_i) g(u_j) P(X = t_i, Y = u_j) \\ &= \sum_{i=1}^{n} I_M(t_i) [\sum_{j=1}^{m} g(u_j) P(Y = u_j | X = t_i)] P(X = t_i) \\ &= \sum_{i=1}^{n} I_M(t_i) e(t_i) P(X = t_i) = E[I_M(X) e(X)] \end{aligned}$$

We have the pattern

$$(A)\ E[I_M(X)g(Y)] = E[I_M(X)e(X)] \text{ where } e(t_i) = E[g(Y)|X = t_i]$$

for all t_i in the range of X.

We return to examine this property later. But first, consider an example to display the nature of the concept.

Example 14.1.2 Basic calculations and interpretation

Suppose the pair $\{X, Y\}$ has the joint distribution

$P(X = t_i, Y = u_j)$

$X =$	0	1	4	9
$Y = 2$	0.05	0.04	0.21	0.15
0	0.05	0.01	0.09	0.10
-1	0.10	0.05	0.10	0.05
PX	0.20	0.10	0.40	0.30

Calculate $E[Y|X = t_i]$ for each possible value t_i taken on by X

$$\begin{aligned} E[Y|X = 0] &= -1\frac{0.10}{0.20} + 0\frac{0.05}{0.20} + 2\frac{0.05}{0.20} \\ &= (-1 \cdot 0.10 + 0 \cdot 0.05 + 2 \cdot 0.05)/0.20 = 0 \\ E[Y|X = 1] &= (-1 \cdot 0.05 + 0 \cdot 0.01 + 2 \cdot 0.04)/0.10 = 0.30 \\ E[Y|X = 4] &= (-1 \cdot 0.10 + 0 \cdot 0.09 + 2 \cdot 0.21)/0.40 = 0.80 \\ E[Y|X = 9] &= (-1 \cdot 0.05 + 0 \cdot 0.10 + 2 \cdot 0.15)/0.10 = 0.83 \end{aligned}$$

The pattern of operation in each case can be described as follows:

- For the i th column, multiply each value u_j by $P(X = t_i, Y = u_j)$, sum, then divide by $P(X = t_i)$.

The following interpretation helps visualize the conditional expectation and points to an important result in the general case.

- For each t_i we use the mass distributed "above" it. This mass is distributed along a vertical line at values u_j taken on by Y. The result of the computation is to determine the *center of mass* for the *conditional distribution* above $t = t_i$. As in the case of ordinary expectations, this should be the best estimate, in the mean-square sense, of Y when $X = ti$. We examine that possibility in the treatment of the regression problem in Section: The regression problem.

Although the calculations are not difficult for a problem of this size, the basic pattern can be implemented simply with MATLAB, making the handling of much larger problems quite easy. This is particularly useful in dealing with the simple approximation to an absolutely continuous pair.

```
X = [0 1 4 9];            % Data for the joint distribution
Y = [-1 0 2];
P = 0.01*[ 5  4 21 15; 5  1  9 10; 10  5 10  5];
jcalc                     % Setup for calculations
Enter JOINT PROBABILITIES (as on the plane)  P
Enter row matrix of VALUES of X  X
Enter row matrix of VALUES of Y  Y
 Use array operations on matrices X, Y, PX, PY, t, u, and P
EYX = sum(u.*P)./sum(P);   % sum(P) = PX  (operation sum yields column sums)
disp([X;EYX]')             % u.*P = u_j P(X = t_i, Y = u_j) for all i, j
         0         0
    1.0000    0.3000
    4.0000    0.8000
    9.0000    0.8333
```

The calculations extend to $E[g(X, Y)|X = t_i]$. Instead of values of u_j we use values of $g(t_i, u_j)$ in the calculations. Suppose $Z = g(X, Y) = Y^2 - 2XY$.

```
G = u.^2 - 2*t.*u;         % Z = g(X,Y) = Y^2 - 2XY
EZX = sum(G.*P)./sum(P);   % E[Z|X=x]
disp([X;EZX]')
         0    1.5000
    1.0000    1.5000
    4.0000   -4.0500
    9.0000  -12.8333
```

Conditioning by a random vector — absolutely continuous case

Suppose the pair $\{X, Y\}$ has joint density function f_{XY}. We seek to use the concept of a conditional distribution, given $X = t$. The fact that $P(X = t) = 0$ for each t requires a modification of the approach adopted in the discrete case. Intuitively, we consider the *conditional density*

$$f_{Y|X}(u|t) \geq 0, \int f_{Y|X}(u|t)\, du = \frac{1}{f_X(t)} \int f_{XY}(t, u)\, du = f_X(t)/f_X(t) = 1$$

We define, in this case,

$$E[g(Y)|X = t] = \int g(u) f_{Y|X}(u|t)\, du = e(t)$$

The function $e(\cdot)$ is defined for $f_X(t) > 0$, hence effectively on the range of X. For any reasonable set M on the real line,

$$E[I_M(X)g(Y)] = \int\int I_M(t)g(u)f_{XY}(t,u)\,dudt = \int I_M(t)[\int g(u)f_{Y|X}(u|t)\,du]f_X(u)\,dt$$
$$= \int I_M(t)e(t)f_X(t)\,dt \text{ , where } e(t) = E[g(Y)|X=t]$$

Thus we have, as in the discrete case, for each t in the range of X.

$$(A)\ E[I_M(X)g(Y)] = E[I_M(X)e(X)] \text{ where } e(t) = E[g(Y)|X=t]$$

Again, we postpone examination of this pattern until we consider a more general case.

Example 14.1.3 Basic calculation and interpretation

Suppose the pair $\{X, Y\}$ has joint density $f_{XY}(t,u) = \frac{6}{5}(t+2u)$ on the triangular region bounded by $t=0$, $u=1$, and $u=t$ (see Figure 14.1.1). Then

$$f_X(t) = \frac{6}{5}\int_t^1 (t+2u)\,du = \frac{6}{5}(1+t-2t^2) \quad , 0 \le t \le 1$$

By definition, then,

$$f_{Y|X}(u|t) = \frac{t+2u}{1+t-2t^2} \text{ on the triangle (zero elsewhere)}$$

We thus have

$$E[Y|X=t] = \int u f_{Y|X}(u|t)\,du = \frac{1}{1+t-2t^2}\int_t^1 (tu+2u^2)\,du = \frac{4+3t-7t^3}{6(1+t-2t^2)} \quad (0 \le t < 1)$$

Theoretically, we must rule out $t=1$ since the denominator is zero for that value of t. This causes no problem in practice.

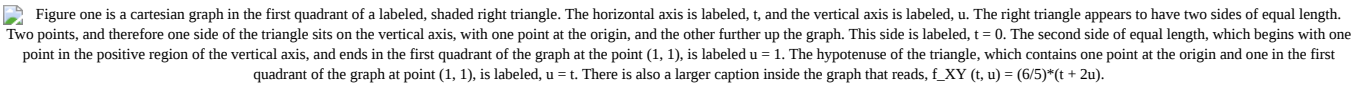

Figure 14.1.1. The density function for Example 14.1.3

We are able to make an interpretation quite analogous to that for the discrete case. This also points the way to practical MATLAB calculations.

- For any t in the range of X (between 0 and 1 in this case), consider a narrow vertical strip of width Δt with the vertical line through t at its center. If the strip is narrow enough, then $f_{XY}(t,u)$ does not vary appreciably with t for any u.
- The mass in the strip is approximately
$$\text{Mass} \approx \Delta t \int f_{XY}(t,u)\,du = \Delta t f_X(t)$$
- The moment of the mass in the strip about the line $u=0$ is approximately
$$\text{Momemt} \approx \Delta t \int u f_{XY}(t,u)\,du$$
- The center of mass in the strip is
$$\text{Center of mass} = \frac{\text{Moment}}{\text{Mass}} \approx \frac{\Delta \int u f_{XY}(t,u)\,du}{\Delta t f_X(t)} = \int u f_{Y|X}(u|t)\,du = e(t)$$

This interpretation points the way to the use of MATLAB in approximating the conditional expectation. The success of the discrete approach in approximating the theoretical value in turns supports the validity of the interpretation. Also, this points to the general result on regression in the section, "The Regression Problem".

In the MATLAB handling of joint absolutely continuous random variables, we divide the region into narrow vertical strips. Then we deal with each of these by dividing the vertical strips to form the grid structure. The center of mass of the discrete distribution over one of the *t* chosen for the approximation must lie close to the actual center of mass of the probability in the strip. Consider the MATLAB treatment of the example under consideration.

```
f = '(6/5)*(t + 2*u).*(u>=t)';                  % Density as string variable
tuappr
Enter matrix [a b] of X-range endpoints  [0 1]
```

```
Enter matrix [c d] of Y-range endpoints  [0 1]
Enter number of X approximation points  200
Enter number of Y approximation points  200
Enter expression for joint density  eval(f)     % Evaluation of string variable
Use array operations on X, Y, PX, PY, t, u, and P
EYx = sum(u.*P)./sum(P);                        % Approximate values
eYx = (4 + 3*X - 7*X.^3)./(6*(1 + X - 2*X.^2)); % Theoretical expression
plot(X,EYx,X,eYx)
% Plotting details             (see Figure 14.1.2)
```

—□

Figure two is a graph titled, theoretical and approximate conditional expectation. The horizontal axis is labeled, t, and the vertical axis is labeled E[X | Y = t]. The values on the horizontal axis are from 0 to 1 in increments of 0.1. The values on the vertical axis range from 0.65 to 1 in increments of 0.05. There is a caption inside the graph that reads fXY (t, u) = (6/5)*(t + 2u), for 0 ≤ t ≤ u ≤ 1. There are two plots on this graph. The first is a solid, smooth line labeled Approximate. the second is a smooth, dashed line, labeled theoretical. Both lines follow the same path on the graph, and are so closely fitted that they are nearly indistinguishable. They begin on the lower left side, at approximately (0, 0.67), and continue towards the right with a slightly negative slope for a very small segment, until approximately (0.08, 0.66), where the plots begin gradually increasing at an increasing rate. By midway across the graph, at approximately (0.4, 0.74), the slope of the graph remains positive and constant, and continues in a linear fashion from this point to the top-right corner of the graph, at (1, 1).

Figure 14.1.2. Theoretical and approximate conditional expectation for above.

The agreement of the theoretical and approximate values is quite good enough for practical purposes. It also indicates that the interpretation is reasonable, since the approximation determines the center of mass of the discretized mass which approximates the center of the actual mass in each vertical strip.

Extension to the general case

Most examples for which we make numerical calculations will be one of the types above. Analysis of these cases is built upon the intuitive notion of conditional distributions. However, these cases and this interpretation are rather limited and do not provide the basis for the range of applications—theoretical and practical—which characterize modern probability theory. We seek a basis for extension (which includes the special cases). In each case examined above, we have the property

$$(A)\ E[I_M(X)g(Y)] = E[I_M(X)e(X)] \text{ where } e(t) = E[g(Y)|X = t]$$

for all t in the range of X.

We have a tie to the simple case of conditioning with respect to an event. If $C = \{X \in M\}$ has positive probability, then using $I_C = I_M(X)$ we have

$(B)\ E[I_M(X)g(Y)] = E[g(Y)|X \in M]P(X \in M)$

wo properties of expectation are crucial here:

By the uniqueness property (E5), since (A) holds for all reasonable (Borel) sets, then $e(X)$ is unique a.s. (i.e., except for a set of ω of probability zero).
By the special case of the Radon Nikodym theorem (E19), the function $e(\cdot)$ *always exists* and is such that random variable $e(X)$ is unique a.s.

We make a definition based on these facts.

Definition

The *conditional expectation* $E[g(Y)|Y = t] = e(t)$ is the a.s. unique function defined on the range of X such that

$$(A)\ E[I_M(X)g(Y)] = E[I_M(X)e(X)] \text{ for all Borel sets } M$$

Note that $e(X)$ is a random variable and $e(\cdot)$ is a function. Expectation $E[g(Y)]$ is always a constant. The concept is abstract. At this point it has little apparent significance, except that it must include the two special cases studied in the previous sections. Also, it is not clear why the term conditional *expectation* should be used. The justification rests in certain formal properties which are based on the defining condition (A) and other properties of expectation.

In Appendix F we tabulate a number of key properties of conditional expectation. The condition (A) is called property (CE1). We examine several of these properties. For a detailed treatment and proofs, any of a number of books on measure-theoretic probability

may be consulted.

(CE1) **Defining condition.** $e(X) = E[g(Y)|X]$ a.s. iff

$$E[I_M(X)g(Y)] = E[I_M(X)e(X)] \text{ for each Borel set } M \text{ on the codomain of } X$$

Note that X and Y do not need to be real valued, although $g(Y)$ is real valued. This extension to possible vector valued X and Y is extremely important. The next condition is just the property (B) noted above.

(CE1a) If $P(X \in M) > 0$, then $E[I_M(X)e(X)] = E[g(Y)|X \in M]P(X \in M)$

The special case which is obtained by setting M to include the entire range of X so that $I_M(X(\omega)) = 1$ for all ω is useful in many theoretical and applied problems.

(CE1b) Law of total probability. $E[g(Y)] = E\{E[g(Y)|X]\}$

It may seem strange that we should complicate the problem of determining $E[g(Y)]$ by first getting the conditional expectation $e(X) = E[g(Y)|X]$ then taking expectation of that function. Frequently, the data supplied in a problem makes this the expedient procedure.

Exercise 14.1.4 Use of the law of total probability

Suppose the time to failure of a device is a random quantity $X \sim$ exponential (μ), where the parameter u is the value of a parameter random variable H. Thus

$$f_{X|H}(t|u) = ue^{-ut} \text{ for } t \geq 0$$

If the parameter random variable $H \sim$ uniform (a, b), determine the expected life $E[X]$ of the device.

Solution

We use the law of total probability:

$$E[X] = E\{E[X|H]\} = \int E[X|H = u] f_H(u)\, du$$

Now by assumption

$$E[X|H = u] = 1/u \text{ and } f_H(u) = \frac{1}{b-a},\ a < u < b$$

Thus

$$E[X] = \frac{1}{b-a} \int_a^b \frac{1}{u} du = \frac{\ln(b/a)}{b-a}$$

For $a = 1/100$, $b = 2/100$, $E[X] = 100\ln(2) \approx 69.31$.

The next three properties, linearity, positivity/monotonicity, and monotone convergence, along with the defining condition provide the "expectation like" character. These properties for expectation yield most of the other essential properties for expectation. A similar development holds for conditional expectation, with some reservation for the fact that $e(X)$ is a random variable, unique a.s. This restriction causes little problem for applications at the level of this treatment.

In order to get some sense of how these properties root in basic properties of expectation, we examine one of them.

(CE2) **Linearity**. For any constants a, b

$$E[ag(Y) + bh(Z)|X] = aE[g(Y)|X] + bE[h(Z)|X] \quad \text{a.s.}$$

VERIFICATION

Let $e_1(X) = E[g(Y)|X]$, $e_2[X] = E[h(Z)|X]$, and $e(X) = E[ag(Y) + bh(Z)|X]$ a.s.

$$\begin{aligned} E[I_M(X)e(X)] &= E\{I_M(X)[ag(Y) + bh(Z)]\} \text{ a.s.} && \text{by(CE1)} \\ &= aE[I_M(X)g(Y)] + bE[I_M(X)h(Z)] \text{ a.s.} && \text{by linearity of expectation} \\ &= aE[I_M(X)e_1(X)] + bE[I_M(X)e_2(X)] \text{ a.s.} && \text{by (CE1)} \\ &= E\{I_M(X)[ae_1(X) + be_2(X)]\} \text{ a.s.} && \text{by linearity of expectation} \end{aligned}$$

Since the equalities hold for any Borel M, the uniqueness property (E5) for expectation implies

$$e(X) = ae_1(X) = be_2(X) \text{ a.s.}$$

This is property (CE2). An *extension to any finite linear combination* is easily established by mathematical induction.

— □

Property (CE5) provides another condition for independence.

(CE5) **Independence**. $\{X, Y\}$ is an independent pair

iff $E[g(Y)|X] = E[g(Y)]$ a.s. for all Borel functions g
iff $E[I_N(Y)|X] = E[I_N(Y)]$ a.s. for all Borel sets N on the codomain of Y

Since knowledge of X does not affect the likelihood that Y will take on any set of values, then conditional expectation should not be affected by the value of X. The resulting constant value of the conditional expectation must be $E[g(Y)]$ in order for the law of total probability to hold. A formal proof utilizes uniqueness (E5) and the product rule (E18) for expectation.

Property (CE6) forms the basis for the solution of the regresson problem in the next section.

(CE6) $e(X) = E[g(Y)|X]$ a.s. iff $E[h(X)g(Y)] = E[h(X)e(X)]$ a.s. for any Borel function h

Examination shows this to be the result of replacing $I_M(X)$ in (CE1) with arbitrary $h(X)$. Again, Again, to get some insight into how the various properties arise, we sketch the ideas of a proof of (CE6).

IDEAS OF A PROOF OF (CE6)

For $h(X) = I_M(X)$, this is (CE1).

For $h(X) = \sum_{i=1}^{n} a_i I_{M_i}(X)$, the result follows by linearity.

For $h \geq 0$, $g \geq 0$, there is a seqence of nonnegative, simple $h_n nearrow h$. Now by positivity, $e(X) \geq 0$. By monotone convergence (CE4),

$$E[h_n(X)g(Y)] \nearrow E[h(X)g(Y)] \text{ and } E[h_n(X)e(X)] \nearrow E[h(X)e(X)]$$

Since corresponding terms in the sequences are equal, the limits are equal.
For $h = h^+ - h^-$, $g \geq 0$, the result follows by linearity (CE2).
For $g = g^+ - g^-$, the result again follows by linearity.

— □

Properties (CE8) and (CE9) are peculiar to conditional expectation. They play an essential role in many theoretical developments. They are essential in the study of Markov sequences and of a class of random sequences known as submartingales. We list them here (as well as in Appendix F) for reference.

(CE8) $E[h(X)g(Y)|X] = h(X)E[g(Y)|X]$ a.s. for any Borel function h

This property says that any function of the conditioning random vector may be treated as a constant factor. This combined with (CE10) below provide useful aids to computation.

(CE9) **Repeated conditioning**

If $X = h(W)$, then $E\{E[g(Y)|X|W\} = E\{E[g(Y)|W|X\} = E[g(Y)|X]$ a.s.

This somewhat formal property is highly useful in many theoretical developments. We provide an interpretation after the development of regression theory in the next section.

The next property is highly intuitive and very useful. It is easy to establish in the two elementary cases developed in previous sections. Its proof in the general case is quite sophisticated.

(CE10) Under conditions on g that are nearly always met in practice

$E[g(X, Y)|X = t] = E[g(t, Y)|X = t]$ a.s. $[P_X]$
If $\{X, Y\}$ is independent, then $E[g(X, Y)|X = t] = E[g(t, Y)]$ a.s. $[P_X]$

It certainly seem reasonable to suppose that if $X = t$, then we should be able to replace X by t in $E[g(X, Y)|X = t]$ to get $E[g(t, Y)|X = t]$. Property (CE10) assures this. If $\{X, Y\}$ is an independent pair, then the value of X should not affect the value

of Y, so that $E[g(t,Y)|X=t] = E[g(t,Y)]$ a.s.

Example 14.1.5 Use of property (CE10)

Consider again the distribution for Example 14.1.3. The pair $\{X, Y\}$ has density

$f_{XY}(t,u) = \frac{6}{5}(t+2u)$ on the triangular region bounded by $t=0$, $u=1$, and $u=t$

We show in Example 14.1.3 that

$$E[Y|X=t] = \frac{4+3t-7t^3}{6(1+t-2t^2)} \quad 0 \le t < 1$$

Let $Z = 3X^2 + 2XY$. Determine $E[Z|X=t]$.

Solution

By linearity, (CE8), and (CE10)

$$E[Z|X=t] = 3t^2 + 2tE[Y|X=t] = 3t^2 + \frac{4t+3t^2-7t^4}{3(1+t-2t^2)}$$

Conditional probability

In the treatment of mathematical expectation, we note that probability may be expressed as an expectation

$$P(E) = E[I_E]$$

For conditional probability, given an event, we have

$$E[I_E|C] = \frac{E[I_E I_C]}{P(C)} = \frac{P(EC)}{P(C)} = P(E|C)$$

In this manner, we extend the concept conditional expectation.

Definition

The *conditional probability* of event E, given X, is

$$P(E|X) = E[I_E|X]$$

Thus, there is no need for a separate theory of conditional probability. We may define the conditional distribution function

$$F_{Y|X}(u|X) = P(Y \le u|X) = E[I_{(-\infty,u]}(Y)|X]$$

Then, by the law of total probability (CE1b),

$$F_Y(u) = E[F_{Y|X}(u|X)] = \int F_{Y|X}(u|t) F_X(dt)$$

If there is a conditional density $f_{Y|X}$ such that

$$P(Y \in M|X=t) = \int_M f_{Y|X}(r|t)\, dr$$

then

$$F_{Y|X}(u|t) = \int_{-\infty}^{u} f_{Y|X}(r|t)\, dr \text{ so that } f_{Y|X}(u|t) = \frac{\partial}{\partial u} F_{Y|X}(u|t)$$

A careful, measure-theoretic treatment shows that it may *not* be true that $F_{Y|X}(\cdot|t)$ is a distribution function for all t in the range of X. However, in applications, this is seldom a problem. Modeling assumptions often start with such a family of distribution functions or density functions.

Example 14.1.6 The conditional distribution function

As in Example 14.1.4, suppose X ~ exponential (u), where the parameter u is the value of a parameter random variable H. If the parameter random variable H ~ uniform (a,b), determine the distribution fuction F_X.

Solution

As in Example 14.1.4, take the assumption on the conditional distribution to mean

$$f_{X|H}(t|u) = ue^{-ut} \quad t \geq 0$$

Then

$$F_{X|H}(t|u) = \int_0^1 ue^{-us}\, ds = 1 - e^{-ut} \quad 0 \leq t$$

By the law of total probability

$$F_X(t) = \int F_{X|H}(t|u) f_H(u)\, du = \frac{1}{b-a}\int_a^b (1 - e^{-ut})\, du = 1 - \frac{1}{b-a}\int_a^b e^{-ut}\, du$$

$$= 1 - \frac{1}{t(b-a)}[e^{-bt} - e^{-at}]$$

Differentiation with respect to t yields the expression for $f_X(t)$

$$f_X(t) = \frac{1}{b-a}[(\frac{1}{t^2} + \frac{b}{t})e^{-bt} - (\frac{1}{t^2} + \frac{a}{t})e^{-at}] \quad t > 0$$

The following example uses a discrete conditional distribution and marginal distribution to obtain the joint distribution for the pair.

Example 14.1.7 A random number N of Bernoulli trials

A number N is chosen by a random selection from the integers from 1 through 20 (say by drawing a card from a box). A pair of dice is thrown N times. Let S be the number of "matches" (i.e., both ones, both twos, etc.). Determine the joint distribution for $[N, S]$.

Solution

N ~ uniform on the integers 1 through 20. $P(N = i) = 1/20$ for $1 \leq i \leq 20$. Since there are 36 pairs of numbers for the two dice and six possible matches, the probability of a match on any throw is 1/6. Since the i throws of the dice constitute a Bernoulli sequence with probability 1/6 of a success (a match), we have S conditionally binomial (i, 1/6), given $N = i$. For any pair (i, j), $0 \leq j \leq i$,

$$P(N = i, S = j) = P(S = j|N = i)P(N = i)$$

Now $E[S|N = i) = i/6$, so that

$$E[S] = \frac{1}{6} \cdot \frac{1}{20}\sum_{i=1}^{20} i = \frac{20 \cdot 21}{6 \cdot 20 \cdot 2} = \frac{7}{4} = 1.75$$

The following MATLAB procedure calculates the joint probabilities and arranges them "as on the plane."

```
% file randbern.m
p  = input('Enter the probability of success  ');
N  = input('Enter VALUES of N  ');
PN = input('Enter PROBABILITIES for N  ');
n  = length(N);
m  = max(N);
S  = 0:m;
P  = zeros(n,m+1);
for i = 1:n
  P(i,1:N(i)+1) = PN(i)*ibinom(N(i),p,0:N(i));
end
PS = sum(P);
P  = rot90(P);
disp('Joint distribution N, S, P, and marginal PS')
```

```
randbern                            % Call for the procedure
Enter the probability of success  1/6
Enter VALUES of N  1:20
Enter PROBABILITIES for N  0.05*ones(1,20)
Joint distribution N, S, P, and marginal PS
ES = S*PS'
ES =  1.7500                          % Agrees with the theoretical value
```

The regression problem

We introduce the regression problem in the treatment of linear regression. Here we are concerned with more general regression. A pair $\{X, Y\}$ of real random variables has a joint distribution. A value $X(\omega)$ is observed. We desire a rule for obtaining the "best" estimate of the corresponding value $Y(\omega)$. If $Y(\omega)$ is the actual value and $r(X(\omega))$ is the estimate, then $Y(\omega) - r(X(\omega))$ is the error of estimate. The best estimation rule (function) $r(\cdot)$ is taken to be that for which the average square of the error is a minimum. That is, we seek a function r such that

$$E[(Y - r(X))^2] \text{ is a minimum}$$

In the treatment of linear regression, we determine the best affine function, $u = at + b$. The optimum function of this form defines the regression *line* of Y on X. We now turn to the problem of finding the best function r, which may in some cases be an affine function, but more often is not.

We have some hints of possibilities. In the treatment of expectation, we find that the best constant to approximate a random variable in the mean square sense is the mean value, which is the center of mass for the distribution. In the interpretive Example 14.2.1 for the discrete case, we find the conditional expectation $E[Y|X = t_i]$ is the center of mass for the conditional distribution at $X = t_i$. A similar result, considering thin vertical strips, is found in Example 14.1.3 for the absolutely continuous case. This suggests the possibility that $e(t) = E[Y|X = t]$ might be the best estimate for Y when the value $X(\omega) = t$ is observed. We investigate this possibility. The property (CE6) proves to be key to obtaining the result.

Let $e(X) = E[Y|X]$. We may write (CE6) in the form $E[h(X)(Y - e(X))] = 0$ for any reasonable function h. Consider

$$E[(Y - r(X))^2] = E[(Y - e(X) + e(X) - r(X))^2]$$
$$= E[(Y - e(X))^2] + E[(e(X) - r(X))^2] + 2E[(Y - e(X))(r(X) - e(X))]$$

Now $e(X)$ is fixed (a.s.) and for any choice of r we may take $h(X) = r(X) - e(X)$ to assert that

$$E[Y - e(X))(r(X) - e(X))] = E[(Y - e(X))h(X)] = 0$$

Thus

$$E[(Y - r(X))^2] = E[(Y - e(X))^2] + E[(e(X) - r(X))^2]$$

The first term on the right hand side is fixed; the second term is nonnegative, with a minimum at zero iff $r(X) = e(X)$ a.s. Thus, $r = e$ is the best rule. For a given value $X(\omega) = t$ the best mean square esitmate of Y is

$$u = e(t) = E[Y|X = t]$$

The graph of $u = e(t)$ vs t is known as the *regression curve of Y on X*. This is defined for argument t in the range of X, and is unique except possibly on a set N such that $P(X \in N) = 0$. Determination of the regression curve is thus determination of the conditional expectation.

Example 14.1.8 Regression curve for an independent pair

If the pair $\{X, Y\}$ is independent, then $u = E[Y|X = t] = E[Y]$, so that the regression curve of Y on X is the horizontal line through $u = E[Y]$. This, of course, agrees with the regression line, since $\text{Cov}[X, Y] = 0$ and the regression line is $u = 0 = E[Y]$.

The result extends to functions of X and Y. Suppose $Z = g(X, Y)$. Then the pair $\{X, Z\}$ has a joint distribution, and the best mean square estimate of Z given $X = t$ is $E[Z|X = t]$.

Example 14.1.9 Estimate of a function of $\{X, Y\}$

Suppose the pair $\{X, Y\}$ has joint density $f_{XY}(t,u) = 60t^2u$ for $0 \le t \le 1$, $0 \le u \le 1-t$. This is the triangular region bounded by $t = 0$, $u = 0$, and $u = 1 - t$ (see Figure 14.1.3). Integration shows that

$$f_X(t) = 30t^2(1-t)^2,\ 0 \le t \le 1 \text{ and } f_{Y|X}(u|t) = \frac{2u}{(1-t)^2} \text{ on the triangle}$$

Consider

$$Z = \begin{cases} X^2 & \text{for } X \le 1/2 \\ 2Y & \text{for } X > 1/2 \end{cases} = I_M(X)X^2 + I_N(X)2Y$$

where $M = [0, 1/2]$ and $N = (1/2, 1]$. Determine $E[Z|X = t]$.

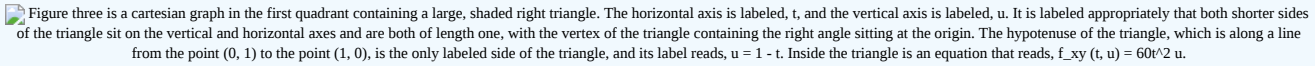
Figure three is a cartesian graph in the first quadrant containing a large, shaded right triangle. The horizontal axis is labeled, t, and the vertical axis is labeled, u. It is labeled appropriately that both shorter sides of the triangle sit on the vertical and horizontal axes and are both of length one, with the vertex of the triangle containing the right angle sitting at the origin. The hypotenuse of the triangle, which is along a line from the point (0, 1) to the point (1, 0), is the only labeled side of the triangle, and its label reads, u = 1 - t. Inside the triangle is an equation that reads, f_xy (t, u) = 60t^2 u.

Figure 14.1.3. The density function for Example 14.1.9.

Solution By linearity and (CE8).

$$E[Z|X = t] = E[I_M(X)X^2||X = t] + E[I_N(X)2Y||X = t] = I_M(t)t^2 + I_N(t)2E[Y|X = t]$$

Now

$$E[Y|X = t] = \int u f_{Y|X}(u|t)\,du = \frac{1}{(1-t)^2}\int_0^{1-t} 2u^2\,du = \frac{2}{3}\cdot\frac{(1-t)^3}{(1-t)^2} = \frac{2}{3}(1-t) \quad , 0 \le t < 1$$

so that

$$E[Z|X = t] = I_M(t)t^2 + I_N(t)\frac{4}{3}(1-t)$$

Note that the indicator functions separate the two expressions. The first holds on the interval $M = [0, 1/2]$ and the second holds on the interval $N = (1/2, 1]$. The two expressions $t^2 \backslash 0 and (4/3) \backslash ((1-t)$ *must not be added*, for this would give an expression *incorrect for all t* in the range of X.

APPROXIMATION

```
tuappr
Enter matrix [a b] of X-range endpoints  [0 1]
Enter matrix [c d] of Y-range endpoints  [0 1]
Enter number of X approximation points  100
Enter number of Y approximation points  100
Enter expression for joint density  60*t.^2.*u.*(u<=1-t)
Use array operations on X, Y, PX, PY, t, u, and P
G = (t<=0.5).*t.^2 + 2*(t>0.5).*u;
EZx = sum(G.*P)./sum(P);                        % Approximation
eZx = (X<=0.5).*X.^2 + (4/3)*(X>0.5).*(1-X);    % Theoretical
plot(X,EZx,'k-',X,eZx,'k-.')
% Plotting details                              % See Figure 14.1.4
```

The fit is quite sufficient for practical purposes, in spite of the moderate number of approximation points. The difference in expressions for the two intervals of X values is quite clear.

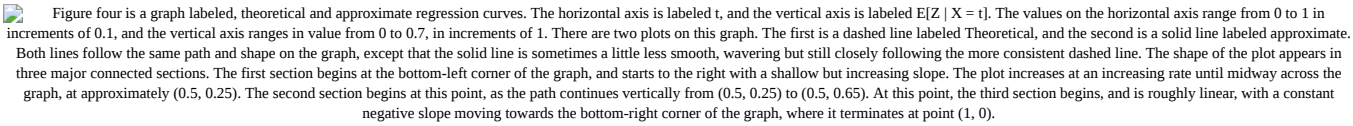
Figure four is a graph labeled, theoretical and approximate regression curves. The horizontal axis is labeled t, and the vertical axis is labeled E[Z | X = t]. The values on the horizontal axis range from 0 to 1 in increments of 0.1, and the vertical axis ranges in value from 0 to 0.7, in increments of 1. There are two plots on this graph. The first is a dashed line labeled Theoretical, and the second is a solid line labeled approximate. Both lines follow the same path and shape on the graph, except that the solid line is sometimes a little less smooth, wavering but still closely following the more consistent dashed line. The shape of the plot appears in three major connected sections. The first section begins at the bottom-left corner of the graph, and starts to the right with a shallow but increasing slope. The plot increases at an increasing rate until midway across the graph, at approximately (0.5, 0.25). The second section begins at this point, as the path continues vertically from (0.5, 0.25) to (0.5, 0.65). At this point, the third section begins, and is roughly linear, with a constant negative slope moving towards the bottom-right corner of the graph, where it terminates at point (1, 0).

Figure 14.1.4. Theoretical and approximate regression curves for Example 14.1.9

Example 14.1.10 Estimate of a function of $\{X, Y\}$

Suppose the pair $\{X, Y\}$ has joint density $f_{XY}(t,u) = \frac{6}{5}(t^2 + u)$, on the unit square $0 \le t \le 1$, $0 \le u \le 1$ (see Figure 14.1.5). The usual integration shows

$$f_X(t) = \frac{3}{5}(2t^2 + 1),\ 0 \le t \le 1, \text{ and } f_{Y|X}(u|t) = 2\frac{t^2 + u}{2t^2 + 1} \text{ on the square}$$

Consider

$$Z = \begin{cases} 2X^2 & \text{for } X \le Y \\ 3XY & \text{for } X > Y \end{cases} \quad I_Q(X,Y)2X^2 + I_{Q^c}(X,Y)3XY \text{, where } Q = \{(t,u) : u \ge t\}$$

Determine $E[Z|X = t]$.

Solution

$$E[Z|X = t] = 2t^2 \int I_Q(t,u) f_{Y|X}(u|t) + 3t \int I_{Q^c}(t,u) u f_{Y|X}(u|t)\, du$$

$$= \frac{4t^2}{2t^2 + 1} \int_t^1 (t^2 + u)\, du + \frac{6t}{2t^2 + 1} \int_0^t (t^2 u + u^2)\, du = \frac{-t^5 + 4t^4 + 2t^2}{2t^2 + 1}, \quad 0 \le t \le 1$$

Figure five is a cartesian graph containing two equal right triangles that put together at their hypotenuse create a large square. The horizontal axis is labeled, t, and the vertical axis is labeled, u. Each axis marked only with the value 1. The points (0, 0), (0, 1), (1, 1), and (1, 0) are vertices of the square. A diagonal dashed line from point (0, 0) through point (1, 1) is labeled u = t and divides the square into two triangles. The two sides of the triangle not sitting on an axis are labeled, with the horizontal side from (0, 1) to (1, 1) labeled, u = 1, and the vertical side from (1, 0) to (1, 1) labeled, t = 1. The triangle above the diagonal line is labeled, Q, and the triangle below is labeled Q^C. A large equation is printed below the graph that reads, f_XY (t, u) = (6/5)*(t^2 + u).

Figure 14.1.5. The density and regions for Example 14.1.10

Note the different role of the indicator functions than in Example 14.1.9. There they provide a separation of two parts of the result. Here they serve to set the effective limits of integration, but sum of the two parts is needed for each t.

Figure six is a graph labeled, theoretical and approximate regression curves. The horizontal axis is labeled, t, and the vertical axis is labeled, E[Z | X = t]. The values on the horizontal axis range from 0 to 1 in increments of 0.1. The values on the vertical axis range from 0 to 1.8 in increments of 0.2. There are two plots on the graph, but both follow the same shape so closely that they are indistinguishable. One is a solid line, labeled Approximate, and the other is a dashed line, labeled Theoretical. The shape begins at the bottom-right corner of the graph at (0, 0). It initially moves to the right at a shallow positive slope. As it continues to move to the right, it begins increasing at an increasing rate until approximately (0.6, 0.7) where it maintains a constant positive slope. The plot continues this slope up to the upper-right corner of the graph, where it terminates at approximately (1, 1.65).

Figure 14.1.6. Theoretical and approximate regression curves for Example 14.1.10

APPROXIMATION

```
tuappr
Enter matrix [a b] of X-range endpoints  [0 1]
Enter matrix [c d] of Y-range endpoints  [0 1]
Enter number of X approximation points  200
Enter number of Y approximation points  200
Enter expression for joint density  (6/5)*(t.^2 + u)
Use array operations on X, Y, PX, PY, t, u, and P
G = 2*t.^2.*(u>=t) + 3*t.*u.*(u<t);
EZx = sum(G.*P)./sum(P);                          % Approximate
eZx = (-X.^5 + 4*X.^4 + 2*X.^2)./(2*X.^2 + 1);  % Theoretical
plot(X,EZx,'k-',X,eZx,'k-.')
% Plotting details                                % See Figure 14.1.4
```

The theoretical and approximate are barely distinguishable on the plot. Although the same number of approximation points are use as in Figure 14.1.4 (Example 14.1.9), the fact that the entire region is included in the grid means a larger number of effective points in this example.

Given our approach to conditional expectation, the fact that it solves the regression problem is a matter that requires proof using properties of of conditional expectation. An alternate approach is simply to *define* the conditional expectation to be the solution to the regression problem, then determine its properties. This yields, in particular, our defining condition (CE1). Once that is established, properties of expectation (including the uniqueness property (E5)) show the essential equivalence of the two concepts.

There are some technical differences which do not affect most applications. The alternate approach assumes the second moment $E[X^2]$ is finite. Not all random variables have this property. However, those ordinarily used in applications at the level of this treatment will have a variance, hence a finite second moment.

We use the interpretation of $e(X) = E[g(Y)|X]$ as the best mean square estimator of $g(Y)$, given X, to interpret the formal property (CE9). We examine the special form

(CE9a) $E\{E[g(Y)|X]|X, Z\} = E\{E[g(Y)|X, Z]|X\} = E[g(Y)|X]$

Put $e_1(X, Z) = E[g(Y)|X, Z]$, the best mean square estimator of $g(Y)$, given (X, Z). Then (CE9b) can be expressed

$$E[e(X)|X, Z] = e(X) \text{ a.s. and } E[e_1(X, Z)|X] = e(X) \text{ a.s.}$$

In words, if we take the best estimate of $g(Y)$, given X, then take the best mean sqare estimate of that, given (X, Z), we do not change the estimate of $g(Y)$. On the other hand if we first get the best mean sqare estimate of $g(Y)$, given (X, Z), and then take the best mean square estimate of that, given X, we get the best mean square estimate of $g(Y)$, given X.
