Probability and Stochastic Processes I - Lecture 20

Michael Evans
University of Toronto
http://www.utstat.utoronto.ca/mikevans/stac62/STAC622023.html

2023

III.8 Conditional Expectation

- consider r.v. Y where $E(|Y|) < \infty$ and random vector $\mathbf{X}$

- from the joint distribution of $(\mathbf{X}, Y)$ we get the conditional distribution of Y given that $\mathbf{X} = \mathbf{x}$ and now we want the conditional mean of Y given that $\mathbf{X} = \mathbf{x}$

discrete case

- the joint distribution of $(\mathbf{X}, Y) \in R^{k+1}$ is given by the prob. function

$$p_{(\mathbf{X},Y)}(\mathbf{x}, y) = P_{(\mathbf{X},Y)}(\{(\mathbf{x}, y)\}) = P(\mathbf{X} = \mathbf{x}, Y = y)$$

and so the conditional distribution of $Y \,|\, \mathbf{X} = \mathbf{x}$, namely, the probability measure $P_{Y\,|\,\mathbf{X}}$, has prob. function

$$p_{Y\,|\,\mathbf{X}}(y \,|\, \mathbf{x}) = p_{(\mathbf{X},Y)}(\mathbf{x}, y)/p_{\mathbf{X}}(\mathbf{x})$$

when $p_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{X}}(\{\mathbf{x}\}) = P(\mathbf{X} = \mathbf{x}) = \sum_y p_{(\mathbf{X},Y)}(\mathbf{x}, y) > 0$ (otherwise cond. dist. not defined)

- then the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$ is given by

$$E_{p_{Y\,|\,\mathbf{X}}}(Y\,|\,\mathbf{X})(\mathbf{x}) = \sum_{y>0} y p_{Y\,|\,\mathbf{X}}(y\,|\,\mathbf{x}) - \sum_{y<0} y p_{Y\,|\,\mathbf{X}}(y\,|\,\mathbf{x})$$

when defined and note that

$$\begin{aligned}
\sum_{y} |y| p_{Y\,|\,\mathbf{X}}(y\,|\,\mathbf{x}) &= \sum_{y} |y| \frac{p_{(\mathbf{X},Y)}(\mathbf{x},y)}{p_{\mathbf{X}}(\mathbf{x})} \\
&= \frac{1}{p_{\mathbf{X}}(\mathbf{x})} \sum_{y,p_{(\mathbf{X},Y)}(\mathbf{x},y)>0} |y| p_{(\mathbf{X},Y)}(\mathbf{x},y) \\
&\leq \frac{1}{p_{\mathbf{X}}(\mathbf{x})} \sum_{(\mathbf{z},y)} |y| p_{(\mathbf{X},Y)}(\mathbf{z},y) = \frac{1}{p_{\mathbf{X}}(\mathbf{x})} E(|Y|) < \infty
\end{aligned}$$

- so when $E(|Y|) < \infty$, the conditional expectation is also finite

- sometimes we write $E_{p_{Y\,|\,\mathbf{X}}}(Y\,|\,\mathbf{X}=\mathbf{x}) = E_{p_{Y\,|\,\mathbf{X}}}(Y\,|\,\mathbf{X})(\mathbf{x})$

- but we want to think of $E_{p_{Y\,|\,\mathbf{X}}}(Y\,|\,\mathbf{X}) : (R^k, \mathcal{B}^k) \to (R^1, \mathcal{B}^1)$ and then define $E(Y\,|\,\mathbf{X}) : (\Omega, A) \to (R^1, \mathcal{B}^1)$ by

$$E(Y\,|\,\mathbf{X})(\omega) = E_{P_{Y\,|\,\mathbf{X}}}(Y\,|\,\mathbf{X})(\mathbf{X}(\omega))$$

Proposition III.8.1 If $h : (R^k, \mathcal{B}^k) \to (R^1, \mathcal{B}^1)$ is s.t. $E(|Yh(\mathbf{X})|) < \infty$, then

$$E(Yh(\mathbf{X})) = E(h(\mathbf{X})E(Y\,|\,\mathbf{X})).$$

Proof:

$$E(Yh(\mathbf{X})) = \sum_{(\mathbf{x},y)} yh(\mathbf{x})p_{(\mathbf{X},Y)}(\mathbf{x},y) = \sum_{(\mathbf{x},y)} yh(\mathbf{x})p_{\mathbf{X}}(\mathbf{x})\frac{p_{(\mathbf{X},Y)}(\mathbf{x},y)}{p_{\mathbf{X}}(\mathbf{x})}$$

$$= \sum_{(\mathbf{x},y)} yh(\mathbf{x})p_{\mathbf{X}}(\mathbf{x})p_{Y\,|\,\mathbf{X}}(y\,|\,\mathbf{x}) = \sum_{\mathbf{x}} h(\mathbf{x})\left(\sum_{y} yp_{Y\,|\,\mathbf{X}}(y\,|\,\mathbf{x})\right)p_{\mathbf{X}}(\mathbf{x})$$

$$= \sum_{\mathbf{x}} h(\mathbf{x})E_{p_{Y\,|\,\mathbf{X}}}(Y\,|\,\mathbf{X})(\mathbf{x})p_{\mathbf{X}}(\mathbf{x}) = E(h(\mathbf{X})E(Y\,|\,\mathbf{X})). \blacksquare$$

Corollary III.8.2 $E(Yh(\mathbf{X}) \mid \mathbf{X}) = h(\mathbf{X})E(Y \mid \mathbf{X})$

note - $E(Y \mid \mathbf{X})$ has all the properties of E as it is an expectation

Corollary III.8.3 (*Theorem of Total Expectation*) For random vector $(\mathbf{X}, Y)$ such that $E(|Y|) < \infty$,

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

Proof: Put $h(\mathbf{x}) \equiv 1$.

- if $Y = I_A$ for $A \in \mathcal{A}$, then

$$E(Y \mid \mathbf{X})(\mathbf{x}) = \sum y p_{Y \mid \mathbf{X}}(y \mid \mathbf{x}) = 0 p_{Y \mid \mathbf{X}}(0 \mid \mathbf{x}) + 1 p_{Y \mid \mathbf{X}}(1 \mid \mathbf{x}) = P(A \mid \mathbf{X})(\mathbf{x})$$

Corollary III.8.3 (*Theorem of Total Probability*) If $A \in \mathcal{A}$, then

$$P(A) = E(P(A \mid \mathbf{X})).$$

Corollary III.8.4 If also $E(Y^2) < \infty$, then

$$Var(Y) = E(Var(Y \mid \mathbf{X})) + Var(E(Y \mid \mathbf{X})).$$

Proof: We have

$$
\begin{aligned}
& Var(Y) = E((Y - E(Y))^2) \overset{\text{TTE}}{=} E(E((Y - E(Y))^2 \mid \mathbf{X})) \\
= \quad & E(E((Y - E(Y \mid \mathbf{X}) + E(Y \mid \mathbf{X}) - E(Y))^2 \mid \mathbf{X})) \\
& \text{and} \\
& E((Y - E(Y \mid \mathbf{X}) + E(Y \mid \mathbf{X}) - E(Y))^2 \mid \mathbf{X}) \\
= \quad & E((Y - E(Y \mid \mathbf{X}))^2 \mid \mathbf{X}) + \\
& 2E((Y - E(Y \mid \mathbf{X}))(E(Y \mid \mathbf{X}) - E(Y)) \mid \mathbf{X}) + \\
& E((E(Y \mid \mathbf{X}) - E(Y))^2 \mid \mathbf{X}) \\
= \quad & Var(Y \mid \mathbf{X}) + 2(E(Y \mid \mathbf{X}) - E(Y \mid \mathbf{X}))(E(Y \mid \mathbf{X}) - E(Y)) + \\
& (E(Y \mid \mathbf{X}) - E(Y))^2 \\
= \quad & Var(Y \mid \mathbf{X}) + (E(Y \mid \mathbf{X}) - E(Y))^2
\end{aligned}
$$

and applying E to both sides gives the result. ∎

Corollary III.8.5 The random variable $E(Y \mid \mathbf{X})$ is the best predictor of Y from $\mathbf{X}$ in the sense that it minimizes $E((Y - h(\mathbf{X}))^2)$ among all $h : (R^k, \mathcal{B}^k) \to (R^1, \mathcal{B}^1)$ and smallest residual error is $E(Var(Y \mid \mathbf{X}))$.

Proof:

$$\begin{aligned} & E((Y - h(\mathbf{X}))^2) = E((Y - E(Y \mid \mathbf{X}) + E(Y \mid \mathbf{X}) - h(\mathbf{X}))^2) \\ = \; & E((Y - E(Y \mid \mathbf{X}))^2) + \\ & 2E((Y - E(Y \mid \mathbf{X}))(E(Y \mid \mathbf{X}) - h(\mathbf{X}))) + E(E(Y \mid \mathbf{X}) - h(\mathbf{X}))^2) \\ & \text{and} \\ & E((Y - E(Y \mid \mathbf{X}))(E(Y \mid \mathbf{X}) - h(\mathbf{X}))) \\ & \overset{\text{TTE}}{=} E(E((Y - E(Y \mid \mathbf{X}))(E(Y \mid \mathbf{X}) - h(\mathbf{X})) \mid \mathbf{X})) = 0 \end{aligned}$$

and so

$$\begin{aligned} E((Y - h(\mathbf{X}))^2) &= E((Y - E(Y \mid \mathbf{X}))^2 + E(E(Y \mid \mathbf{X}) - h(\mathbf{X}))^2) \\ &\geq E((Y - E(Y \mid \mathbf{X}))^2) = E(Var(Y \mid \mathbf{X})) \end{aligned}$$

with equality when $h(\mathbf{X}) = E(Y \mid \mathbf{X})$. ■

- in general, if r.v. Y satisfies $E(|Y|) < \infty$, then $E(Y \,|\, \mathbf{X})$ is defined as the r.v. $E(Y \,|\, \mathbf{X}) : (\Omega, A) \to (R^1, \mathcal{B}^1)$ that satisfies

$$E(Yh(\mathbf{X})) = E(h(\mathbf{X})E(Y \,|\, \mathbf{X})) \tag{1}$$

for every $h : (R^k, \mathcal{B}^k) \to (R^1, \mathcal{B}^1)$ such that $E(|Yh(\mathbf{X})|) < \infty$

- it can be proven that $E(Y \,|\, \mathbf{X})$ exists and two versions are the same wp1

- this can be generalized to define $E(Y \,|\, \{(t, X_t) : t \in T\})$ the conditional expectation of Y given the process $\{(t, X_t) : t \in T\}$

- all the results proved here also apply to these general contexts

Exercise III.9.1 If $(\mathbf{X}, Y)$ has density $f_{(\mathbf{X},Y)}$ and $E(|Y|) < \infty$, then prove

$$\begin{aligned} E(Y \,|\, \mathbf{X})(\mathbf{x}) &= \int_{-\infty}^{\infty} y f_{Y \,|\, \mathbf{X}}(y \,|\, \mathbf{x})\, dy \text{ where} \\ f_{Y \,|\, \mathbf{X}}(y \,|\, \mathbf{x}) &= \frac{f_{(\mathbf{X},Y)}(\mathbf{x},y)}{f_{\mathbf{X}}(\mathbf{x})} \text{ and } f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty} f_{(\mathbf{X},Y)}(\mathbf{x},y)\, dy. \end{aligned}$$

Hint: use (1).

Example III.9.1 $N_k(\mu, \Sigma)$

- suppose Σ is p.d. and

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} \sim N_k(\mu, \Sigma) \text{ with } \mathbf{Y} \in R^l$$

$$\mu = \begin{pmatrix} \mu_{\mathbf{Y}} \\ \mu_{\mathbf{X}} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{\mathbf{Y}} & \Sigma_{\mathbf{YX}} \\ \Sigma'_{\mathbf{YX}} & \Sigma_{\mathbf{X}} \end{pmatrix}$$

- then

$$\mathbf{Y} \,|\, \mathbf{X} = \mathbf{x} \sim N_k(\mu_{\mathbf{Y}} + \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{X}}^{-1}(\mathbf{x} - \mu_{\mathbf{X}}), \Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{X}}^{-1}\Sigma'_{\mathbf{YX}})$$

so $E(\mathbf{Y} \,|\, \mathbf{X})(\mathbf{x}) = \mu_{\mathbf{Y}} + \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{X}}^{-1}(\mathbf{x} - \mu_{\mathbf{X}})$ and this minimizes

$$\sum_{i=1}^{l} E((Y_i - h_i(\mathbf{X}))^2) = E(||\mathbf{Y} - \mathbf{h}(\mathbf{X})||^2)$$

among all $\mathbf{h} : (R^{k-l}, \mathcal{B}^{k-l}) \to (R^l, \mathcal{B}^l)$ ■

Example III.9.2 *Martingales*

- consider a game of coin tossing where a gambler bets on H which occurs with probability $1/2$, and if the gambler bets $\$x$ the payoff is $\$2x$ so the expected gain on a toss is $0.5(2x - x) - 0.5x = 0$

- the gambler adopts the following strategy: they bet $\$1$ on the first toss, if they lose this bet they bet $\$2$ on the next toss, if they lose this bet they bet $\$4$ on the next toss and generally if they lose the first n bets they bet $\$2^n$ on the next bet and they stop as soon as they win which happens with probability 1

- if the first H occurs at time n then gain is $2^n - (1 + 2 + \cdots + 2^{n-1}) = 2^n - 2^n + 1 = 1$ so this guarantees a profit

- but note that expected loss just before win is

$$\sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n (2^n - 1) = \infty$$

so you need a big bank account if you want to use this strategy

- let X_n denote the gambler's gain (loss) at toss n

- so

$$X_{n+1} = \begin{cases} X_n & \text{if stopped by toss } n \\ X_n + 2^n & \text{if } H \text{ at toss } n \\ X_n - 2^n & \text{if } T \text{ at toss } n \end{cases}$$

- then

$$\begin{aligned} E(X_{n+1} \,|\, X_1, \ldots, X_n)(x_1, \ldots, x_n) &= x_n \text{ and so} \\ E(X_{n+1} \,|\, X_1, \ldots, X_n) &= X_n \end{aligned}$$

- a s. p. $\{(n, X_n) : n \in \mathbb{N}\}$ with this property is called a *martingale*