EE 464

Spring 2003

Lecture Notes Part 11c

Christopher Wayne Walker, Ph.D.

11.4 Conditional Expected Values

Definition: The *conditional mean of* $g(Y)$ *given* $X \le x$ is given by

$$E\left[g(Y)|X \le x\right] = \int_{-\infty}^{\infty} g(y) f(y|X \le x) dy.$$

Definition: The *conditional mean of* $g(Y)$ *given* $X = x$ is given by

$$E\left[g(Y)|X = x\right] = \int_{-\infty}^{\infty} g(y) f(y|x) dy.$$

In particular, we have the conditional mean of Y given $X = x$

$$\mu_{Y|X} = E\left[Y|X = x\right] = \int_{-\infty}^{\infty} y f(y|x) dy$$

and the conditional variance of Y given $X = x$

$$\sigma_{Y|X}^2 = E\left[(Y - \mu_{Y|X})^2 | X = x\right] = \int_{-\infty}^{\infty} (y - \mu_{Y_X})^2 f(y|x) dy.$$

Notation: $E\left[g(Y)|x\right] = E\left[g(Y)|X = x\right].$

Preceding developments lead to the following theorem.

Theorem:

$$E\left[g(X,Y)|M\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y|M) dx dy$$

for an event M.

Special Case: Let $M = \{x < X \le x + \Delta x\}$. Then

$$E\left[g(X,Y)|x < X \le x + \Delta x\right]$$

$$= \int_{-\infty}^{\infty} \int_{x}^{x+\Delta x} g(\alpha, y) f(\alpha, y | x < X \le x + \Delta x) d\alpha dy.$$

Recall,

$$f(x,y|x_1 < X \le x_2) = \frac{f(x,y)}{F_X(x_2) - F_X(x_1)}, \ x_1 < x \le x_2.$$

Let $x_1 = x,\ x_2 = x + \Delta x$. Then

$$f(x, y | x_1 < X \le x + \Delta x) = \frac{f(x, y)}{F_X(x + \Delta x) - F_X(x)}.$$

Therefore,

$$\begin{aligned} &E\left[g(X, Y) | x < X \le x + \Delta x\right] \\ &= \int_{-\infty}^{\infty} \int_{x}^{x+\Delta x} g(\alpha, y) \frac{f(\alpha, y)}{F_X(\alpha + \Delta x) - F_X(\alpha)} d\alpha dy \\ &= \int_{-\infty}^{\infty} \int_{x}^{x+\Delta x} g(\alpha, y) f(\alpha, y) \frac{\frac{1}{\Delta x}}{\frac{F_X(\alpha+\Delta x) - F_X(\alpha)}{\Delta x}} d\alpha dy \\ &\longrightarrow \int_{-\infty}^{\infty} g(x, y) f(x, y) \Delta x \frac{\frac{1}{\Delta x}}{f_X(x)} dy \quad (\text{as } \Delta x \to 0). \end{aligned}$$

Thus,

$$E\left[g(X, Y) | X = x\right] = \int_{-\infty}^{\infty} g(x, y) \frac{f(x, y)}{f_X(x)} dy$$

which becomes

$$E\left[g(X, Y) | X = x\right] = \int_{-\infty}^{\infty} g(x, y) f(y|x) dy.$$

Note that the conditional mean of Y given $X = x$ is itself a function of x:

$$E\left[Y | X = x\right] = \int_{-\infty}^{\infty} y f(y|x) dy.$$

Then $E[Y|X]$ is a random variable and

$$\begin{aligned} E\left[E(Y|X)\right] &= \int_{-\infty}^{\infty} E(Y|X) f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(y|x) f_X(x) dy dx. \end{aligned}$$

But,

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

so

$$E\left[E(Y|X)\right] = \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f(x, y) dx dy = \int_{-\infty}^{\infty} y f_Y(y) dy = E(Y).$$

Similarly,

$$E\left[E\left(g(X,Y)|X\right)\right] = E\left[g(X,Y)\right].$$

11.5 Mean Square Estimation

Recall that the value of b that minimizes $E\left[(X-b)^2\right]$ is $b = E(X)$ (see class notes section 9.4). So if we wish to estimate the value of a random variable Y using only a constant, c, then the mean square error (MSE)

$$e = E\left[(Y-c)^2\right] = \int_{-\infty}^{\infty} (y-c)^2 f_Y(y)dy$$

is minimized if we choose

$$c = E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy.$$

With $c = E(Y)$, our cost function is $E\left[(Y-E(Y))^2\right]$ which is the variance (so we are minimizing the variance in our error).

Nonlinear MS Estimation:

Now consider a possibly nonlinear estimate for Y. Let

$$e = E\left[(Y-c(X))^2\right] = \int_{-\infty}^{\infty} (y-c(x))^2 f(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-c(x))^2 f(y|x) f_X(x)dxdy$$

$$= \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{\infty} (y-c(x))^2 f(y|x)dydx.$$

Now x is a constant in the integral

$$\int_{-\infty}^{\infty} (y-c(x))^2 f(y|x)dy$$

which implies $c(x)$ is a constant in the integral as well. Since $c(x)$ is a constant we can use our prior result to conclude

$$c(x) = E(Y|X=x) = \int_{-\infty}^{\infty} y f(y|x)dy$$

minimizes this integral for any x. Thus, $E(Y|X = x)$ is the best MSE predictor of Y given $X = x$.

Linear MS Estimation:

Sometimes we are willing to not necessarily have the best minimum mean square estimate (or predictor) but instead a predictor that is easier to calculate.

Theorem: Suppose that $E(X^2)$ and $E(Y^2)$ are finite and X and Y are not constant. Then the best (in the MS sense) zero intercept linear predictor of Y ($\hat{Y} = a_0 X$) is obtained by taking

$$a_0 = \frac{E(XY)}{E(X^2)}$$

while the best linear predictor of Y ($\hat{Y} = a_1 X + b_1$) is

$$a_1 = \frac{Cov(X,Y)}{Var(X)}, \quad b_1 = E(Y) - a_1 E(X).$$

Proof:

$$E\left[(Y - aX)^2\right] = E(Y^2) - 2aE(XY) + a^2 E(X^2)$$

$$= E(X^2)\left[a - \frac{E(XY)}{E(X^2)}\right]^2 + \left[E(Y^2) - \frac{[E(XY)]^2}{E(X^2)}\right].$$

Using a we have no control over

$$\left[E(Y^2) - \frac{[E(XY)]^2}{E(X^2)}\right]$$

while

$$E(X^2)\left[a - \frac{E(XY)}{E(X^2)}\right]^2$$

is minimized by taking $a = a_0$. This proves the first part.

Now

$$E\left[(Y - aX - b)^2\right] = E\left[(Y - aX)^2\right] - 2bE(Y - aX) + b^2$$

$$= Var(Y - aX) + [E(Y - aX)]^2 - 2bE(Y - aX) + b^2$$
$$= Var(Y - aX) + [E(Y)]^2 - 2aE(X)E(Y) + a^2 [E(X)]^2$$
$$-2bE(Y) + 2abE(X) + b^2$$
$$= Var(Y - aX) + [E(Y) - aE(X) - b]^2 .$$

Now given any value of a, $[E(Y) - aE(X) - b]^2$ is minimized by taking $b = E(Y) - aE(X) = b_1$. Using this value of b we seek to minimize

$$E\left[(Y - aX - b)^2\right] = E\left[(Y - aX - (E(Y) - aE(X)))^2\right]$$
$$= E\left[(Y - E(Y) - a(X - E(X)))^2\right].$$

Let $Y_0 = Y - E(Y)$, $X_0 = X - E(X)$. Then we want to minimize $[(Y_0 - aX_0)^2]$. From first part of theorem we know

$$a = \frac{E(X_0Y_0)}{E(X_0^2)} = \frac{E\left[(X - E(X))(Y - E(Y))\right]}{E\left[(X - E(X))^2\right]} = \frac{Cov(X, Y)}{Var(X)} = a_1.$$

Thus, $b = b_1 = E(Y) - a_1E(X)$ and $\hat{Y} = a_1X + b_1$ is the best linear mean square error predictor or estimator of Y.

Example: Suppose Z_1 is bernoulli with $E(Z_1) = p$ and $Var(Z_1) = pq$ where $q = 1 - p$. Also, let Z_2 be bernoulli with $E(Z_2) = p$ and $Var(Z_2) = pq$. Assume Z_1 is independent of Z_2. Let $X = Z_1$ and let $Y = Z_1Z_2$. Then

a. $E(Y|X = x) = E(Z_1Z_2|Z_1 = x] = E(xZ_2) = px$.

b. $E[E(Y|X)] = E[pX] = p^2 = E(Y)$.

c. $Var[E(Y|X)] = Var(pX) = p^2Var(X) = p^3q$.

d. $Var(Y|X = x) = Var(Z_1Z_2|Z_1 = x) = Var(xZ_2) = x^2pq$.

e. $E[Var(Y|X)] = E(X^2pq) = pqE(X^2) = pqE(Z_1^2) = pq\left[Var(Z_1) + [E(Z_1)]^2\right] = pq(pq + p^2) = p^2q(q + p) = p^2q$.

f. Best MSE predictor of Y is $E(Y|X) = pX \Rightarrow$best MSE predictor of Y given $X = x$ is px.

g. Best linear MSE predictor of Y is $\hat{Y} = a_1 X + b_1$ where

$$a_1 = \frac{Cov(X,Y)}{Var(X)}, \quad b_1 = E(Y) - a_1 E(X).$$

Now

$$a_1 = \frac{E(XY) - E(X)E(Y)}{Var(X)} = \frac{E(Z_1^2 Z_2) - E(Z_1)E(Z_1 Z_2)}{Var(Z_1)}$$

$$= \frac{E(Z_1^2)E(Z_2) - E(Z_1)E(Z_1)E(Z_2)}{Var(Z_1)}$$

$$= \frac{(pq + p^2)p - p^3}{pq} = \frac{pq + p^2 - p^2}{q} = p.$$

So $b_1 = E(Z_1 Z_2) - a_1 E(Z_1) = E(Z_1)E(Z_2) - a_1 E(Z_1) = p^2 - a_1 p = p^2 - p^2 = 0$. Thus, $\hat{Y} = pX$.

Therefore the best MSE predictor of Y given X is also the best linear MSE predictor in this case (as expected since the best MSE predictor was itself linear).

Orthogonality Principle

Consider

$$e = E\left[(Y - (aX + b))^2\right]$$

where $aX + b$ is a linear estimate of Y given the observed data X. This is minimal where

$$\frac{\partial e}{\partial a} = 0 \text{ and } \frac{\partial e}{\partial b} = 0.$$

Thus

$$\frac{\partial e}{\partial b} = E\left[2(Y - (aX + b))\right] = 0 \Rightarrow E(Y) = aE(X) + b.$$

Also

$$\frac{\partial e}{\partial a} = E\left[2(Y - (aX + b))(-X)\right] = 0 \Rightarrow E\left[(Y - (aX + b))X\right] = 0.$$

This implies the estimation error $(Y - (aX + b))$ is orthogonal to the data. This is called the *orthogonality principle*.

Special case: If $b = 0$ we have $e = E\left[(Y - aX)^2\right]$ and $E\left[(Y - aX)X\right] = 0$ by the orthogonality principle, Thus,

$$E(XY) - aE(X^2) = 0 \Rightarrow a = \frac{E(XY)}{E(X^2)}$$

which is the same as we got for the best zero intercept linear predictor of Y given X.