EE 503

Lecture Notes Part 11

Christopher Wayne Walker, Ph.D.

11.0 Moments and Conditional Distributions

11.1 Joint Moments

Given random variables X and Y, let $Z = g(X, Y)$. The expected value of Z is given by

$$E(Z) = \int_{-\infty}^{\infty} z f_Z(z) dz$$

as usual.

Theorem:

$$E(Z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

Proof: Omitted.

Note that if Z is only a function of X, i.e., $Z = g(X)$, then

$$\begin{aligned} E(Z) = E(g(X)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} g(x) \int_{-\infty}^{\infty} f(x, y) dy dx \\ &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \end{aligned}$$

as expected.

If X and Y are discrete, then

$$E(Z) = E(g(X, Y)) = \sum_{i,k} g(x_i, y_k) p_{ik}.$$

Linearity: As in the 1-dim case, the 2-dim expectation operator is linear. In particular,

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

We would like to measure in some meaningful way the degree of association between random variables X and Y. The covariance helps us do this.

Definition: The *covariance* of two random variables X and Y is

$$C_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right],$$

where $\mu_x = E(X),\ \mu_Y = E(Y)$.

Note:

$$C_{XY} = E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_y$$

or

$$C_{XY} = E(XY) - E(X)E(Y).$$

One can write C for C_{XY} when clear to do so.

In order to compare the degree of association of different pairs of random variables it is convenient to normalize the covariance measure to produce the correlation coefficient.

Definition: The *correlation coefficient*, r_{XY}, of random variables X and Y is

$$r_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y},\ \sigma_X \sigma_y \neq 0.$$

Note: We define $C_{XY} = 0$ if $\sigma_X = 0$ or $\sigma_Y = 0$.

One can write r for r_{XY} when clear.

Note: Some authors use ρ_{XY} instead of r_{XY}.

Claim: $|r_{XY}| \leq 1$ (i.e., $|C_{XY}| \leq \sigma_X \sigma_Y$).

Proof: Consider

$$q(t) = E\left[(t(X - \mu_X) + (Y - \mu_Y))^2\right].$$

Then $q(t) \geq 0$. Expanding we get

$$q(t) = E\left[t^2(X - \mu_X)^2 + 2t(X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2\right] \geq 0$$

or

$$q(t) = t^2\sigma_X^2 + 2tC_{XY} + \sigma_Y^2 \geq 0.$$

This is a quadratic in t. Setting $q(t) = 0$ we find

$$t = \frac{-2C_{XY} \pm \sqrt{4C_{XY}^2 - 4\sigma_X^2\sigma_Y^2}}{2\sigma_X^2}.$$

Since $q(t) \geq 0$, $q(t)$ cannot have two distinct real roots. Thus, the discriminant is less than or equal to zero. Hence,

$$4C_{XY}^2 - 4\sigma_X^2\sigma_Y^2 \leq 0 \Rightarrow |C_{XY}| \leq |\sigma_X\sigma_Y| \Rightarrow |C_{XY}| \leq \sigma_X\sigma_Y$$

$$\Rightarrow -\sigma_X\sigma_y \leq C_{XY} \leq \sigma_X\sigma_y \Rightarrow -1 \leq r_{XY} \leq 1 \Rightarrow |r_{XY}| \leq 1.$$

If $r_{XY}^2 = 1$, i.e., $r_{XY} = \pm 1$, we can establish a functional relationship between X and Y. We will now show this. We need the following theorem.

Theorem: Suppose $\sigma_X^2 = 0$. Then $P(X = \mu_X) = 1$ (we say $X = \mu_X$ with probability 1).

Proof: Recall Tchebycheff's inequality

$$P\left(|X - \mu_X| \geq \epsilon\right) \leq \frac{\sigma_X^2}{\epsilon^2}, \quad \forall\ \epsilon > 0.$$

So,

$$P\left(|X - \mu_X| \geq \epsilon\right) = 0 \Rightarrow P\left(|X - \mu_X| < \epsilon\right) = 1$$

$$\Rightarrow P(-\epsilon < X - \mu_X < \epsilon) = 1.$$

Let $\epsilon \to 0$ to get $P(X - \mu_X = 0) = 1$ or $P(X = \mu_X) = 1$.

Theorem: $r_{XY}^2 = 1$ if and only if $Y = aX + b$ for some constants a, b.

Proof:
"$\Longrightarrow$" (Necessity or "only if" part).

Assume $r_{XY}^2 = 1$. Recall,

$$q(t) = E\left[\left(t(X - \mu_X) + (Y - \mu_Y)\right)^2\right]$$

or

$$q(t) = t^2\sigma_X^2 + 2tC_{XY} + \sigma_Y^2 \geq 0.$$

If $q(t) > 0$, then the discriminant

$$4C_{XY}^2 - 4\sigma_X^2\sigma_Y^2 < 0 \Rightarrow \frac{C_{XY}^2}{\sigma_X^2\sigma_Y^2} < 1 \Rightarrow r_{XY}^2 < 1$$

contrary to the assumption $r_{XY}^2 = 1$. So if $r_{XY}^2 = 1$ there exists some $t_0 \in \mathbf{R}$ such that $q(t_0) = 0$. Thus,

$$q(t_0) = E\left[(t_0(X - \mu_X) + (Y - \mu_Y))^2\right] = 0.$$

Note that

$$E\left[t_0(X - \mu_X) + (Y - \mu_Y)\right] = 0.$$

So

$$E\left[(t_0(X - \mu_X) + (Y - \mu_Y))^2\right] = Var\left(t_0(X - \mu_X) + (Y - \mu_Y)\right) = 0.$$

The last theorem implies

$$P\left(t_0(X - \mu_X) + (Y - \mu_Y) = 0\right) = 1$$

which imples

$$t_0(X - \mu_X) + (Y - \mu_Y) = 0 \text{ with probability 1.}$$

We suppress the expression "with probability 1" in equations. Thus,

$$Y = -t_0 X + t_0\mu_X + \mu_Y \Rightarrow Y = ax + b$$

with $a = -t_0$ and $b = t_0\mu_X + \mu_Y$.

"$\Longleftarrow$" (Sufficiency or "if" part).

Assume $Y = aX + b$. Then,

$$E(Y) = aE(X) + b$$

and

$$Var(Y) = a^2 Var(X).$$

Also,

$$E(XY) = E\left[X(aX + b)\right] = aE(X^2) + bE(X).$$

Thus,

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

$$
\begin{aligned}
&= \frac{\left[aE(X^2) + bE(X) - E(X)(aE(X) + b)\right]^2}{Var(X)a^2Var(X)} \\
&= \frac{\left[aE(X^2) + bE(X) - a(E(X))^2 - bE(X))\right]^2}{a^2(Var(X))^2} \\
&= \frac{a^2\left[E(X^2) - (E(X))^2)\right]^2}{a^2(Var(X))^2} = 1.
\end{aligned}
$$

Note that if $a > 0$, $r_{XY} = 1$ and if $a < 0$, $r_{XY} = -1$.

Observe that the correlation coefficient is a measure of the degree of linearity between X and Y: $|r_{XY}| \approx 1$ indicates a high degree of linearity while $|r_{XY}| \approx 0$ indicates a lack of linearity. Positive values of r_{XY} means Y tends to increase with increasing X (and vice versa) while negative values of r_{XY} means that Y tends to decrease with increasing X (and vice versa).

Caution: $r_{XY} = 0$ does not mean there is no relationship between X and Y, it only means there is no linear relationship. There could still be a nonlinear relationship.

Example: Suppose $X \sim N(0, 1)$. Let $Y = X^2$. Here Y is very much related to X, in fact, Y is a (nonlinear) function of X. Now

$$r_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y}$$

but

$$C_{XY} = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) = 0$$

since $E(X^3) = E(X) = 0$. Thus $r_{XY} = 0$.

Definition: Two random variables X and Y are called *uncorrelated* if their covariance is zero, i.e., $C_{XY} = 0$.

Definition: Two random variables X and Y are called *orthogonal* if $E(XY) = 0$.

Notation: $X \perp Y$ means X and Y are orthogonal.

Note: If X and Y are uncorrelated then $(X - \mu_X) \perp (Y - \mu_Y)$.

Theorem: If X and Y are independent then they are uncorrelated.

Proof: We will show $E(XY) = E(X)E(Y)$ which proves the theorem.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy = E(X)E(Y). \end{aligned}$$

Note: If X and Y are independent then $g(X)$ and $h(Y)$ are also independent and $E\left[g(X)h(Y)\right] = E\left[g(X)\right]E\left[h(Y)\right]$ (so $g(X)$ and $h(Y)$ are also uncorrelated) but if X and Y are just uncorrelated it does not follow that $g(X)$ and $h(Y)$ are necessarily uncorrelated. Also, random variables can be uncorrelated without being independent, but in the normal case we that independence implies uncorrelated and uncorrelated implies independence.

Variance of $X + Y$

Consider $Z = X + Y$. Then $E[Z] = \mu_Z = \mu_X + \mu_Y$.

$$\begin{aligned} Var(Z) &= \sigma_Z^2 = E\left[(Z-\mu_Z)^2\right] = E\left[\left((X-\mu_X)+(Y-\mu_Y)\right)^2\right] \\ &= E\left[(X-\mu_X)^2\right] + 2E\left[(X-\mu_X)(Y-\mu_Y)\right] + E\left[(Y-\mu_Y)^2\right] \end{aligned}$$

or

$$\sigma_Z^2 = \sigma_X^2 + 2r_{XY}\sigma_X\sigma_Y + \sigma_Y^2.$$

If X and Y are uncorrelated then $r_{XY} = 0$ and

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2.$$

It then follows that if X and Y are independent

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2.$$

Moments

Definition: A *joint moment* of the random variables X and Y of order $k + r = n$ is

$$m_{kr} = E\left(X^k Y^r\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^r f(x,y)dxdy.$$

Note: The first order moments are

$$m_{10} = E(X)$$

$$m_{01} = E(Y)$$

and the second order moments are

$$m_{20} = E(X^2)$$

$$m_{11} = E(XY)$$

$$m_{02} = E(Y^2).$$

Definition: The *joint central moments* of the random variables X and Y are the moments of $(X - \mu_X)$ and $(Y - \mu_Y)$, i.e.,

$$\mu_{kr} = E\left[(X-\mu_X)^k (Y-\mu_Y)^r\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)^k (y-\mu_Y)^r f(x,y)dxdy.$$

Here

$$\mu_{10} = 0,\ \mu_{01} = 0,\ \mu_{11} = C_{XY},\ \mu_{20} = \sigma_X^2,\ \mu_{02} = \sigma_Y^2.$$

In general, the joint density of X and Y is required in order to determine the joint statistics. However, often for many applications it is sufficient to use only the first- and second-order moments which can be easily estimated numerically given sampled data.

11.2 Joint Characteristic Functions

Definition: The *joint characteristic function* of the random variables X and Y is

$$\phi_{XY}(\omega_1, \omega_2) = E\left[e^{i(\omega_1 X + \omega_2 Y)}\right].$$

So,

$$\phi_{XY} : \mathbf{R}^2 \to \mathbf{C}^2.$$

For X, Y discrete we have

$$\phi_{XY}(\omega_1, \omega_2) = \sum_{k,l} e^{i(\omega_1 x_k + \omega_2 y_l)} P(X = x_k, Y = y_l).$$

For X, Y continuous we have

$$\phi_{XY}(\omega_1, \omega_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{i(\omega_1 x + \omega_2 y)} dx dy.$$

Using the inversion formula for the 2-dimensional Fourier transform (with a sign change) we get

$$f(x, y) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_{XY}(\omega_1, \omega_2) e^{-i(\omega_1 x + \omega_2 y)} d\omega_1 d\omega_2.$$

Definitions: The *marginal characteristic functions* of the random variables X and Y are

$$\phi_X(\omega) = E\left(e^{i\omega X}\right) \text{ and } \phi_Y(\omega) = E\left(e^{i\omega Y}\right).$$

Note that $\phi_X(\omega) = \phi_{XY}(\omega, 0)$ and $\phi_Y(\omega) = \phi_{XY}(0, \omega)$.

Claim: X and Y are independent if and only if

$$\phi_{XY}(\omega_1, \omega_2) = \phi_X(\omega_1)\phi_Y(\omega_2).$$

Proof:
"$\Longrightarrow$" (Necessity or "only if" part).

Assume that X and Y are independent. Then

$$\begin{aligned}
\phi_{XY}(\omega_1, \omega_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{i(\omega_1 x + \omega_2 y)} dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x) f_Y(y) e^{i\omega_1 x} e^{i\omega_2 y} dx dy \\
&= \int_{-\infty}^{\infty} f_X(x) e^{i\omega_1 x} dx \int_{-\infty}^{\infty} f_Y(y) e^{i\omega_2 y} dx \\
&= E\left[e^{i\omega_1 x}\right] E\left[e^{i\omega_2 y}\right] = \phi_X(\omega_1)\phi_Y(\omega_2).
\end{aligned}$$

"$\Longleftarrow$" (Sufficiency or "if" part).

Assume $\phi_{XY}(\omega_1, \omega_2) = \phi_X(\omega_1)\phi_Y(\omega_2)$. Then

$$\begin{aligned} f(x,y) &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi_{XY}(\omega_1,\omega_2)e^{-i(\omega_1 x+\omega_2 y)}d\omega_1 d\omega_2 \\ &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi_X(\omega_1)\phi_Y(\omega_2)e^{-i\omega_1 x}e^{-i\omega_2 y}d\omega_1 d\omega_2 \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_X(\omega_1)e^{-i\omega_1 x}d\omega_1 \cdot \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_Y(\omega_2)e^{-i\omega_2 y}d\omega_2 \\ &= f_X(x)f_Y(y) \Rightarrow \ X \text{ and } Y \text{ are independent.} \end{aligned}$$

Theorem: If the random variables X and Y are independent and $Z = X+Y$ then

$$\phi_Z(\omega) = \phi_X(\omega)\phi_Y(\omega).$$

Proof:

$$\begin{aligned} \phi_Z(\omega) &= E\left[e^{i\omega Z}\right] = E\left[e^{i\omega(X+Y)}\right] = E\left[e^{i\omega X}e^{i\omega Y}\right] \\ &= E\left[e^{i\omega X}\right]E\left[e^{i\omega Y}\right] = \phi_X(\omega)\phi_Y(\omega). \end{aligned}$$

Recall for X and Y independent that

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx \quad \text{(convolution)}.$$

Theorem: Let $X \sim N(\mu_X, \sigma_X^2),\ Y \sim N(\mu_Y, \sigma_Y^2)$ and let X and Y be independent. Then $Z = X + Y$ is also normal and

$$E(Z) = \mu_X + \mu_Y, \quad Var(Z) = Var(X) + Var(Y).$$

Proof:

$$\phi_X(\omega) = E\left(e^{i\omega X}\right) = \frac{1}{\sqrt{2\pi}\sigma_X}\int_{-\infty}^{\infty} e^{i\omega x}e^{-(x-\mu_X)^2/2\sigma_X^2}dx.$$

Recall

$$M_X(s) = E\left(e^{sX}\right) = e^{s\mu_X + \frac{1}{2}\sigma_X^2 s^2}.$$

So

$$M_X(i\omega) = E\left(e^{i\omega X}\right) = \phi_X(\omega)$$

when $M_X(s)$ exists (recall $\phi_X(\omega)$ always exists). Thus,

$$\phi_X(\omega) = e^{i\mu_X\omega - \frac{1}{2}\sigma_X^2\omega^2}.$$

Similarly,

$$\phi_Y(\omega) = e^{i\mu_Y\omega - \frac{1}{2}\sigma_Y^2\omega^2}.$$

Therefore,

$$\phi_Z(\omega) = \phi_X(\omega)\phi_Y(\omega) = e^{i(\mu_X+\mu_Y)\omega - \frac{1}{2}(\sigma_X^2+\sigma_Y^2)\omega^2}.$$

Hence,

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

More generally, if we have the above conditions and $Z = aX + bY$, then

$$\phi_{aX}(\omega) = e^{ia\mu_X\omega - \frac{1}{2}a^2\sigma_X^2\omega^2} = \phi_X(a\omega),$$

$$\phi_{bX}(\omega) = e^{ib\mu_X\omega - \frac{1}{2}b^2\sigma_X^2\omega^2} = \phi_X(b\omega),$$

and

$$\phi_Z(\omega) = \phi_{aX}(\omega)\phi_{bY}(\omega).$$

Thus,

$$Z \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2).$$

11.3 Conditional Distributions

Consider

$$F_Y(y|X \le x) = P(Y \le y|X \le x) = \frac{P(X \le x,\ Y \le y)}{P(X \le x)} = \frac{F_{XY}(x,y)}{F_X(x)}.$$

Thus,

$$f_Y(y|X \le x) = \frac{d}{dy}F_Y(y|X \le x) = \frac{\frac{\partial F_{XY}(x,y)}{\partial y}}{F_X(x)}.$$

Now

$$F_{XY}(x,y|x_1 < X \le x_2) = P(X \le x,\ Y \le y\ |x_1 < X \le x_2)$$

$$= \frac{P(X \le x,\ Y \le y,\ x_1 < X \le x_2)}{P(x_1 < X \le x_2)}. \quad (\star)$$

If $x > x_2$,

$$(\star) = \frac{P(x_1 < X \le x_2,\ Y \le y)}{P(x_1 < X \le x_2)} = \frac{F_{XY}(x_2,y) - F_{XY}(x_1,y)}{F_X(x_2) - F_X(x_1)}.$$

If $x_1 < x \leq x_2)$,

$$(\star) = \frac{P(x_1 < X \leq x,\ Y \leq y)}{P(x_1 < X \leq x_2)} = \frac{F_{XY}(x,y) - F_{XY}(x_1,y)}{F_X(x_2) - F_X(x_1)}.$$

If $x < x_1$,

$$(\star) = 0.$$

Now

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

so

$$f_{XY}(x,y|x_1 < X \leq x_2) = \begin{cases} \dfrac{f_{XY}(x,y)}{F_X(x_2) - F_X(x_1)}, & x_1 < x \leq x_2, \\ 0, & \text{elsewhere.} \end{cases}$$

Consider next

$$\begin{aligned} F_Y(y|x_1 < X \leq x_2) &= P(Y \leq y|x_1 < X \leq x_2) \\ &= \frac{P(x_1 < X \leq x_2,\ Y \leq y)}{P(x_1 < X \leq x_2)} \\ &= \frac{F_{XY}(x_2,y) - F_{XY}(x_1,y)}{F_X(x_2) - F_X(x_1)}. \end{aligned}$$

Thus

$$f_Y(y|x_1 < X \leq x_2) = \frac{\partial}{\partial y} \frac{\int_{-\infty}^{x_2} \int_{-\infty}^{y} f_{XY}(x,\theta) d\theta dx - \int_{-\infty}^{x_1} \int_{-\infty}^{y} f_{XY}(x,\theta) d\theta dx}{F_X(x_2) - F_X(x_1)}$$

$$= \frac{\int_{x_1}^{x_2} f_{XY}(x,y) dx}{F_X(x_2) - F_X(x_1)}.$$

Now let $x_1 = x,\ x_2 = x + \Delta x$. Then

$$f_Y(y|x < X \leq x + \Delta x) = \frac{\int_{x}^{x+\Delta x} f_{XY}(\alpha,y) d\alpha}{F_X(x + \Delta x) - F_X(x)}.$$

Now let $\Delta x \to 0$ to get

$$\frac{\int_{x}^{x+\Delta x} f_{XY}(\alpha,y) d\alpha}{F_X(x + \Delta x) - F_X(x)} \longrightarrow \frac{f_{XY}(x,y)\Delta x}{f_X(x)\Delta x}.$$

Thus,

$$f_Y(y|X=x) = \lim_{\Delta x \to 0} f_Y(y|x < X \leq x + \Delta x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

Notation:

$$f(y|x) = f_Y(y|x) = f_Y(y|X=x), \quad f(x|y) = f_X(x|y) = f_X(x|Y=y),$$

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

Compare to

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{for events } A \text{ and } B,\ P(B) \neq 0.$$

Notation: If X and Y are independent

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

and similarly $f_X(x|y) = f_X(x)$.

Now

$$f_X(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_Y(y|x)f_X(x)}{f_Y(y)}.$$

Also

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx$$

and

$$f_{XY}(x,y) = f_Y(y|x)f_X(x).$$

Hence,

$$f_Y(y) = \int_{-\infty}^{\infty} f_Y(y|x)f_X(x)dx.$$

This is total probability. We thus get Bayes' theorem for densities

$$f_X(x|y) = \frac{f_Y(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_Y(y|x)f_X(x)dx}.$$

Discrete Type: Let $p_i = P(X = x_i),\ p_{ik} = P(X = x_i,\ Y = y_k)$. Then.

$$P(Y = y_k|X = x_i) = \frac{P(X = x_i,\ Y = y_k)}{P(X = x_i)} = \frac{p_{ik}}{p_i}.$$

11.4 Conditional Expected Values

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

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

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 $X = x$

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

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

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

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

Preceding developments lead to the following theorem.

Theorem:

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

for an event M.

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

$$\begin{aligned} &E\left[g(X,Y)|x < X \leq x + \Delta x\right] \\ &\quad = \int_{-\infty}^{\infty} \int_{x}^{x+\Delta x} g(\alpha, y) f(\alpha, y|x < X \leq x + \Delta x) d\alpha dy. \end{aligned}$$

Recall,

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

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

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

Therefore,

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

$$\longrightarrow \int_{-\infty}^{\infty} g(x, y) f(x, y) \Delta x \frac{\frac{1}{\Delta x}}{f_X(x)} dy \quad (\text{as } \Delta x \to 0).$$

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

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

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

$$\begin{aligned} e &= E\left[(Y-c(X))^2\right] = \int_{-\infty}^{\infty} (y-c(x))^2 f(x,y) dx dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-c(x))^2 f(y|x) f_X(x) dx dy \\ &= \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{\infty} (y-c(x))^2 f(y|x) dy dx. \end{aligned}$$

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:

$$\begin{aligned} E\left[(Y-aX)^2\right] &= E(Y^2) - 2aE(XY) + a^2 E(X^2) \\ &= E(X^2)\left[a - \frac{E(XY)}{E(X^2)}\right]^2 + \left[E(Y^2) - \frac{[E(XY)]^2}{E(X^2)}\right]. \end{aligned}$$

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

$$\begin{aligned} E\left[(Y - aX - b)^2\right] &= E\left[(Y - aX)^2\right] - 2bE(Y - aX) + b^2 \\ &= Var(Y - aX) + [E(Y - aX)]^2 - 2bE(Y - aX) + b^2 \\ &= Var(Y - aX) + [E(Y)]^2 - 2aE(X)E(Y) + a^2[E(X)]^2 \\ &\quad -2bE(Y) + 2abE(X) + b^2 \\ &= Var(Y - aX) + [E(Y) - aE(X) - b]^2. \end{aligned}$$

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

$$\begin{aligned} E\left[(Y - aX - b)^2\right] &= E\left[(Y - aX - (E(Y) - aE(X)))^2\right] \\ &= E\left[(Y - E(Y) - a(X - E(X)))^2\right]. \end{aligned}$$

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

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

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 $\hat{Y} = a_1X + b_1$ where

$$a_1 = \frac{Cov(X,Y)}{Var(X)}, \quad b_1 = E(Y) - a_1E(X).$$

Now

$$a_1 = \frac{E(XY) - E(X)E(Y)}{Var(X)} = \frac{E(Z_1^2Z_2) - E(Z_1)E(Z_1Z_2)}{Var(Z_1)}$$

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

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

So $b_1 = E(Z_1Z_2) - a_1E(Z_1) = E(Z_1)E(Z_2) - a_1E(Z_1) = p^2 - a_1p = p^2 - p^2 = 0$. 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.