

4.3 Expectations and conditional expectations

Definition. Consider (X, Y) with joint pdf/pmf $f_{X,Y}$. Suppose that the integrals and sums given below exist.

$$(i) \quad E g(X, Y) = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy & \text{(continuous),} \\ \sum_x \sum_y g(x, y) f_{X,Y}(x, y) & \text{(discrete).} \end{cases}$$

$$(ii) \quad E\{g(X, Y) | Y = y\} = \begin{cases} \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) dx & \text{(continuous),} \\ \sum_x g(x, y) f_{X|Y}(x|y) & \text{(discrete).} \end{cases}$$

Definition. The function $m(x) = E(Y|X = x)$ is called the **regression** of Y on X , or simply “regression function”.

Examples: $m(x) = ax + b$ (linear regression),
 $m(x) = 3x^2 + e^{\theta x}$ (nonlinear regression),
 $m(x)$ not specified / “smooth” (nonparametric regression).

Theorem. Suppose X, Y independent. Then, provided the expectations exist, $E\{g(X)|Y = y\} = E g(X)$, $E\{h(Y)|X = x\} = E h(Y)$. In particular, $E(Y|X = x) = EY = \text{constant} (= m(x))$, $E(X|Y = y) = EX = \text{constant}$.

Proof. Use the definition and $f_{X|Y} = f_X$.

Theorem (iterated expectation). Suppose Y has pdf f_Y . For any function $g(X, Y)$ for which $E g(X, Y)$ exists, $E\{g(X, Y)\} = \int_{-\infty}^{\infty} E\{g(X, Y)|Y = y\} f_Y(y) dy$ (analogously for discrete distributions). Write

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

Proof. Use $f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y)$ for all x, y . (Note that $f_Y(y) = 0$ implies $f_{X,Y}(x, y) = 0$.)

$$\begin{aligned} E g(X, Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \underbrace{\int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) dx}_{E\{g(X, Y)|Y=y\}} f_Y(y) dy. \quad \square \end{aligned}$$

Example (cont., marked Poisson): $Y \sim \text{Poisson}(\lambda)$, $X|Y = y \sim \text{binomial}(y, p)$,

$$EX = \sum_y E(X|Y = y) f_Y(y) = \sum_y yp f_Y(y) = p \sum_y y f_Y(y) = p \cdot EY = p\lambda.$$

Even faster: write $E(X|Y) = Yp$, then $EX = E\{E(X|Y)\} = E(Yp) = \lambda p$.

Similarly, $EX^2 = E\{E(X^2|Y)\} = E\{Yp(1-p) + (Yp)^2\} = p(1-p)\lambda + p^2 EY^2 = p(1-p)\lambda + p^2(\lambda + \lambda^2) = p\lambda - p^2\lambda + p^2\lambda + (p\lambda)^2 = p\lambda + (EX)^2$. Therefore,

$$\text{Var}(X) = EX^2 - (EX)^2 = p\lambda \quad (\text{Poisson}(\lambda p) \text{ variance}).$$

Theorem.

- (i) Suppose $E g(X, Y)$ exists. Then $E\{g(X, Y)|Y = y\} = E\{g(X, y)|Y = y\}$.
- (ii) $E\{g(X)h(Y)\} = E[h(Y)E\{g(X)|Y\}]$.

Proof. (i) $E\{g(X, Y)|Y = y\} = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) dx = E\{g(X, y)|Y = y\}$.
(ii) $E\{g(X)h(Y)\} = E[E\{g(X)h(Y)|Y\}] = \int_{-\infty}^{\infty} E\{g(X)h(Y)|Y = y\} f_Y(y) dy$
 $= \int_{-\infty}^{\infty} E\{g(X)h(y)|Y = y\} f_Y(y) dy = \int_{-\infty}^{\infty} h(y) E\{g(X)|Y = y\} f_Y(y) dy$
 $= E[h(Y)E\{g(X)|Y\}].$ \square

Remark. (ii) implies for *independent* r.v.'s X, Y :

$$E\{g(X)h(Y)\} = E[g(X) \underbrace{E\{h(Y)|X\}}_{=Eh(Y) \text{ const.}}] = Eh(Y) \cdot Eg(X).$$

In particular, X, Y independent $\Rightarrow E(XY) = EX \cdot EY$.

Example (linear regression). Let $Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2)$ (or any other distribution with this mean and variance). If $X \sim N(\mu_x, \sigma_x^2)$ then

$$\mu_y = EY = E\{E(Y|X)\} = E(\beta_0 + \beta_1 X) = \beta_0 + \beta_1 \mu_x,$$

$$EY^2 = E\{E(Y^2|X)\} = E\{Var(Y|X) + E^2(Y|X)\} = E\{\sigma^2 + (\beta_0 + \beta_1 X)^2\}$$

$$= \sigma^2 + \beta_1^2 \sigma_x^2 + (\beta_0 + \beta_1 \mu_x)^2,$$

$$Var Y = EY^2 - E^2 Y = EY^2 - (\beta_0 + \beta_1 \mu_x)^2 = \sigma^2 + \beta_1^2 \sigma_x^2.$$

Example (insurance claims). Consider $N =$ "number of claims" \sim Poisson(λ), and suppose that the claim values are independent r.v.'s from an $\text{expo}(\beta)$ (= gamma(1, β)) distribution. Let $Y = \sum$ claims.

If $N = n$ then $Y|N = n \sim$ gamma(n, β) and

$$EY = E\{E(Y|N)\} = E(N\beta) = \beta EN = \beta \lambda,$$

$$Var Y = EY^2 - E^2 Y = E\{E(Y^2|N)\} - E^2 Y = E\{Var(Y|N) + E^2(Y|N)\} - E^2 Y$$

$$= E\{N\beta^2 + (N\beta)^2\} - \lambda^2 \beta^2 = \beta^2(\lambda + \lambda + \lambda^2 - \lambda^2) = 2\lambda\beta^2.$$

Definition. The conditional variance of Y given $X = x$ is the variance of the conditional distribution of Y given $X = x$, if it exists. Write $Var(Y|X = x)$, and $Var(Y|X)$ for the random variable.

Theorem. $Var Y = \underbrace{E\{Var(Y|X)\}}_A + \underbrace{Var\{E(Y|X)\}}_B$.

Proof. We have $A + B = EY^2 - E^2 Y (= Var Y)$ since

$$A = E\{Var(Y|X)\} = E\{E(Y^2|X) - E^2(Y|X)\} = EY^2 - E\{E^2(Y|X)\},$$

$$B = Var Z \text{ with } Z = E(Y|X), \text{ i.e. } Var Z = EZ^2 - E^2 Z = E\{E^2(Y|X)\} - E^2 Y.$$

\square

Example (cont.) $N \sim$ Poisson(λ), $Y|N \sim$ gamma(N, β).

$$Var\{E(Y|N)\} = \beta^2 Var N = \beta^2 \lambda;$$

we know that $Var(Y|N) = N\beta^2$, therefore $E\{Var(Y|N)\} = E(N\beta^2) = \lambda\beta^2$.

In conclusion, $Var Y = E\{Var(Y|N)\} + Var\{E(Y|N)\} = 2\lambda\beta^2$.

Example (linear regression, cont.).

Assume $E(Y|X) = \beta_0 + \beta_1 X$, $Var(Y|X) = \sigma^2$ (or $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$).

$$Var Y = Var\{E(Y|X)\} + E\{Var(Y|X)\} = Var(\beta_0 + \beta_1 X) + E\sigma^2$$

$$= \beta_1^2 Var X + \sigma^2 = \beta_1^2 \sigma_x^2 + \sigma^2.$$

Theorem (best predictor).

Assume X, Y jointly distributed with finite second moments. The function of X that best predicts Y , in the sense of minimizing $E\{Y - g(X)\}^2$ (mean squared

prediction error), is $g(X) = E(Y|X)$.

Proof. We will show: $E\{Y - g(X)\}^2 - E\{Y - E(Y|X)\}^2 \geq 0$.

For the left-hand side write $Ef(X, Y)$ with

$$f(X, Y) = Y^2 + g(X)^2 - 2Yg(X) - Y^2 - E^2(Y|X) + 2YE(Y|X)$$

and $Ef(X, Y) = E[E\{f(X, Y)|X\}]$

$$\begin{aligned} &= E\{g(X)^2 - 2g(X)E(Y|X) - E^2(Y|X) + 2E^2(Y|X)\} \\ &= E\{g(X) - E(Y|X)\}^2 \geq 0. \end{aligned}$$

□

4.4 Covariance and correlation

Definition.

Assume X, Y jointly distributed with means μ_x, μ_y and finite variances σ_x^2, σ_y^2 .

The **covariance** of X and Y is $Cov(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\}$.

The **correlation** of X and Y is $Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$.

If $Cov(X, Y) = 0$ we say that X and Y are *uncorrelated* (and “*correlated*” otherwise).

Properties.

(i) $Cov(X, Y) = Cov(Y, X)$, $Corr(X, Y) = Corr(Y, X)$.

(ii) $Cov(X, Y) = E(XY) - \mu_x \mu_y = E\{X(Y - \mu_y)\}$.

(iii) $Cov(X, X) = Var X$.

(iv) $Cov(aX + b, Y) = a Cov(X, Y)$, $Corr(aX + b, Y) = sign(a) \cdot Corr(X, Y)$.

Proof of (ii), (iv).

$$\begin{aligned} \text{(ii)} \quad Cov(X, Y) &= E\{(X - \mu_x)(Y - \mu_y)\} = E(XY + \mu_x \mu_y - \mu_x Y - \mu_y X) \\ &= E(XY) + \mu_x \mu_y - \mu_x EY - \mu_y EX = E(XY) + \mu_x \mu_y - 2\mu_x \mu_y = E(XY) - \mu_x \mu_y. \end{aligned}$$

$$\begin{aligned} \text{(iv)} \quad Cov(aX + b, Y) &= E\{(aX + b - aEX - b)(Y - EY)\} = a E\{(X - EX)(Y - EY)\} \\ &= a E\{(X - \mu_x)(Y - \mu_y)\} = a Cov(X, Y). \end{aligned}$$

$$\begin{aligned} Corr(aX + b, Y) &= \frac{Cov(aX + b, Y)}{\sqrt{Var(aX + b)}\sigma_y} = \frac{a Cov(X, Y)}{\sqrt{a^2 Var(X)}\sigma_y} = \frac{a Cov(X, Y)}{|a|\sigma_x \sigma_y} \\ &= sign(a) \cdot Corr(X, Y). \end{aligned}$$

□

Example (linear regression).

$E(Y|X) = \beta_0 + \beta_1 X$, $\beta_1 \neq 0$, $Var(Y|X) = \sigma^2$ (“noise” of Y).

We know $\mu_y = \beta_0 + \beta_1 \mu_x$, $\sigma_y^2 = \sigma^2 + \beta_1^2 \sigma_x^2$.

$$\begin{aligned} Cov(X, Y) &= E(XY) - \mu_x \mu_y = E\{E(XY|X)\} - \mu_x \mu_y = E\{XE(Y|X)\} - \mu_x \mu_y \\ &= E\{X(\beta_0 + \beta_1 X)\} - \mu_x(\beta_0 + \beta_1 \mu_x) = E(\beta_1 X^2 - \beta_1 \mu_x^2) = \beta_1 Var X = \beta_1 \sigma_x^2. \end{aligned}$$

$$Corr(X, Y) = Cov(X, Y) / (\sigma_x \sigma_y) = \beta_1 \sigma_x^2 / (\sigma_x \sigma_y) = \beta_1 \sigma_x / \sigma_y.$$

$$\sigma^2 \text{ small } (\sigma^2 \approx 0): Corr(X, Y) = \beta_1 \sigma_x / \sigma_y = \beta_1 \sigma_x / \sqrt{\sigma^2 + \beta_1^2 \sigma_x^2},$$

i.e. $Corr(X, Y) \approx 1$ if $\beta_1 > 0$ and $Corr(X, Y) \approx -1$ if $\beta_1 < 0$.

σ^2 large: $Corr(X, Y) \approx 0$.

More properties (same assumptions.)

(i) $Cov(X + W, Y) = Cov(X, Y) + Cov(W, Y)$,

(ii) $Var(aX + bY) = a^2 Var X + b^2 Var Y + 2ab Cov(X, Y)$,

(iii) X, Y independent $\Rightarrow Cov(X, Y) (= E(X - \mu_x)E(Y - \mu_y)) = 0$,

(iv) $|Cov(X, Y)| \leq \sigma_x \sigma_y$ ($\Rightarrow |Corr(X, Y)| \leq 1$),

(v) $|Corr(X, Y)| = 1 \iff P(Y = aX + b) = 1$ for some a, b
with $sign(a) = sign\{Corr(X, Y)\}$.

Proof of (iv). Use (ii) with $a = \sigma_y$ and $b = -\sigma_x$:

$$\begin{aligned} 0 &\leq \text{Var}(\sigma_y X - \sigma_x Y) = 2\sigma_x^2\sigma_y^2 + 2(-\sigma_x)\sigma_y \text{Cov}(X, Y) \\ \iff 2\sigma_x\sigma_y \text{Cov}(X, Y) &\leq 2\sigma_x^2\sigma_y^2 \iff \text{Cov}(X, Y) \leq \sigma_x\sigma_y. \end{aligned}$$

Now use (ii) with $a = \sigma_y$ and $b = \sigma_x$ to obtain $\text{Cov}(X, Y) \geq -\sigma_x\sigma_y$.
 $\text{Cov}(X, Y) \leq \sigma_x\sigma_y$ and $\text{Cov}(X, Y) \geq -\sigma_x\sigma_y \Rightarrow |\text{Cov}(X, Y)| \leq \sigma_x\sigma_y$.

Proof of (v).

$$\begin{aligned} \text{Corr}(X, Y) = 1 &\iff \text{Cov}(X, Y) = \sigma_x\sigma_y \iff \underbrace{\text{Var}(\sigma_y X - \sigma_x Y)}_Z \stackrel{(ii)}{=} 2\sigma_x^2\sigma_y^2 - \\ 2\sigma_x\sigma_y \underbrace{\text{Cov}(X, Y)}_{=\sigma_x\sigma_y} &= 0 \iff P(Z = EZ) = 1, \end{aligned}$$

i.e. $Z = \sigma_y X - \sigma_x Y$ is degenerate (“a constant”) and $Y = \frac{\sigma_y}{\sigma_x} X - \frac{Z}{\sigma_x}$ is linear with positive slope σ_y/σ_x and intercept Z/σ_x . Analogously, $\text{Corr}(X, Y) = -1$ implies Y linear with negative slope. \square .

Theorem. Let X be a r.v. and suppose that $g(x)$ and $h(x)$ are nondecreasing functions on an interval $I \supset \text{support}(X)$. Then $g(X)$ and $h(X)$ are positively correlated, provided the expectations exist.

Proof. We show $\text{Cov}\{g(X), h(X)\} \geq 0$ (\Rightarrow statement).

Write $\mu = Eg(X)$, $\nu = Eh(X)$. The function h is nondecreasing

\Rightarrow there is a number a with $h(x) \leq \nu \forall x \leq a$, $h(x) \geq \nu \forall x \geq a$

$\Rightarrow \{g(x) - g(a)\}\{h(x) - \nu\}$ (= “++” or “--”) ≥ 0 for every x (g is also nondecreasing). This gives

$$\begin{aligned} \text{Cov}\{g(X), h(X)\} &= E[\{g(X) - g(a) + g(a) - \mu\}\{h(X) - \nu\}] \\ &= \underbrace{E[\{g(X) - g(a)\}\{h(X) - \nu\}]}_{\geq 0} + \{g(a) - \mu\} \underbrace{E\{h(X) - \nu\}}_{=0} \\ &\geq 0. \end{aligned}$$

4.5 Bivariate normal distribution

Example (linear regression): $Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2)$, $X \sim N(\mu_x, \sigma_x^2)$. Then

$$\begin{aligned} f_{X,Y}(x, y) &= f_{Y|X=x}(y|x)f_X(x) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y - \beta_0 - \beta_1 x}{\sigma}\right)^2\right\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu_x}{\sigma_x}\right)^2\right\}. \end{aligned}$$

Use $\mu_y (= \beta_0 + \beta_1\mu_x)$ and $\sigma_y^2 (= \sigma^2 + \beta_1^2\sigma_x^2)$ to rewrite the numerator (exercise), and $\rho = \text{Corr}(X, Y) (= \beta_1\sigma_x/\sigma_y)$ for the denominator: $\sqrt{2\pi}\sigma\sqrt{2\pi}\sigma_x = 2\pi\sigma\sigma_x = 2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}$, since $\sigma = \sqrt{\sigma^2} = \sqrt{\sigma_y^2 - \beta_1^2\sigma_x^2} = \sigma_y\sqrt{1 - \beta_1^2\sigma_x^2/\sigma_y^2} = \sigma_y\sqrt{1 - \rho^2}$.

This gives the following general form for $f_{X,Y}$:

$$f_{X,Y}(x, y) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right)}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}.$$

This is the **bivariate normal**($\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho$) **distribution**.