

# HW 2

Kevin Lin

1/23/2026

Let  $X$  and  $Y$  be bivariate Gaussian with mean  $(\mu_X, \mu_Y)$ , common variance  $\sigma^2$ , and correlation coefficient  $\rho > 0$ . The bivariate mean is random and determined by  $Z$ , where

$$(\mu_X, \mu_Y) = \begin{cases} (-\mu, \mu) & Z = -1, \\ (\mu, -\mu) & Z = 1 \end{cases}$$

with equal probability for  $Z = -1$  and  $Z = 1$ .

## 1

(a) The conditional expectation  $E[Y|X, Z]$  can be calculated as:

$$E[Y|X, Z] = \mu_Y + \frac{Cov(X, Y)}{Var(X)}(X - \mu_X)$$

Since  $\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$ , we have:

$$\begin{aligned} \rho &= \frac{Cov(X, Y)}{\sqrt{\sigma^2 \dot{\sigma}^2}} \\ Cov(X, Y) &= \rho \sigma^2 \end{aligned}$$

Substituting this back into our equation:

$$\begin{aligned} E[Y|X, Z] &= \mu_Y + \frac{\rho \sigma^2}{\sigma^2}(X - \mu_X) \\ E[Y|X, Z] &= \mu_Y + \rho(X - \mu_X) \end{aligned}$$

$\therefore$

$$E[Y|X, Z] = \begin{cases} \mu - \rho(X + \mu) & Z = -1, \\ -\mu + \rho(X - \mu) & Z = 1 \end{cases}$$

The CEF consequently is linear in  $X$  for each fixed  $Z$ .

The conditional expectation  $E[Y|Z]$  is:

$$\begin{aligned} E[Y|Z] &= \mu_Y + \frac{Cov(X, Y)}{Var(X)}(E[X|Z] - \mu_X) \\ E[Y|Z] &= \mu_Y + \rho(\mu_X - \mu_Y) \\ E[Y|Z] &= \mu_Y \\ E[Y|Z] &= \begin{cases} \mu & Z = -1, \\ -\mu & Z = 1 \end{cases} \end{aligned}$$

The CEF consequently is constant for each fixed  $Z$ .

The conditional expectation  $E[Y|X]$  is:

$$\begin{aligned} E[Y|X] &= E[E[Y|X, Z]|X] = [-\mu + \rho(X - \mu)] \cdot P(Z = 1|X) + [\mu - \rho(X + \mu)] \cdot P(Z = -1|X) \\ P(Z = 1|X) &= \frac{P(X|Z = 1)P(Z = 1)}{P(X|Z = 1)P(Z = 1) + P(X|Z = -1)P(Z = -1)} \\ P(Z = 1|X) &= \frac{\exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) \cdot \frac{1}{2}}{\exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) \cdot \frac{1}{2} + \exp\left(-\frac{(X+\mu)^2}{2\sigma^2}\right) \cdot \frac{1}{2}} \\ P(Z = 1|X) &= \frac{1}{1 + \exp\left(-\frac{2\mu}{\sigma^2}X\right)} \\ P(Z = -1|X) &= 1 - P(Z = 1|X) \therefore \\ E[Y|X] &= [-\mu + \rho(X - \mu)] \cdot \frac{1}{1 + \exp\left(-\frac{2\mu}{\sigma^2}X\right)} + [\mu - \rho(X + \mu)] \cdot \left(1 - \frac{1}{1 + \exp\left(-\frac{2\mu}{\sigma^2}X\right)}\right) \\ \text{Let } p &= \frac{1}{1 + \exp\left(-\frac{2\mu}{\sigma^2}X\right)} \therefore \\ E[Y|X] &= [-\mu + \rho(X - \mu)] \cdot p + [\mu - \rho(X + \mu)] \cdot (1 - p) \\ E[Y|x] &= -\mu p + \rho X p - \rho \mu p + \mu - \mu p - \rho X + \rho X p - \rho \mu + \rho \mu p \\ E[Y|X] &= \rho X + \mu(1 - \rho)(1 - 2p) \\ E[Y|X] &= \rho X + \mu(1 - \rho) \left(1 - 2 \cdot \frac{1}{1 + \exp\left(-\frac{2\mu}{\sigma^2}X\right)}\right) \end{aligned}$$

The CEF consequently is not linear in  $X$ .

- (b) See attached Jupyter notebook.

## 2

- (a) The best linear approximations for the CEFs  $E[Y|X, Z], E[Y|Z]$  are their literal lines as they are already linear, and respectively thus are:

$$E^*[Y|X, Z] = E[Y|X, Z] = \begin{cases} \mu - \rho(X + \mu) = 1 - \frac{1}{2}(X + 1) & Z = -1, \\ -\mu + \rho(X - \mu) = -1 + \frac{1}{2}(X - 1) & Z = 1 \end{cases}$$

$$E^*[Y|Z] = E[Y|Z] = \begin{cases} \mu = 1 & Z = -1, \\ -\mu = -1 & Z = 1 \end{cases}$$

The best linear approximation for the CEF  $E[Y|X]$  can be found as:

$$\begin{aligned} b &= \frac{Cov(X, E[Y|X])}{Var(X)} = \frac{E[Cov(X, Y|Z)] + Cov(E[X|Z], E[Y|Z])}{Var(X)} \\ &= \frac{\rho\sigma^2 - \mu^2}{\sigma^2 + \mu^2} \\ a &= E[E[Y|X]] - bE[X] = 0 - b \cdot 0 = 0 \\ &\quad \ddots \\ E^*[Y|X] &= \frac{\rho\sigma^2 - \mu^2}{\sigma^2 + \mu^2} X \end{aligned}$$

- (b) See attached Jupyter notebook.

## 3

- (a) The conditional residual variance  $Var[Y - E[Y|X, Z]|X, Z]$  can be calculated as:

$$\begin{aligned} Var[Y - E[Y|X, Z]|X, Z] &= Var[Y|X, Z] \\ &= Var[Y] - \frac{Cov(X, Y)^2}{Var(X)} \\ &= \sigma^2 - \frac{(\rho\sigma^2)^2}{\sigma^2} \\ &= \sigma^2(1 - \rho^2) \end{aligned}$$

The conditional residual variance  $Var[Y - E[Y|Z]|Z]$  can be calculated as:

$$\begin{aligned} Var[Y - E[Y|Z]|Z] &= Var[Y|Z] \\ &= Var[Y] = \sigma^2 \end{aligned}$$

The conditional residual variance  $Var[Y - E[Y|X]|X]$  can be calculated as:

$$\begin{aligned}
 Var[Y - E[Y|X]|X] &= E[Var[Y|X, Z]|X] + Var[E[Y|X, Z] - E[Y|X]|X] \\
 &= E[Var[Y|X, Z]|X] + Var[E[Y|X, Z]|X] \\
 &= \sigma^2(1 - \rho^2) + p(1 - p)(E[Y|X, Z = 1] - E[Y|X, Z = -1])^2 \\
 &= \sigma^2(1 - \rho^2) + p(1 - p)(2\mu(1 + \rho))^2
 \end{aligned}$$

where  $p$  is defined as in (1a)

- (b) See attached Jupyter notebook.

## 4

- (a) When ignoring  $Z$  and looking at each subgroup, the best linear predictor coincides with each CEF, showing a seemingly perfect linear relationship between  $X$  and  $Y$  for each fixed  $Z$ . However, when you actually condition on  $Z$ , we realize that the opposite is true: there is no linear relationship between  $X$  and  $Y$  and we see a negative slope which extremely poorly fits the data.  $Z$  is the confounding variable that causes this paradoxical situation.
- (b) As we can already see when calculating the best linear predictor for  $E[Y|X]$ , the slope becomes negative even though the data trend is positive. Thus we can conclude that Simpson's paradox arises when  $\mu^2 > \rho\sigma^2$  and  $\rho > 0$ .