HE2003 Econometrics I

Tutorial 1 (Week 2) – Problems & Solutions

Academic Year 2026/2027, Semester 1

Quantitative Research Society @NTU

October 12, 2026

Overview

- **Lecture coverage:** Lecture 1 — *Introduction: why we study econometrics, correlation vs. causality, the linear regression model previewed via the California class-size example, Type I/Type II errors, and randomised controlled trials.*
- **Questions:** 4 questions from Tutorial Question Set No. 1.
- **Notation:** as in lectures — β_0, β_1 population parameters; u_i population error term (never $\hat{u}_i$ at this stage, since no data has been fitted yet); $E[\cdot]$ population expectation; $E[\cdot \mid \cdot]$ conditional expectation.
- **Question themes:** Q1 the three sources of a correlation; Q2 parameters and error term of the class-size regression; Q3 conditional versus unconditional moments and zero correlation under nonlinear dependence; Q4 Type I/Type II errors and the threshold trade-off.

For academic reference only. This document is provided for educational purposes and does not constitute an official question paper, answer key, or any formally authorised academic material. Its content is not guaranteed to be complete or accurate.

Question 1 (Correlation vs. causality: the three cases)

Problem

Simply observing a correlation between two variables X and Y is not enough to conclude that X has a causal effect on Y. When X and Y are correlated, there are several possible cases. Explain these cases with specific examples. *(Tutorial Question Set No. 1, Q1.)*

Solution

Method 1: Case enumeration with applied examples

Let X be the variable suspected to be the cause and Y the outcome, and suppose the population correlation $\text{Corr}(X, Y)$ is not zero. A sample correlation computed from data estimates this population number, but even the exact population value does not say which way the causal arrow points. Three cases are possible.

- **(i) Forward causality** $X \to Y$. Wind speed (X, metres per second) causes windmill rotation (Y, revolutions per minute): stronger wind pushes the blades faster, and the blades cannot change the wind. In the model $\text{windmill}_i = \beta_0 + \beta_1 \text{wind}_i + u_i$ the slope β_1 is a causal effect. If $\beta_1 = 2$ rpm per m/s, a gust that raises wind speed by 3 m/s raises rotation by $\beta_1 \times 3 = 2 \times 3 = 6$ rpm.
- **(ii) Reverse causality** $Y \to X$. Let X be police officers per 1,000 residents and Y crimes per 1,000 residents. Cities hire more police after crime rises, so high crime causes high police staffing. Across cities the two variables are positively correlated, and a regression of crime on police shows a positive slope that would wrongly suggest that police generate crime; the arrow actually runs from Y to X.
- **(iii) Common factor** $X \leftarrow C \to Y$, with no arrow between X and Y. Hot weather (C, daily temperature in °C) raises ice-cream sales (X) and also sends more people swimming, which raises drowning deaths (Y). The two series move together across days, yet banning ice cream would not prevent a single drowning. In the same way, the COVID-19 outbreak (C) raised both mask use (X) and online-working hours (Y), although wearing a mask does not make anyone work from home.
- All three structures produce a non-zero population correlation, so a correlation coefficient alone (population or sample) cannot tell them apart. Separating case (i) from cases (ii) and (iii) requires knowledge of the mechanism or a design, such as a randomised controlled trial, that sets X independently of everything else.

3 cases when X, Y are correlated:
(i) $X \to Y$, (ii) $Y \to X$, (iii) $C \to X$, $C \to Y$.

Method 2: Covariance algebra for each causal structure

Write each case as population equations with disturbances that are independent of the other right-hand-side variables, and compute $\mathrm{Cov}(X, Y)$ with the rules

$$\mathrm{Cov}(a + bW, V) = b\,\mathrm{Cov}(W, V), \qquad \mathrm{Cov}(W_1 + W_2, V) = \mathrm{Cov}(W_1, V) + \mathrm{Cov}(W_2, V),$$
$$\mathrm{Cov}(W, W) = \mathrm{Var}(W).$$

Case (i): $X \to Y$. Let $Y = \beta_0 + \beta_1 X + v$ with v independent of X. Then

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{Cov}(X,\ \beta_0 + \beta_1 X + v)\\ &= \beta_1\,\mathrm{Cov}(X, X) + \mathrm{Cov}(X, v)\\ &= \beta_1\,\mathrm{Var}(X) + 0\\ &= \beta_1\,\mathrm{Var}(X).\end{aligned}$$

Since $\mathrm{Var}(X) > 0$, the covariance is non-zero exactly when $\beta_1 \neq 0$, so here the correlation comes from the causal arrow $X \to Y$ (wind $\to$ windmill).

Case (ii): $Y \to X$. Let $X = \gamma_0 + \gamma_1 Y + w$ with w independent of Y; the equation for Y does not contain X. Then

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{Cov}(\gamma_0 + \gamma_1 Y + w,\ Y)\\ &= \gamma_1\,\mathrm{Cov}(Y, Y) + \mathrm{Cov}(w, Y)\\ &= \gamma_1\,\mathrm{Var}(Y) + 0\\ &= \gamma_1\,\mathrm{Var}(Y).\end{aligned}$$

With $\gamma_1 > 0$ (more police hired when crime rises) the covariance is positive, although crime does not depend on police at all.

Case (iii): $C \to X$ and $C \to Y$. Let

$$X = a_0 + a_1 C + e_1, \qquad Y = b_0 + b_1 C + e_2,$$

with C, e_1, e_2 mutually independent; neither equation contains the other variable. Then

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{Cov}(a_0 + a_1 C + e_1,\ b_0 + b_1 C + e_2)\\ &= a_1 b_1\,\mathrm{Cov}(C, C) + a_1\,\mathrm{Cov}(C, e_2) + b_1\,\mathrm{Cov}(e_1, C) + \mathrm{Cov}(e_1, e_2)\\ &= a_1 b_1\,\mathrm{Var}(C) + 0 + 0 + 0\\ &= a_1 b_1\,\mathrm{Var}(C).\end{aligned}$$

For a hypothetical numerical illustration take temperature with standard deviation 5°C, so $\mathrm{Var}(C) = 25$; ice-cream sales (hundreds of cones) with $a_1 = 3$ and $\mathrm{Var}(e_1) = 25$; drownings (deaths per day) with $b_1 = 0.1$ and $\mathrm{Var}(e_2) = 0.25$. Then

$$\begin{aligned}\mathrm{Cov}(X, Y) &= 3 \times 0.1 \times 25 = 7.5,\\ \mathrm{Var}(X) &= a_1^2\,\mathrm{Var}(C) + \mathrm{Var}(e_1) = 9 \times 25 + 25 = 250,\\ \mathrm{Var}(Y) &= b_1^2\,\mathrm{Var}(C) + \mathrm{Var}(e_2) = 0.01 \times 25 + 0.25 = 0.5,\\ \mathrm{Corr}(X, Y) &= \frac{7.5}{\sqrt{250 \times 0.5}} = \frac{7.5}{\sqrt{125}} = 0.671.\end{aligned}$$

The correlation is 0.67 although ice cream has no effect on drowning. Each of the three structures generates a non-zero covariance through a different channel (β_1, γ_1, or $a_1 b_1$), so all three must be listed.

3 cases when X, Y are correlated:
(i) $X \to Y$, (ii) $Y \to X$, (iii) $C \to X$, $C \to Y$.

Unofficial - QRS@NTU - qrsntu.org

Method 3: Intervention check via a randomised experiment

Suppose a researcher sets X by a coin flip: a treatment group receives $X = x_1$ and a control group receives $X = x_0$, independently of every other variable. Under random assignment $E[\varepsilon \mid X = x_1] = E[\varepsilon \mid X = x_0] = E[\varepsilon]$ for any disturbance ε that is not caused by X. Compute the treatment–control difference in mean outcomes in each case.

Case (i). With $Y = \beta_0 + \beta_1 X + v$,

$$\begin{aligned} E[Y \mid X = x_1] - E[Y \mid X = x_0] &= (\beta_0 + \beta_1 x_1 + E[v]) - (\beta_0 + \beta_1 x_0 + E[v]) \\ &= \beta_1 (x_1 - x_0). \end{aligned}$$

For the windmill with $\beta_1 = 2$ rpm per m/s and $x_1 - x_0 = 3$ m/s, the difference is $2 \times 3 = 6$ rpm. The relation survives the experiment.

Case (ii). Crime is generated by its own causes, $Y = \mu_Y + \varepsilon$, and does not contain X. Once police staffing is set by the coin, it no longer responds to crime, and

$$\begin{aligned} E[Y \mid X = x_1] - E[Y \mid X = x_0] &= (\mu_Y + E[\varepsilon]) - (\mu_Y + E[\varepsilon]) \\ &= 0. \end{aligned}$$

The positive observational correlation disappears.

Case (iii). With $Y = b_0 + b_1 C + e_2$ and X assigned independently of C and e_2,

$$\begin{aligned} E[Y \mid X = x_1] - E[Y \mid X = x_0] &= (b_0 + b_1 E[C] + E[e_2]) - (b_0 + b_1 E[C] + E[e_2]) \\ &= 0. \end{aligned}$$

With the numbers of the ice-cream example, the observational regression slope of drownings on sales is $\mathrm{Cov}(X, Y)/\mathrm{Var}(X) = 7.5/250 = 0.03$ deaths per hundred cones, while handing out free ice cream at randomly chosen beaches changes drownings by 0.

Only case (i) leaves a treatment effect under randomisation; cases (ii) and (iii) produce a correlation in observational data that vanishes once X is set from outside. The three cases are therefore distinct.

3 cases when X, Y are correlated:
(i) $X \to Y$, (ii) $Y \to X$, (iii) $C \to X$, $C \to Y$.

Question 2 (The class-size / test-score linear regression model)

Problem

Explain the linear regression model in the example of class size and student performance. What are the parameters of the model? How to interpret these parameters? What is the meaning of the error term? *(Tutorial Question Set No. 1, Q2.)*

Solution

Method 1: Population model and parameter interpretation

The unit of observation is a California school district, $i = 1, \ldots, n$ with $n = 420$. The population linear regression model is

$$\text{TestScore}_i = \beta_0 + \beta_1 \, \text{STR}_i + u_i, \qquad i = 1, \ldots, n \quad (n = 420 \text{ districts}).$$

- TestScore_i is the average test score of district i, measured in points; it is the dependent variable. STR_i is the student–teacher ratio of district i, measured in students per teacher; it is the regressor.
- The parameters β_0 and β_1 are fixed, unknown population numbers that define the population regression line $\beta_0 + \beta_1 \text{STR}$. They are not computed from the 420 districts; the sample estimates $\hat{\beta}_0, \hat{\beta}_1$ obtained later by OLS are approximations to them.
- β_0 is the intercept: the height of the population line at $\text{STR} = 0$, in points. No district has zero students per teacher, so β_0 has no policy meaning by itself; it anchors the level of the line.
- $\beta_1 = \Delta\text{TestScore}/\Delta\text{STR}$ is the slope: the change in test score, in points, when STR rises by one student per teacher with all other factors held fixed. For example, if $\beta_1 = -2$ and a district cuts its class size so that $\Delta\text{STR} = -2$, then

$$\Delta\text{TestScore} = \beta_1 \times \Delta\text{STR} = (-2) \times (-2) = 4 \text{ points}.$$

- u_i is the error term: all factors other than STR that affect district i's score, such as family income, parental education, the share of English learners, teacher quality and luck on the test day. It is the vertical gap between district i's actual score and the population line, $u_i = \text{TestScore}_i - (\beta_0 + \beta_1 \text{STR}_i)$, and it is never observed because β_0, β_1 are unknown.

$$\boxed{\begin{array}{c} \text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i, \qquad \beta_1 = \Delta\text{TestScore}/\Delta\text{STR}, \\ u_i = \text{all other factors affecting TestScore}_i. \end{array}}$$

Method 2: Conditional-mean (population regression function) reading

(uses Lecture 3: Least Squares Assumption 1.)

Start from $\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$ and assume $E[u_i \,|\, \text{STR}_i] = 0$. Take the conditional expectation of both sides given STR_i:

$$\begin{aligned} E[\text{TestScore}_i \,|\, \text{STR}_i] &= E[\beta_0 + \beta_1 \text{STR}_i + u_i \,|\, \text{STR}_i] \\ &= \beta_0 + \beta_1 \text{STR}_i + E[u_i \,|\, \text{STR}_i] \qquad \text{(linearity)} \\ &= \beta_0 + \beta_1 \text{STR}_i. \end{aligned}$$

The population line is therefore the mean test score among all districts with a given class size.

Slope. Compare two class sizes x and $x + 1$:

$$\begin{aligned} &E[\text{TestScore} \,|\, \text{STR} = x + 1] - E[\text{TestScore} \,|\, \text{STR} = x] \\ &\quad = \big(\beta_0 + \beta_1 (x + 1)\big) - \big(\beta_0 + \beta_1 x\big) \\ &\quad = \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\ &\quad = \beta_1. \end{aligned}$$

For a change of Δ students per teacher the same subtraction gives $\beta_1 \Delta$, so

$$\beta_1 = \frac{\Delta \text{TestScore}}{\Delta \text{STR}} \quad \text{(points per student per teacher)}.$$

With $\beta_1 = -2$ and $\Delta = -2$ the average score rises by $(-2)(-2) = 4$ points.

Intercept. Setting $x = 0$,

$$E[\text{TestScore} \,|\, \text{STR} = 0] = \beta_0 + \beta_1 \cdot 0 = \beta_0,$$

so β_0 is the mean score the line assigns to a class size of zero, a point outside the data.

Error term. Rearranging the model and using the conditional mean,

$$u_i = \text{TestScore}_i - (\beta_0 + \beta_1 \text{STR}_i) = \text{TestScore}_i - E[\text{TestScore}_i \,|\, \text{STR}_i],$$

and

$$E[u_i \,|\, \text{STR}_i] = E[\text{TestScore}_i \,|\, \text{STR}_i] - E[\text{TestScore}_i \,|\, \text{STR}_i] = 0.$$

So u_i is how far district i lies above or below the average district with the same STR; it collects all other factors and averages to zero at every class size.

Sample counterpart. The population objects are estimated from the 420 districts by the OLS line $\widehat{\text{TestScore}} = 698.9 - 2.28\,\text{STR}$ (Lecture 2). At $\text{STR} = 20$,

$$\widehat{\text{TestScore}} = 698.9 - 2.28 \times 20 = 698.9 - 45.6 = 653.3 \text{ points},$$

an estimate of $E[\text{TestScore} \,|\, \text{STR} = 20]$; $\hat{\beta}_1 = -2.28$ estimates β_1, and the residual $\hat{u}_i$ is the sample analogue of u_i.

$$\boxed{\begin{array}{c} \text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i, \qquad \beta_1 = \Delta\text{TestScore}/\Delta\text{STR}, \\ u_i = \text{all other factors affecting TestScore}_i. \end{array}}$$

Method 3: Two-district comparison holding other factors fixed

Write the model for two districts A and B:

$$\begin{aligned}\text{TestScore}_A &= \beta_0 + \beta_1 \text{STR}_A + u_A,\\ \text{TestScore}_B &= \beta_0 + \beta_1 \text{STR}_B + u_B.\end{aligned}$$

Subtract the second line from the first:

$$\begin{aligned}\text{TestScore}_A - \text{TestScore}_B &= (\beta_0 - \beta_0) + \beta_1(\text{STR}_A - \text{STR}_B) + (u_A - u_B)\\ &= \beta_1(\text{STR}_A - \text{STR}_B) + (u_A - u_B).\end{aligned}$$

Slope. If the two districts are identical in every factor other than class size, then $u_A = u_B$ and

$$\beta_1 = \frac{\text{TestScore}_A - \text{TestScore}_B}{\text{STR}_A - \text{STR}_B} = \frac{\Delta\text{TestScore}}{\Delta\text{STR}},$$

the effect on the score (points) of one more student per teacher with everything else held fixed.

Intercept. β_0 cancels in the difference, so it plays no role in comparing districts; it only fixes the level of the line, the score the line assigns at $\text{STR} = 0$ when $u = 0$.

Error term. When the districts differ in other factors, $u_A - u_B \neq 0$ and the observed gap mixes the class-size effect with everything else. For example, let $\beta_1 = -2$, $\text{STR}_A = 18$, $\text{STR}_B = 20$, and suppose district A has wealthier families so that $u_A - u_B = 5$ points. Then

$$\begin{aligned}\text{TestScore}_A - \text{TestScore}_B &= (-2)(18 - 20) + 5\\ &= 4 + 5 = 9 \text{ points},\end{aligned}$$

and the naive ratio $9/(18-20) = -4.5$ is more than twice the true slope -2 in magnitude. The error term u_i is exactly this collection of other factors.

$$\boxed{\begin{array}{c}\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i, \qquad \beta_1 = \Delta\text{TestScore}/\Delta\text{STR},\\ u_i = \text{all other factors affecting TestScore}_i.\end{array}}$$

Question 3 (Conditional and unconditional moments of $Y = X^2 + Z$)

Problem

Let X and Z be two independently distributed standard normal random variables, and let $Y = X^2 + Z$.

(a) Show that $E[Y \mid X] = X^2$.

(b) Show that $E[Y] = 1$.

(c) Show that $E[XY] = 0$.

(Tutorial Question Set No. 1, Q3.)

Solution

Method 1: Linearity, properties of conditional expectation and independence

Recall the facts used below. Since $X, Z \sim N(0,1)$,

$$E[X] = E[Z] = 0,$$
$$E[X^2] = \mathrm{Var}(X) + \big(E[X]\big)^2 = 1 + 0^2 = 1.$$

The standard normal density ϕ satisfies $\phi(-x) = \phi(x)$, so $x^3\phi(x)$ is an odd function; it is absolutely integrable because every moment of the normal distribution is finite, so its integral over $\mathbb{R}$ is zero:

$$E[X^3] = \int_{-\infty}^{\infty} x^3\phi(x)\,dx = 0.$$

Because X and Z are independent,

$$E[Z \mid X] = E[Z] = 0, \qquad E[XZ] = E[X]\,E[Z] = 0 \times 0 = 0.$$

(a) Substitute $Y = X^2 + Z$ and use linearity of conditional expectation:

$$\begin{aligned} E[Y \mid X] &= E[X^2 + Z \mid X] \\ &= E[X^2 \mid X] + E[Z \mid X] && \text{(linearity)} \\ &= X^2 + E[Z \mid X] && (X^2 \text{ known given } X) \\ &= X^2 + E[Z] && (X \perp Z) \\ &= X^2 + 0. \end{aligned}$$

$$\boxed{E[Y \mid X] = X^2.}$$

(b) Take the unconditional expectation of $Y = X^2 + Z$:

$$\begin{aligned} E[Y] &= E[X^2 + Z] \\ &= E[X^2] + E[Z] && \text{(linearity)} \\ &= 1 + 0. \end{aligned}$$

$$\boxed{E[Y] = 1.}$$

(c) Multiply out $XY = X(X^2 + Z) = X^3 + XZ$:

$$\begin{aligned} E[XY] &= E[X^3 + XZ] \\ &= E[X^3] + E[XZ] && \text{(linearity)} \\ &= E[X^3] + E[X]E[Z] && (X \perp Z) \\ &= 0 + 0 \times 0 = 0. \end{aligned}$$

$$\boxed{E[XY] = 0.}$$

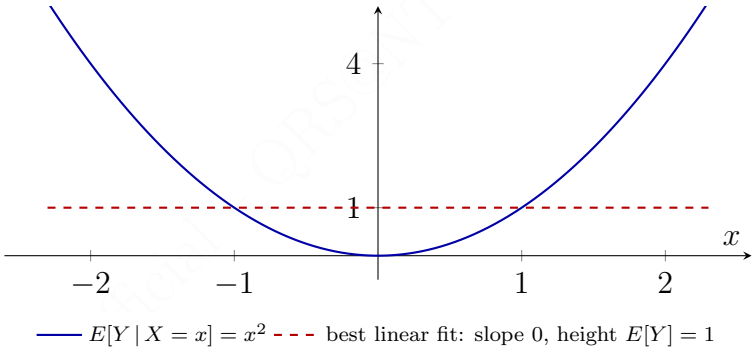
Correlation. By the definition of covariance and the results (b), (c),

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= 0 - 0 \times 1 = 0. \end{aligned}$$

Also $\mathrm{Var}(X) = 1$ and $\mathrm{Var}(Y) = \mathrm{Var}(X^2) + \mathrm{Var}(Z) + 2\,\mathrm{Cov}(X^2, Z) = \mathrm{Var}(X^2) + 1 + 0 \geq 1 > 0$, so

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{0}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 0.$$

Yet by (a) the conditional mean $E[Y \mid X] = X^2$ changes with X: at $X = 0$ it is 0, at $X = 2$ it is 4. The dependence is a U-shape, and correlation measures only the linear part of it.



$$\boxed{\begin{array}{c} E[Y \mid X] = X^2, \qquad E[Y] = 1, \qquad E[XY] = 0, \\ \text{yet } \mathrm{Corr}(X, Y) = 0 \text{ while } Y \text{ depends strongly on } X. \end{array}}$$

Method 2: Law of iterated expectations

(a) Condition on the event $X = x$. Then X^2 is the constant x^2, and because Z is independent of X, conditioning on $X = x$ leaves the distribution of Z unchanged:

$$\begin{aligned} E[Y \mid X = x] &= E[x^2 + Z \mid X = x] \\ &= x^2 + E[Z \mid X = x] \\ &= x^2 + E[Z] \\ &= x^2. \end{aligned}$$

Replacing x by X,

$$\boxed{E[Y \mid X] = X^2.}$$

(b) Recall the law of iterated expectations, $E[Y] = E\big[E[Y \mid X]\big]$. Using (a),

$$\begin{aligned} E[Y] &= E\big[E[Y \mid X]\big] && \text{(LIE)} \\ &= E[X^2] && \text{(by (a))} \\ &= \mathrm{Var}(X) + \big(E[X]\big)^2 = 1 + 0 = 1. \end{aligned}$$

$$\boxed{E[Y] = 1.}$$

(c) Given X, the factor X is known and can be taken outside the conditional expectation:

$$E[XY \mid X] = X\,E[Y \mid X] = X \cdot X^2 = X^3.$$

Apply the LIE to XY:

$$\begin{aligned} E[XY] &= E\big[E[XY \mid X]\big] && \text{(LIE)} \\ &= E[X^3] \\ &= 0 && \text{(odd moment of } N(0,1)\text{)}. \end{aligned}$$

$$\boxed{E[XY] = 0.}$$

Correlation. The LIE also gives the covariance directly through the conditional mean. Since $E[XY] = E\big[X\,E[Y \mid X]\big]$ and $E[Y] = E\big[E[Y \mid X]\big]$,

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E\big[X\,E[Y \mid X]\big] - E[X]\,E\big[E[Y \mid X]\big] \\ &= E[X \cdot X^2] - E[X]\,E[X^2] \\ &= \mathrm{Cov}(X, X^2) \\ &= 0 - 0 \times 1 = 0. \end{aligned}$$

So $\mathrm{Corr}(X, Y) = 0$: the only part of Y that is related to X is X^2, and X^2 is uncorrelated with X because X is symmetric about 0, even though X^2 is an exact function of X.

$$\boxed{\begin{array}{c} E[Y \mid X] = X^2, \qquad E[Y] = 1, \qquad E[XY] = 0, \\ \text{yet } \mathrm{Corr}(X, Y) = 0 \text{ while } Y \text{ depends strongly on } X. \end{array}}$$

Method 3: Conditional distribution and density integrals

Let $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. Three integrals are needed; each is computed from $\phi'(t) = -t\phi(t)$, so that $-\phi$ is an antiderivative of $t\phi(t)$, and $t^k\phi(t) \to 0$ as $t \to \pm\infty$ for every k:

$$\begin{aligned}
\int_{-\infty}^{\infty} t\,\phi(t)\,dt &= \Big[-\phi(t)\Big]_{-\infty}^{\infty} = 0 - 0 = 0,\\
\int_{-\infty}^{\infty} t^2\phi(t)\,dt &= \int_{-\infty}^{\infty} t\cdot t\phi(t)\,dt = \Big[-t\phi(t)\Big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(t)\,dt = 0 + 1 = 1,\\
\int_{-\infty}^{\infty} t^3\phi(t)\,dt &= \int_{-\infty}^{\infty} t^2\cdot t\phi(t)\,dt = \Big[-t^2\phi(t)\Big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 2t\,\phi(t)\,dt = 0 + 2\times 0 = 0.
\end{aligned}$$

The second and third lines are integration by parts with $dv = t\phi(t)\,dt$, $v = -\phi(t)$.

(a) Given $X = x$, $Y = x^2 + Z$ with $Z \sim N(0,1)$ independent of X, so $Y \mid X = x \sim N(x^2, 1)$ with density $\phi(y - x^2)$. Substitute $t = y - x^2$, $dy = dt$:

$$\begin{aligned}
E[Y \mid X = x] &= \int_{-\infty}^{\infty} y\,\phi(y - x^2)\,dy\\
&= \int_{-\infty}^{\infty} (t + x^2)\phi(t)\,dt\\
&= \int_{-\infty}^{\infty} t\phi(t)\,dt + x^2\int_{-\infty}^{\infty} \phi(t)\,dt\\
&= 0 + x^2\times 1 = x^2.
\end{aligned}$$

$$\boxed{E[Y \mid X] = X^2.}$$

(b) The joint density of (X, Z) is $\phi(x)\phi(z)$ by independence. The integrand $(x^2 + z)\phi(x)\phi(z)$ is absolutely integrable because the normal moments are finite, so by Fubini's theorem the double integral splits into products of single integrals:

$$\begin{aligned}
E[Y] &= \iint (x^2 + z)\,\phi(x)\phi(z)\,dx\,dz\\
&= \int x^2\phi(x)\,dx \int \phi(z)\,dz + \int \phi(x)\,dx \int z\phi(z)\,dz\\
&= 1\times 1 + 1\times 0 = 1.
\end{aligned}$$

$$\boxed{E[Y] = 1.}$$

(c) In the same way, with $x(x^2 + z) = x^3 + xz$,

$$\begin{aligned}
E[XY] &= \iint (x^3 + xz)\,\phi(x)\phi(z)\,dx\,dz\\
&= \int x^3\phi(x)\,dx \int \phi(z)\,dz + \int x\phi(x)\,dx \int z\phi(z)\,dz\\
&= 0\times 1 + 0\times 0 = 0.
\end{aligned}$$

$$\boxed{E[XY] = 0.}$$

Correlation. One more integration by parts gives the fourth moment:

$$\int_{-\infty}^{\infty} t^4\phi(t)\,dt = \Big[-t^3\phi(t)\Big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 3t^2\phi(t)\,dt = 0 + 3\times 1 = 3.$$

Hence, using $\mathrm{Cov}(X^2, Z) = 0$ by independence,

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(X^2) + \mathrm{Var}(Z) = \big(E[X^4] - (E[X^2])^2\big) + 1 = (3-1) + 1 = 3,\\
\mathrm{Cov}(X, Y) &= E[XY] - E[X]E[Y] = 0 - 0\times 1 = 0,\\
\mathrm{Corr}(X, Y) &= \frac{0}{\sqrt{1\times 3}} = 0.
\end{aligned}$$

The variance of Y is 3, of which $\mathrm{Var}(X^2) = 2$ is explained exactly by X, yet the correlation is 0.

$$\boxed{\begin{array}{c} E[Y \mid X] = X^2, \qquad E[Y] = 1, \qquad E[XY] = 0, \\ \text{yet } \mathrm{Corr}(X, Y) = 0 \text{ while } Y \text{ depends strongly on } X. \end{array}}$$

Unofficial QRS@NTU material

Question 4 (Type I and Type II errors and the trade-off)

Problem

Explain the Type I and Type II errors and the trade-off between the two errors by using a specific example. *(Tutorial Question Set No. 1, Q4.)*

Solution

Method 1: COVID ART-test example

A COVID-19 antibody rapid test (ART) measures a person's antibody level L (in units). The hypotheses and the decision rule are

$$H_0: \text{ the person is healthy}, \qquad H_1: \text{ the person is infected},$$
$$\text{reject } H_0 \iff L > c, \qquad c = 10 \text{ units}.$$

- **Type I error** (false positive): rejecting H_0 when H_0 is true. A healthy person whose antibody level happens to exceed 10 units is told that they are infected. Its probability is $\alpha = P(L > c \,|\, H_0)$, the significance level of the test. If $\alpha = 0.05$, then among 1,000 healthy people tested about $0.05 \times 1{,}000 = 50$ are wrongly sent into quarantine.
- **Type II error** (false negative): failing to reject H_0 when H_1 is true. An infected person whose antibody level is at most 10 units is cleared and may go on to spread the virus. Its probability is $\beta = P(L \leq c \,|\, H_1)$.
- **Raising the threshold** $c : 10 \to 100$. Rejection now requires $L > 100$. Every healthy person with $L \in (10, 100]$, who was falsely flagged before, is now cleared, so α falls. Every infected person with $L \in (10, 100]$, who was caught before, is now missed, so β rises.
- **Lowering the threshold** $c : 10 \to 1$. Rejection now happens whenever $L > 1$. Infected people with $L \in (1, 10]$ are now caught, so β falls; healthy people with $L \in (1, 10]$ are now flagged, so α rises.
- The antibody levels of healthy and infected people overlap, so a single threshold cannot remove both errors at once. Choosing c means choosing which mistake to tolerate: unnecessary quarantine (Type I) against undetected spread (Type II).

$$\text{Type I error} = P(\text{reject } H_0 \,|\, H_0 \text{ true}) = \alpha,$$
$$\text{Type II error} = P(\text{fail to reject } H_0 \,|\, H_1 \text{ true}) = \beta,$$
$$\alpha \downarrow \Rightarrow \ \beta \uparrow \ \text{(and vice versa)}.$$

Method 2: Tail areas of the densities

Let the antibody level L have density f_0 and CDF F_0 when H_0 is true, and density f_1 and CDF F_1 when H_1 is true. The test rejects H_0 when $L > c$.

Type I error. It is the probability of the rejection region under H_0, an upper-tail area:

$$\alpha(c) = P(L > c \,|\, H_0) = 1 - P(L \leq c \,|\, H_0) = 1 - F_0(c).$$

Differentiate with respect to c, using $F_0' = f_0$:

$$\alpha'(c) = -F_0'(c) = -f_0(c) \leq 0.$$

Type II error. It is the probability of the acceptance region under H_1, a lower-tail area:

$$\beta(c) = P(L \leq c \,|\, H_1) = F_1(c).$$

Differentiate:

$$\beta'(c) = F_1'(c) = f_1(c) \geq 0.$$

Trade-off. At any threshold c where both densities are positive, $\alpha'(c) < 0 < \beta'(c)$, so raising c lowers α and raises β, and lowering c does the reverse. Along the family of tests indexed by c, the rate at which one error is exchanged for the other is

$$\frac{d\beta}{d\alpha} = \frac{\beta'(c)}{\alpha'(c)} = \frac{f_1(c)}{-f_0(c)} = -\frac{f_1(c)}{f_0(c)} < 0.$$

To make $\alpha = 0$ the threshold must be at least the largest antibody level of a healthy person, $\bar{\ell}_0$; to make $\beta = 0$ it must be below the smallest level of an infected person, $\underline{\ell}_1$. Because the two distributions overlap, $\underline{\ell}_1 < \bar{\ell}_0$, so no c satisfies $c \geq \bar{\ell}_0$ and $c < \underline{\ell}_1$ simultaneously, and the two errors cannot both be zero.

$$\begin{array}{c} \text{Type I error} = P(\text{reject } H_0 \,|\, H_0 \text{ true}) = \alpha, \\ \text{Type II error} = P(\text{fail to reject } H_0 \,|\, H_1 \text{ true}) = \beta, \\ \alpha \downarrow \Rightarrow \ \beta \uparrow \ \text{(and vice versa)}. \end{array}$$

Method 3: Numerical computation in a normal model

For a numerical illustration, suppose that antibody levels are normal: $L \sim N(5, 3^2)$ for a healthy person and $L \sim N(30, 10^2)$ for an infected person (hypothetical values, in units). Standardise with Φ, the standard normal CDF:

$$\alpha(c) = P(L > c \,|\, H_0) = P\left(\frac{L-5}{3} > \frac{c-5}{3}\right) = 1 - \Phi\left(\frac{c-5}{3}\right),$$

$$\beta(c) = P(L \le c \,|\, H_1) = P\left(\frac{L-30}{10} \le \frac{c-30}{10}\right) = \Phi\left(\frac{c-30}{10}\right).$$

Original threshold $c = 10$.

$$\alpha(10) = 1 - \Phi\left(\frac{10-5}{3}\right) = 1 - \Phi(1.667) = 1 - 0.9522 = 0.0478,$$

$$\beta(10) = \Phi\left(\frac{10-30}{10}\right) = \Phi(-2) = 0.0228.$$

Raised threshold $c = 100$.

$$\alpha(100) = 1 - \Phi\left(\frac{100-5}{3}\right) = 1 - \Phi(31.67) \approx 0.0000,$$

$$\beta(100) = \Phi\left(\frac{100-30}{10}\right) = \Phi(7) \approx 1.0000.$$

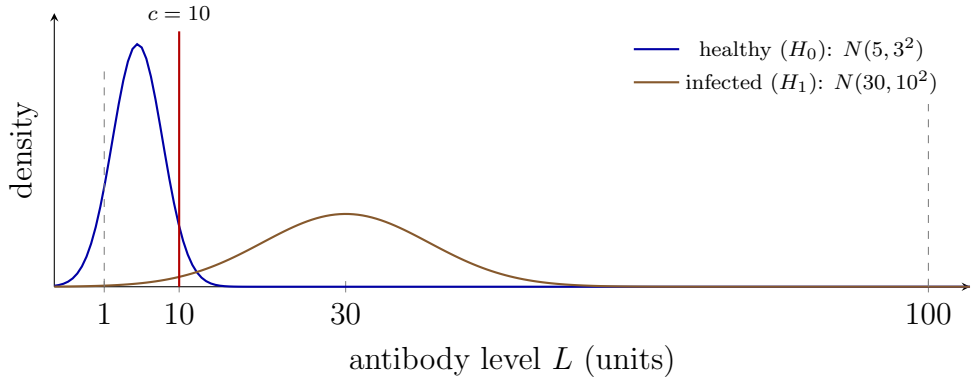
Lowered threshold $c = 1$.

$$\alpha(1) = 1 - \Phi\left(\frac{1-5}{3}\right) = 1 - \Phi(-1.333) = 1 - 0.0912 = 0.9088,$$

$$\beta(1) = \Phi\left(\frac{1-30}{10}\right) = \Phi(-2.9) = 0.0019.$$

threshold c (units)	$\alpha = P$(Type I)	$\beta = P$(Type II)
1	0.9088	0.0019
10	0.0478	0.0228
100	0.0000	1.0000

Moving from $c = 10$ to $c = 100$ cuts α from 4.8% to essentially 0 but raises β from 2.3% to essentially 100%: almost every infected person is cleared. Moving from $c = 10$ to $c = 1$ cuts β to 0.2% but raises α to 90.9%: almost every healthy person is flagged. The two densities below overlap, roughly between 5 and 15 units, which is why no threshold drives both columns to zero.



Type I error $= P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$,
Type II error $= P(\text{fail to reject } H_0 \mid H_1 \text{ true}) = \beta$,
$\alpha \downarrow \Rightarrow \ \beta \uparrow$ (and vice versa).

Unofficial - QRS@NTU - qrsntu.org

Final Answers

1. Three cases when X, Y are correlated: (i) $X \to Y$ (e.g. wind $\to$ windmill rotation); (ii) $Y \to X$ (reverse causality, e.g. crime $\to$ police staffing); (iii) common factor $C \to X$, $C \to Y$ (e.g. temperature $\to$ {ice-cream sales, drowning deaths}; COVID $\to$ {mask use, online-working hours}).

2. $\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$; β_0 is the intercept (value at STR $= 0$); $\beta_1 = \Delta\text{TestScore}/\Delta\text{STR}$ (points per unit STR; e.g. $\beta_1 = -2 \Rightarrow \Delta\text{TestScore} = 4$ for $\Delta\text{STR} = -2$); u_i is all other factors affecting test score.

3. $E[Y \mid X] = X^2$; $E[Y] = 1$; $E[XY] = 0$; yet $\text{Corr}(X, Y) = 0$ while Y depends strongly (nonlinearly) on X.

4. Type I error $= P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$; Type II error $= P(\text{fail to reject } H_0 \mid H_1 \text{ true}) = \beta$. Raising the ART threshold ($10 \to 100$) lowers α but raises β; lowering it ($10 \to 1$) reverses this.

Unofficial QRS@NTU – qrsntu.org