HE2003 Econometrics I

Quiz Practice 2 (Solutions)

AY 2026/2027, Semester 1

Quantitative Research Society @NTU

October 12, 2026

Overview

- **Coverage:** Lectures 1–5 and Lecture 6 slides 1–3 (large-sample distribution of $\hat{\beta}_1$ and its standard error); no t-tests, confidence intervals, p-values or Stata. This paper is slightly harder than Practice 1: it includes an estimator-comparison-with-weights derivation, a Least Squares Assumption 1 counterexample, a $TSS/SSR/R^2$ computation, and a sampling-distribution question about the difference of two independent sample means.
- **Format:** 5 questions, 50 marks total, approximately 60 minutes.
- **Question mix:** Q1 derivation (efficient combination of two independent unbiased estimators, with weights); Q2 regression-output interpretation (prediction, change in prediction, and $TSS/SSR/R^2$); Q3 sampling distribution of a difference of two independent sample means (mean, variance, a standardised probability, and the effect of quadrupling n on the standard deviation); Q4 binary regressor (group means, and a common-slope/different-intercepts extension with a second regressor); Q5 conceptual (selection bias, a Least Squares Assumption 1 counterexample, Type I/II trade-off).
- **Standard normal table** ($\Phi(z) = P(Z \leq z)$ for $Z \sim N(0,1)$), needed in Question 3:

z	0.50	1.00	1.28	1.50	1.645	1.96	2.00	2.50	2.58	3.00
$\Phi(z)$	0.6915	0.8413	0.8997	0.9332	0.9500	0.9750	0.9772	0.9938	0.9951	0.9987

Use $P(Z > z) = 1 - \Phi(z)$ and, by symmetry, $\Phi(-z) = 1 - \Phi(z)$.

Instructions

Answer **all** questions. Show all working clearly; answers without justification may receive zero marks.

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.

Contents

Question 1 **(10 Marks)** **3**
Solution . . . 3
Method 1: Variance rules and the first-order condition . . . 3
Method 2: Deviations from μ and completing the square . . . 5
Method 3: Constrained minimisation over linear combinations (Lagrange multiplier) . . . 7
Marking scheme (10 marks) . . . 8

Question 2 **(10 Marks)** **9**
Solution . . . 9
Method 1: Substitution into the fitted line and ESS = TSS − SSR . . . 9
Method 2: Slope times change in Size and $R^2 = 1 - \text{SSR}/\text{TSS}$ . . . 11
Method 3: R^2 as the squared correlation between Y_i and $\hat{Y}_i$ . . . 11
Marking scheme (10 marks) . . . 13

Question 3 **(10 Marks)** **14**
Solution . . . 14
Method 1: Expectation and variance rules for the two sample means . . . 14
Method 2: Sampling formula from the individual observations and $n \to 4n$ scaling 16
Marking scheme (10 marks) . . . 17

Question 4 **(10 Marks)** **18**
Solution . . . 18
Method 1: Substituting $D = 0$ and $D = 1$ into the fitted line . . . 18
Method 2: OLS formula with group means and coefficient matching . . . 21
Method 3: First-order conditions and partial effects . . . 23
Marking scheme (10 marks) . . . 24

Question 5 **(10 Marks)** **26**
Solution . . . 26
Method 1: Least Squares Assumption 1 as a conditional-mean condition . . . 26
Method 2: Omitted-variable-bias formula, covariance and a threshold model . . . 29
Marking scheme (10 marks) . . . 31

Question 1 (10 Marks)

A government agency wants to estimate the true population mean μ of weekly working hours in a sector. It combines two **independent** unbiased surveys: survey 1 has $\hat{\mu}_1$ with $\text{Var}(\hat{\mu}_1) = v_1$, and survey 2 has $\hat{\mu}_2$ with $\text{Var}(\hat{\mu}_2) = v_2$, where both are unbiased for μ, i.e. $E[\hat{\mu}_1] = E[\hat{\mu}_2] = \mu$. For a weight $w \in [0, 1]$, consider the combined estimator

$$\hat{\mu}_w = w\hat{\mu}_1 + (1 - w)\hat{\mu}_2.$$

(a) (3 marks) Show that $\hat{\mu}_w$ is unbiased for μ for *every* choice of $w \in [0, 1]$.

(b) (4 marks) Using independence of $\hat{\mu}_1$ and $\hat{\mu}_2$, derive $\text{Var}(\hat{\mu}_w)$ as a function of w, v_1, and v_2.

(c) (3 marks) Find the weight w^* that minimizes $\text{Var}(\hat{\mu}_w)$, and show that it can be written as $w^* = \dfrac{v_2}{v_1 + v_2}$ (i.e. the more precise survey, with the smaller variance, receives the larger weight).

Solution

Method 1: Variance rules and the first-order condition

(a) Fix any $w \in [0, 1]$. The weights w and $1 - w$ are constants, so linearity of expectation applies to the combined estimator:

$$\begin{aligned} E[\hat{\mu}_w] &= E\big[w\hat{\mu}_1 + (1 - w)\hat{\mu}_2\big] \\ &= wE[\hat{\mu}_1] + (1 - w)E[\hat{\mu}_2] && \text{(linearity)} \\ &= w\mu + (1 - w)\mu && \text{(both unbiased)} \\ &= w\mu + \mu - w\mu \\ &= \mu. \end{aligned}$$

The calculation used only that w is a constant, so it holds for every $w \in [0, 1]$. Therefore

$$\boxed{E[\hat{\mu}_w] = \mu \text{ for all } w \in [0, 1].}$$

(b) Recall the variance of a linear combination: for constants a, b and random variables X, Y,

$$\text{Var}(aX + bY) = a^2 \,\text{Var}(X) + b^2 \,\text{Var}(Y) + 2ab \,\text{Cov}(X, Y).$$

Apply it with $a = w$, $b = 1 - w$, $X = \hat{\mu}_1$, $Y = \hat{\mu}_2$:

$$\begin{aligned} \text{Var}(\hat{\mu}_w) &= \text{Var}\left(w\hat{\mu}_1 + (1 - w)\hat{\mu}_2\right) \\ &= w^2 \,\text{Var}(\hat{\mu}_1) + (1 - w)^2 \,\text{Var}(\hat{\mu}_2) + 2w(1 - w) \,\text{Cov}(\hat{\mu}_1, \hat{\mu}_2) \\ &= w^2 v_1 + (1 - w)^2 v_2 + 2w(1 - w) \,\text{Cov}(\hat{\mu}_1, \hat{\mu}_2). \end{aligned}$$

The two surveys are independent, so $E[\hat{\mu}_1\hat{\mu}_2] = E[\hat{\mu}_1]E[\hat{\mu}_2]$, and the covariance vanishes:

$$\begin{aligned} \text{Cov}(\hat{\mu}_1, \hat{\mu}_2) &= E[\hat{\mu}_1\hat{\mu}_2] - E[\hat{\mu}_1]E[\hat{\mu}_2] \\ &= E[\hat{\mu}_1]E[\hat{\mu}_2] - E[\hat{\mu}_1]E[\hat{\mu}_2] && \text{(independence)} \\ &= 0. \end{aligned}$$

Substituting $\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) = 0$,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_w) &= w^2 v_1 + (1-w)^2 v_2 + 2w(1-w)\cdot 0 \\ &= w^2 v_1 + (1-w)^2 v_2.\end{aligned}$$

Therefore

$$\boxed{\mathrm{Var}(\hat{\mu}_w) = w^2 v_1 + (1-w)^2 v_2.}$$

(c) Let $V(w) = w^2 v_1 + (1-w)^2 v_2$. Differentiate, using the chain rule on $(1-w)^2$:

$$\begin{aligned}V'(w) &= 2wv_1 + 2(1-w)\cdot(-1)\cdot v_2 \\ &= 2wv_1 - 2(1-w)v_2.\end{aligned}$$

Set the first-order condition $V'(w) = 0$ and solve for w:

$$\begin{aligned}2wv_1 - 2(1-w)v_2 &= 0 \\ wv_1 - v_2 + wv_2 &= 0 \qquad \text{(divide by 2)} \\ w(v_1+v_2) &= v_2 \\ w &= \frac{v_2}{v_1+v_2}.\end{aligned}$$

For the second-order condition, differentiate $V'(w)$ once more:

$$V''(w) = 2v_1 + 2v_2 = 2(v_1+v_2) > 0,$$

because both variances are positive. Hence V is a strictly convex quadratic in w, and its unique stationary point is its global minimiser. Since $0 < v_2 < v_1 + v_2$, the minimiser satisfies $0 < w^* < 1$, so it lies in the admissible range $[0, 1]$.

The weight on survey 2 is $1 - w^* = \dfrac{v_1}{v_1+v_2}$, so

$$\frac{w^*}{1-w^*} = \frac{v_2}{v_1}.$$

Thus $w^* > 1 - w^*$ exactly when $v_1 < v_2$: the survey with the smaller variance receives the larger weight. For example, with $v_1 = 1$ and $v_2 = 3$ (hours^2), $w^* = 3/4$, so the more precise survey 1 gets weight 0.75. The minimised variance is

$$\begin{aligned}V(w^*) &= \frac{v_2^2}{(v_1+v_2)^2}\, v_1 + \frac{v_1^2}{(v_1+v_2)^2}\, v_2 \\ &= \frac{v_1 v_2 (v_2+v_1)}{(v_1+v_2)^2} \\ &= \frac{v_1 v_2}{v_1+v_2},\end{aligned}$$

which in the example is $3/4$ hours^2, below both $v_1 = 1$ and $v_2 = 3$. Therefore

$$\boxed{w^* = \frac{v_2}{v_1+v_2}.}$$

Method 2: Deviations from μ and completing the square

(a) Write $\mu = w\mu + (1-w)\mu$ and subtract it from $\hat{\mu}_w$, so that the estimation error splits into the two surveys' errors:

$$\begin{aligned}\hat{\mu}_w - \mu &= w\hat{\mu}_1 + (1-w)\hat{\mu}_2 - w\mu - (1-w)\mu \\ &= w(\hat{\mu}_1 - \mu) + (1-w)(\hat{\mu}_2 - \mu).\end{aligned}$$

Each survey error has mean zero, because $E[\hat{\mu}_k - \mu] = E[\hat{\mu}_k] - \mu = \mu - \mu = 0$ for $k = 1, 2$. Taking expectations,

$$\begin{aligned}E[\hat{\mu}_w - \mu] &= wE[\hat{\mu}_1 - \mu] + (1-w)E[\hat{\mu}_2 - \mu] \\ &= w \cdot 0 + (1-w) \cdot 0 \\ &= 0,\end{aligned}$$

so the bias is zero for every w, i.e.

$$\boxed{E[\hat{\mu}_w] = \mu \text{ for all } w \in [0,1].}$$

(b) Let $d_1 = \hat{\mu}_1 - \mu$ and $d_2 = \hat{\mu}_2 - \mu$. By part (a), $E[\hat{\mu}_w] = \mu$, so the definition of variance gives $\mathrm{Var}(\hat{\mu}_w) = E[(\hat{\mu}_w - \mu)^2]$. Expand the square:

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_w) &= E\big[(wd_1 + (1-w)d_2)^2\big] \\ &= E\big[w^2 d_1^2 + 2w(1-w)d_1 d_2 + (1-w)^2 d_2^2\big] \\ &= w^2 E[d_1^2] + 2w(1-w)E[d_1 d_2] + (1-w)^2 E[d_2^2]. \qquad \text{(linearity)}\end{aligned}$$

Evaluate each expectation. Since $E[\hat{\mu}_k] = \mu$, $E[d_k^2] = E[(\hat{\mu}_k - E[\hat{\mu}_k])^2] = \mathrm{Var}(\hat{\mu}_k)$, so $E[d_1^2] = v_1$ and $E[d_2^2] = v_2$. The errors d_1, d_2 are functions of independent surveys, so they are independent:

$$E[d_1 d_2] = E[d_1]\,E[d_2] = 0 \cdot 0 = 0.$$

Substituting the three values,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_w) &= w^2 v_1 + 2w(1-w) \cdot 0 + (1-w)^2 v_2 \\ &= w^2 v_1 + (1-w)^2 v_2,\end{aligned}$$

so

$$\boxed{\mathrm{Var}(\hat{\mu}_w) = w^2 v_1 + (1-w)^2 v_2.}$$

(c) Expand $V(w) = w^2 v_1 + (1-w)^2 v_2$ as a quadratic in w:

$$\begin{aligned}V(w) &= w^2 v_1 + (1 - 2w + w^2)v_2 \\ &= (v_1 + v_2)w^2 - 2v_2 w + v_2.\end{aligned}$$

Factor out the leading coefficient $v_1 + v_2 > 0$ from the first two terms and complete the square:

$$\begin{aligned}V(w) &= (v_1 + v_2)\Big[w^2 - \frac{2v_2}{v_1 + v_2}\,w\Big] + v_2 \\ &= (v_1 + v_2)\Big[\Big(w - \frac{v_2}{v_1 + v_2}\Big)^2 - \frac{v_2^2}{(v_1 + v_2)^2}\Big] + v_2 \\ &= (v_1 + v_2)\Big(w - \frac{v_2}{v_1 + v_2}\Big)^2 - \frac{v_2^2}{v_1 + v_2} + v_2\end{aligned}$$

$$= (v_1 + v_2)\Big(w - \frac{v_2}{v_1 + v_2}\Big)^2 + \frac{v_2(v_1 + v_2) - v_2^2}{v_1 + v_2}$$
$$= (v_1 + v_2)\Big(w - \frac{v_2}{v_1 + v_2}\Big)^2 + \frac{v_1 v_2}{v_1 + v_2}.$$

The squared term is non-negative and its coefficient $v_1 + v_2$ is positive, so

$$V(w) \geq \frac{v_1 v_2}{v_1 + v_2} \quad \text{for every } w,$$

with equality if and only if the square is zero, i.e. $w = \dfrac{v_2}{v_1 + v_2}$. This value lies in $(0, 1)$ because $0 < v_2 < v_1 + v_2$, and it exceeds $\frac{1}{2}$ exactly when $v_2 > v_1$, so the more precise survey gets the larger weight. Therefore

$$\boxed{w^* = \frac{v_2}{v_1 + v_2}.}$$

Unofficial QRS@NTU - qrsntu.org

Method 3: Constrained minimisation over linear combinations (Lagrange multiplier)

This method should be used only if the course allows this interpretation.

(a) Consider any linear combination $\tilde{\mu} = a\hat{\mu}_1 + b\hat{\mu}_2$ with constants a, b. By linearity,

$$\begin{aligned} E[\tilde{\mu}] &= aE[\hat{\mu}_1] + bE[\hat{\mu}_2] \\ &= a\mu + b\mu \\ &= (a+b)\mu. \end{aligned}$$

Hence $\tilde{\mu}$ is unbiased for every value of μ if and only if $a + b = 1$. The estimator $\hat{\mu}_w$ has $a = w$ and $b = 1 - w$, and

$$a + b = w + (1 - w) = 1,$$

so the unbiasedness constraint holds for every w:

$$\boxed{E[\hat{\mu}_w] = \mu \text{ for all } w \in [0, 1].}$$

(b) Write the variance as the covariance of $\tilde{\mu}$ with itself and expand by bilinearity of Cov:

$$\begin{aligned} \mathrm{Var}(\tilde{\mu}) &= \mathrm{Cov}(a\hat{\mu}_1 + b\hat{\mu}_2,\ a\hat{\mu}_1 + b\hat{\mu}_2) \\ &= a^2\,\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_1) + ab\,\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) \\ &\quad + ba\,\mathrm{Cov}(\hat{\mu}_2, \hat{\mu}_1) + b^2\,\mathrm{Cov}(\hat{\mu}_2, \hat{\mu}_2) \\ &= a^2 v_1 + 2ab \cdot 0 + b^2 v_2 \qquad \text{(independence)} \\ &= a^2 v_1 + b^2 v_2. \end{aligned}$$

Setting $a = w$ and $b = 1 - w$,

$$\boxed{\mathrm{Var}(\hat{\mu}_w) = w^2 v_1 + (1-w)^2 v_2.}$$

(c) Minimise $a^2 v_1 + b^2 v_2$ subject to the unbiasedness constraint $a + b = 1$. The Lagrangian is

$$\mathcal{L}(a, b, \lambda) = a^2 v_1 + b^2 v_2 - \lambda(a + b - 1).$$

The first-order conditions are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial a} &= 2av_1 - \lambda = 0 && \Longrightarrow\ a = \frac{\lambda}{2v_1}, \\ \frac{\partial \mathcal{L}}{\partial b} &= 2bv_2 - \lambda = 0 && \Longrightarrow\ b = \frac{\lambda}{2v_2}, \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= -(a + b - 1) = 0 && \Longrightarrow\ a + b = 1. \end{aligned}$$

Substitute the first two into the constraint:

$$\begin{aligned} \frac{\lambda}{2}\left(\frac{1}{v_1} + \frac{1}{v_2}\right) &= 1 \\ \frac{\lambda}{2} \cdot \frac{v_1 + v_2}{v_1 v_2} &= 1 \\ \frac{\lambda}{2} &= \frac{v_1 v_2}{v_1 + v_2}. \end{aligned}$$

Unofficial QRS@NTU study aid

Hence

$$a = \frac{1}{v_1} \cdot \frac{v_1 v_2}{v_1 + v_2} = \frac{v_2}{v_1 + v_2}, \qquad b = \frac{1}{v_2} \cdot \frac{v_1 v_2}{v_1 + v_2} = \frac{v_1}{v_1 + v_2}.$$

Equivalently $a = \dfrac{1/v_1}{1/v_1 + 1/v_2}$: each survey is weighted by its precision $1/v_k$. The objective is a sum of squares with positive coefficients (strictly convex) and the constraint is linear, so this stationary point is the global minimum. Since $\hat{\mu}_w$ uses $a = w$,

$$\boxed{w^* = \frac{v_2}{v_1 + v_2}.}$$

Marking scheme (10 marks)

(a) **(3 marks)** Unbiasedness

- 1 mark: correctly applies linearity of expectation to $\hat{\mu}_w$.
- 1 mark: correctly substitutes $E[\hat{\mu}_1] = E[\hat{\mu}_2] = \mu$.
- 1 mark: correct conclusion $E[\hat{\mu}_w] = \mu$ for every w.

(b) **(4 marks)** Variance

- 1 mark: correctly expands $\mathrm{Var}(w\hat{\mu}_1 + (1-w)\hat{\mu}_2)$ including the covariance term.
- 1 mark: correctly sets $\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) = 0$ by independence.
- 1 mark: correct use of $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$ for both terms.
- 1 mark: correct final expression $\mathrm{Var}(\hat{\mu}_w) = w^2 v_1 + (1-w)^2 v_2$.

(c) **(3 marks)** Optimal weight

- 1 mark: correct first-order condition (or completed-square form).
- 1 mark: correct algebra solving for w^*.
- 1 mark: correct final answer $w^* = v_2/(v_1 + v_2)$ with a valid remark on precision-weighting (or second-order/positivity check).

Question 2 (10 Marks)

PUB (Singapore's national water agency) commissions a study of $n = 200$ households and regresses monthly water consumption (in cubic metres, m^3) on household size (number of residents). The estimated regression is

$$\widehat{\text{Water}}_i = 3.8 + 5.4\,\text{Size}_i, \qquad \text{SE}(\hat{\beta}_1) = 1.2.$$

(a) (3 marks) What is the regression's predicted monthly water consumption for a household of 5 residents?

(b) (2 marks) A household grows from 3 to 6 residents. What is the regression's predicted change in monthly water consumption?

(c) (3 marks) For this regression, TSS = 5000 and SSR = 1500 (in units of $(\text{m}^3)^2$). Compute ESS and R^2, and interpret R^2 in words.

(d) (2 marks) Interpret the slope coefficient with units, and comment on whether the intercept is economically meaningful here.

Solution

Method 1: Substitution into the fitted line and ESS = TSS − SSR

(a) The fitted line is $\widehat{\text{Water}} = \hat{\beta}_0 + \hat{\beta}_1\,\text{Size}$ with $\hat{\beta}_0 = 3.8$ and $\hat{\beta}_1 = 5.4$. Substitute Size = 5:

$$\begin{aligned}\widehat{\text{Water}} &= 3.8 + 5.4(5)\\ &= 3.8 + 27.0\\ &= 30.8.\end{aligned}$$

Both coefficients are in m^3/month ($\hat{\beta}_1$ per resident), so

$$\boxed{\widehat{\text{Water}} = 30.8\ \text{m}^3/\text{month}.}$$

(b) Compute the fitted value at each household size and subtract. At Size = 6:

$$\begin{aligned}\widehat{\text{Water}}(6) &= 3.8 + 5.4(6)\\ &= 3.8 + 32.4\\ &= 36.2.\end{aligned}$$

At Size = 3:

$$\begin{aligned}\widehat{\text{Water}}(3) &= 3.8 + 5.4(3)\\ &= 3.8 + 16.2\\ &= 20.0.\end{aligned}$$

The predicted change is

$$\begin{aligned}\Delta\widehat{\text{Water}} &= \widehat{\text{Water}}(6) - \widehat{\text{Water}}(3)\\ &= 36.2 - 20.0\\ &= 16.2,\end{aligned}$$

a positive number, so the prediction rises:

$$\boxed{\Delta\widehat{\text{Water}} = 16.2\ \text{m}^3/\text{month (an increase)}.}$$

(c) The sums of squares satisfy TSS = ESS + SSR (Lecture 4). Solve for ESS and substitute:

$$\begin{aligned} \text{ESS} &= \text{TSS} - \text{SSR} \\ &= 5000 - 1500 \\ &= 3500 \ (\text{m}^3)^2. \end{aligned}$$

By definition $R^2 = \text{ESS}/\text{TSS}$:

$$\begin{aligned} R^2 &= \frac{\text{ESS}}{\text{TSS}} \\ &= \frac{3500}{5000} \\ &= 0.70. \end{aligned}$$

Interpretation of R^2:

- $R^2 = 0.70$ means that 70% of the sample variation of monthly water consumption around its sample mean (the TSS of 5000 $(\text{m}^3)^2$) is accounted for by the fitted line in household size.
- The remaining share SSR/TSS = 1500/5000 = 0.30, i.e. 30%, is left in the residuals $\hat{u}_i$: it reflects factors other than household size, such as appliances, gardens or washing habits.
- R^2 is a sample measure of fit for these $n = 200$ households; it does not say whether $\hat{\beta}_1$ is a causal effect, which is a Least Squares Assumption 1 question.

Therefore

$$\boxed{\text{ESS} = 3500 \ (\text{m}^3)^2, \qquad R^2 = 0.70.}$$

(d) Interpretation of the two coefficients:

- The slope $\hat{\beta}_1 = 5.4$ means that, comparing two households in this sample that differ by one resident, the regression predicts 5.4 m^3 more water per month for the larger household. For example, the prediction is $3.8 + 5.4(4) = 25.4$ m^3/month at 4 residents and 30.8 m^3/month at 5 residents, a gap of 5.4 m^3.
- $\hat{\beta}_1$ is a sample estimate of the population slope β_1, the change in mean monthly consumption $E[\text{Water} \mid \text{Size}]$ per additional resident.
- The intercept $\hat{\beta}_0 = 3.8$ m^3/month is, by construction, the prediction at Size = 0.
- No household has zero residents, so Size = 0 lies outside the range of the data; the value 3.8 is an extrapolation that only fixes the height of the line, and it has no economic meaning as "water used by an empty household".

Therefore

$$\boxed{\text{Slope: } +5.4 \ \text{m}^3\text{/month per resident; intercept not economically meaningful.}}$$

Method 2: Slope times change in Size and $R^2 = 1 - \text{SSR}/\text{TSS}$

Parts (a) and (d) read the fitted line directly and have the single route of Method 1; this method re-derives parts (b) and (c) with a different technique.

(b) Write $\widehat{\text{Water}}(s) = \hat{\beta}_0 + \hat{\beta}_1 s$ for the fitted value at size s. For any two sizes s_1, s_2 the intercept cancels:

$$\begin{aligned}\widehat{\text{Water}}(s_2) - \widehat{\text{Water}}(s_1) &= (\hat{\beta}_0 + \hat{\beta}_1 s_2) - (\hat{\beta}_0 + \hat{\beta}_1 s_1) \\ &= \hat{\beta}_1 (s_2 - s_1).\end{aligned}$$

With $s_1 = 3$, $s_2 = 6$ and $\hat{\beta}_1 = 5.4$:

$$\begin{aligned}\Delta\widehat{\text{Water}} &= 5.4\,(6 - 3) \\ &= 5.4 \times 3 \\ &= 16.2,\end{aligned}$$

so

$$\boxed{\Delta\widehat{\text{Water}} = 16.2 \text{ m}^3\text{/month (an increase).}}$$

(c) Divide the identity $\text{TSS} = \text{ESS} + \text{SSR}$ by $\text{TSS} > 0$:

$$\begin{aligned}1 &= \frac{\text{ESS}}{\text{TSS}} + \frac{\text{SSR}}{\text{TSS}} \\ &= R^2 + \frac{\text{SSR}}{\text{TSS}},\end{aligned}$$

so $R^2 = 1 - \text{SSR}/\text{TSS}$. Substitute the given sums:

$$\begin{aligned}R^2 &= 1 - \frac{1500}{5000} \\ &= 1 - 0.30 \\ &= 0.70.\end{aligned}$$

Recover ESS from the definition $R^2 = \text{ESS}/\text{TSS}$:

$$\begin{aligned}\text{ESS} &= R^2 \times \text{TSS} \\ &= 0.70 \times 5000 \\ &= 3500 \ (\text{m}^3)^2.\end{aligned}$$

Here the unexplained share $\text{SSR}/\text{TSS} = 0.30$ is computed first, and $R^2 = 0.70$ is its complement: household size accounts for 70% of the sample variation in monthly water use. Therefore

$$\boxed{\text{ESS} = 3500 \ (\text{m}^3)^2, \qquad R^2 = 0.70.}$$

Method 3: R^2 as the squared correlation between Y_i and $\hat{Y}_i$

This method re-derives part (c). Write $Y_i = \text{Water}_i$ and $\hat{Y}_i = 3.8 + 5.4\,\text{Size}_i$, so that $Y_i = \hat{Y}_i + \hat{u}_i$.

(c) The OLS first-order conditions give $\sum_{i=1}^{n} \hat{u}_i = 0$ and $\sum_{i=1}^{n} \hat{u}_i \, \text{Size}_i = 0$. Hence the fitted values have the same mean as Y_i:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{u}_i) = \bar{Y} - \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = \bar{Y},$$

and the residuals are orthogonal to the fitted values:

$$\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i \, \text{Size}_i = \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \cdot 0 = 0.$$

First, as in Method 1, $\text{ESS} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \text{TSS} - \text{SSR} = 5000 - 1500 = 3500$. Next, the cross-product of deviations equals ESS:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y} + \hat{u}_i)(\hat{Y}_i - \bar{Y}) \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i \hat{Y}_i - \bar{Y} \sum_{i=1}^{n} \hat{u}_i \\
&= \text{ESS} + 0 - \bar{Y} \cdot 0 \\
&= 3500.
\end{aligned}$$

With $n - 1 = 199$, the sample covariance and variances are

$$\begin{aligned}
s_{Y\hat{Y}} &= \frac{3500}{199} = 17.588, \\
s_Y^2 &= \frac{\text{TSS}}{199} = \frac{5000}{199} = 25.126, \\
s_{\hat{Y}}^2 &= \frac{\text{ESS}}{199} = \frac{3500}{199} = 17.588.
\end{aligned}$$

The sample correlation is

$$\begin{aligned}
r_{Y\hat{Y}} &= \frac{s_{Y\hat{Y}}}{s_Y \, s_{\hat{Y}}} \\
&= \frac{3500/199}{\sqrt{(5000/199)(3500/199)}} \\
&= \frac{3500}{\sqrt{5000 \times 3500}} \\
&= \frac{3500}{4183.30} \\
&= 0.8367,
\end{aligned}$$

and squaring it,

$$\begin{aligned}
r_{Y\hat{Y}}^2 &= \frac{3500^2}{5000 \times 3500} \\
&= \frac{3500}{5000} \\
&= 0.70.
\end{aligned}$$

Since $\hat{Y}_i$ is an increasing linear function of Size_i ($\hat{\beta}_1 = 5.4 > 0$), this is also the correlation between water use and household size: a strong positive association, with size accounting for 70% of the sample variation. Therefore

$$\boxed{\text{ESS} = 3500 \ (\text{m}^3)^2, \qquad R^2 = 0.70.}$$

Marking scheme (10 marks)

(a) **(3 marks)** Predicted consumption

- 1 mark: correct substitution.
- 1 mark: correct arithmetic.
- 1 mark: correct units.

(b) **(2 marks)** Predicted change

- 1 mark: correct method/arithmetic (5.4×3).
- 1 mark: correctly stated with units and direction.

(c) **(3 marks)** ESS and R^2

- 1 mark: correct use of TSS = ESS + SSR to get ESS = 3500.
- 1 mark: correct $R^2 = \text{ESS}/\text{TSS} = 0.70$.
- 1 mark: correct interpretation of R^2 as the share of variation in water consumption explained by household size.

(d) **(2 marks)** Interpretation

- 1 mark: correct slope interpretation with units.
- 1 mark: correct comment on the intercept's lack of economic meaning.

Unofficial QRS@NTU (qrs.ntu.edu.sg)

Question 3 (10 Marks)

Two independent surveys estimate average weekly online-shopping spending (in SGD) in two regions of a large Southeast Asian city. Region A: sample size $n_A = 100$, sample mean $\bar{Y}_A$, with $\text{Var}(\bar{Y}_A) = 16$. Region B: sample size $n_B = 150$, sample mean $\bar{Y}_B$, with $\text{Var}(\bar{Y}_B) = 9$. The two samples are drawn independently, and the true population means satisfy $\mu_A - \mu_B = 10$.

(a) (3 marks) Derive $E[\bar{Y}_A - \bar{Y}_B]$ and $\text{Var}(\bar{Y}_A - \bar{Y}_B)$, using the independence of the two samples.

(b) (4 marks) In large samples, $\bar{Y}_A - \bar{Y}_B$ is approximately normally distributed. Standardise and use the standard normal table in the Overview to compute $P(\bar{Y}_A - \bar{Y}_B > 17.5)$.

(c) (3 marks) Suppose both sample sizes are quadrupled ($n'_A = 400$, $n'_B = 600$) while the population variances of individual spending stay the same. Explain what happens to $\text{Var}(\bar{Y}_A)$ and $\text{Var}(\bar{Y}_B)$, and hence find the new standard deviation of $\bar{Y}_A - \bar{Y}_B$.

Solution

Method 1: Expectation and variance rules for the two sample means

(a) Each sample mean is unbiased for its population mean, $E[\bar{Y}_A] = \mu_A$ and $E[\bar{Y}_B] = \mu_B$. By linearity of expectation,

$$\begin{aligned} E[\bar{Y}_A - \bar{Y}_B] &= E[\bar{Y}_A] - E[\bar{Y}_B] && \text{(linearity)} \\ &= \mu_A - \mu_B \\ &= 10. \end{aligned}$$

For the variance, apply $\text{Var}(aX + bY) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y)$ with $a = 1$, $b = -1$:

$$\begin{aligned} \text{Var}(\bar{Y}_A - \bar{Y}_B) &= 1^2\,\text{Var}(\bar{Y}_A) + (-1)^2\,\text{Var}(\bar{Y}_B) + 2(1)(-1)\,\text{Cov}(\bar{Y}_A, \bar{Y}_B) \\ &= 16 + 9 - 2\,\text{Cov}(\bar{Y}_A, \bar{Y}_B). \end{aligned}$$

The two samples are drawn independently, so $\bar{Y}_A$ and $\bar{Y}_B$ are independent, and independence implies

$$\text{Cov}(\bar{Y}_A, \bar{Y}_B) = 0.$$

Hence

$$\begin{aligned} \text{Var}(\bar{Y}_A - \bar{Y}_B) &= 16 + 9 - 2 \cdot 0 \\ &= 25. \end{aligned}$$

Therefore (in SGD and SGD^2 respectively)

$$\boxed{E[\bar{Y}_A - \bar{Y}_B] = 10, \qquad \text{Var}(\bar{Y}_A - \bar{Y}_B) = 25.}$$

(b) From part (a), the standard deviation of the difference is

$$\text{s.d.}(\bar{Y}_A - \bar{Y}_B) = \sqrt{25} = 5 \text{ SGD}.$$

In large samples, $\bar{Y}_A - \bar{Y}_B \approx N(10, 25)$, so

$$Z = \frac{(\bar{Y}_A - \bar{Y}_B) - 10}{5} \approx N(0, 1).$$

Subtracting 10 and dividing by $5 > 0$ on both sides of the inequality does not change the event:

$$\begin{aligned} P(\bar{Y}_A - \bar{Y}_B > 17.5) &= P\Big(\frac{(\bar{Y}_A - \bar{Y}_B) - 10}{5} > \frac{17.5 - 10}{5}\Big) \\ &= P\Big(Z > \frac{7.5}{5}\Big) \\ &= P(Z > 1.5) \\ &= 1 - \Phi(1.5) \\ &= 1 - 0.9332 \qquad \text{(table)} \\ &= 0.0668. \end{aligned}$$

So when the true regional gap is SGD 10 per week, there is about a 6.7% chance that the two surveys report a gap above SGD 17.5, i.e. more than 1.5 standard deviations above its mean. Therefore

$$\boxed{P(\bar{Y}_A - \bar{Y}_B > 17.5) = 0.0668.}$$

(c) For a sample mean of n independent draws with population variance σ^2, $\text{Var}(\bar{Y}) = \sigma^2/n$. First recover the population variances of individual spending from the original samples:

$$\begin{aligned} \sigma_A^2 &= n_A \,\text{Var}(\bar{Y}_A) = 100 \times 16 = 1600 \text{ SGD}^2, \\ \sigma_B^2 &= n_B \,\text{Var}(\bar{Y}_B) = 150 \times 9 = 1350 \text{ SGD}^2. \end{aligned}$$

These stay the same; only the sample sizes change. With $n'_A = 400$ and $n'_B = 600$:

$$\begin{aligned} \text{Var}(\bar{Y}'_A) &= \frac{\sigma_A^2}{n'_A} = \frac{1600}{400} = 4, \\ \text{Var}(\bar{Y}'_B) &= \frac{\sigma_B^2}{n'_B} = \frac{1350}{600} = 2.25. \end{aligned}$$

Each variance is one quarter of its original value ($16/4 = 4$, $9/4 = 2.25$), because the denominator n is four times larger while σ^2 is unchanged. The new samples are still independent, so as in part (a)

$$\begin{aligned} \text{Var}(\bar{Y}'_A - \bar{Y}'_B) &= \text{Var}(\bar{Y}'_A) + \text{Var}(\bar{Y}'_B) - 2 \cdot 0 \\ &= 4 + 2.25 \\ &= 6.25 \text{ SGD}^2, \end{aligned}$$

and

$$\begin{aligned} \text{s.d.}(\bar{Y}'_A - \bar{Y}'_B) &= \sqrt{6.25} \\ &= 2.5 \text{ SGD}. \end{aligned}$$

The variance falls by the factor 4 and the standard deviation by the factor $\sqrt{4} = 2$, from 5 to 2.5 SGD: four times the data halves the typical sampling error of the estimated gap. Therefore

$$\boxed{\text{New } \text{Var}(\bar{Y}'_A - \bar{Y}'_B) = 6.25, \quad \text{new s.d.} = 2.5 \text{ (half the original s.d. of 5).}}$$

Method 2: Sampling formula from the individual observations and $n \to 4n$ scaling

Let $Y_{A,1}, \ldots, Y_{A,n_A}$ be i.i.d. with mean μ_A and variance σ_A^2, and $Y_{B,1}, \ldots, Y_{B,n_B}$ be i.i.d. with mean μ_B and variance σ_B^2, the two samples independent of each other.

(a) Write the difference as one weighted sum of all $n_A + n_B = 250$ observations:

$$\bar{Y}_A - \bar{Y}_B = \sum_{i=1}^{n_A} \frac{1}{n_A} Y_{A,i} - \sum_{j=1}^{n_B} \frac{1}{n_B} Y_{B,j}.$$

Taking expectations term by term,

$$\begin{aligned} E[\bar{Y}_A - \bar{Y}_B] &= \sum_{i=1}^{n_A} \frac{1}{n_A}\mu_A - \sum_{j=1}^{n_B} \frac{1}{n_B}\mu_B \\ &= n_A \cdot \frac{1}{n_A}\mu_A - n_B \cdot \frac{1}{n_B}\mu_B \\ &= \mu_A - \mu_B \\ &= 10. \end{aligned}$$

All 250 observations are mutually independent, so the variance of the weighted sum is the sum of (weight)$^2\times$(variance):

$$\begin{aligned} \mathrm{Var}(\bar{Y}_A - \bar{Y}_B) &= \sum_{i=1}^{n_A} \Big(\frac{1}{n_A}\Big)^2 \sigma_A^2 + \sum_{j=1}^{n_B} \Big(-\frac{1}{n_B}\Big)^2 \sigma_B^2 \\ &= n_A \cdot \frac{\sigma_A^2}{n_A^2} + n_B \cdot \frac{\sigma_B^2}{n_B^2} \\ &= \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}. \end{aligned}$$

The same formula with one sample gives $\mathrm{Var}(\bar{Y}_A) = \sigma_A^2/n_A = 16$ and $\mathrm{Var}(\bar{Y}_B) = \sigma_B^2/n_B = 9$, i.e. $\sigma_A^2 = 1600$ and $\sigma_B^2 = 1350$. Substituting,

$$\begin{aligned} \mathrm{Var}(\bar{Y}_A - \bar{Y}_B) &= \frac{1600}{100} + \frac{1350}{150} \\ &= 16 + 9 \\ &= 25. \end{aligned}$$

Therefore

$$\boxed{E[\bar{Y}_A - \bar{Y}_B] = 10, \qquad \mathrm{Var}(\bar{Y}_A - \bar{Y}_B) = 25.}$$

(b) By the central limit theorem, each of $\bar{Y}_A$ and $\bar{Y}_B$ is approximately normal in large samples ($n_A = 100$, $n_B = 150$ i.i.d. draws with finite variances). The two sample means are independent, and a difference of independent normal variables is normal, with the mean and variance of part (a):

$$\bar{Y}_A - \bar{Y}_B \approx N\Big(\mu_A - \mu_B,\ \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\Big) = N(10, 25).$$

Standardise with mean 10 and standard deviation $\sqrt{25} = 5$, then use the symmetry of $N(0,1)$: Z and $-Z$ have the same distribution, so an upper-tail probability equals the matching lower-tail probability.

$$P(\bar{Y}_A - \bar{Y}_B > 17.5) = P\Big(Z > \frac{17.5 - 10}{5}\Big)$$

$$
\begin{aligned}
&= P(Z > 1.5) \\
&= P(-Z < -1.5) \\
&= P(Z < -1.5) && \text{(symmetry)} \\
&= \Phi(-1.5) \\
&= 1 - \Phi(1.5) \\
&= 1 - 0.9332 \\
&= 0.0668.
\end{aligned}
$$

Therefore

$$\boxed{P(\bar{Y}_A - \bar{Y}_B > 17.5) = 0.0668.}$$

(c) Replace n_A, n_B by $4n_A, 4n_B$ in the formula from part (a), keeping σ_A^2, σ_B^2 fixed:

$$
\begin{aligned}
\text{Var}(\bar{Y}'_A - \bar{Y}'_B) &= \frac{\sigma_A^2}{4n_A} + \frac{\sigma_B^2}{4n_B} \\
&= \frac{1}{4}\Big(\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\Big) \\
&= \frac{1}{4} \times 25 \\
&= 6.25.
\end{aligned}
$$

Term by term, $\text{Var}(\bar{Y}'_A) = 16/4 = 4$ and $\text{Var}(\bar{Y}'_B) = 9/4 = 2.25$, each divided by 4. Taking the square root,

$$
\begin{aligned}
\text{s.d.}(\bar{Y}'_A - \bar{Y}'_B) &= \sqrt{\tfrac{1}{4} \times 25} \\
&= \tfrac{1}{2} \times 5 \\
&= 2.5,
\end{aligned}
$$

so the standard deviation scales as $1/\sqrt{n}$ and halves when n is quadrupled. Therefore

$$\boxed{\text{New Var}(\bar{Y}'_A - \bar{Y}'_B) = 6.25, \quad \text{new s.d.} = 2.5 \text{ (half the original s.d. of 5).}}$$

Marking scheme (10 marks)

(a) **(3 marks)** Mean and variance

- 1 mark: correct $E[\bar{Y}_A - \bar{Y}_B] = 10$ by linearity of expectation.
- 1 mark: correctly sets $\text{Cov}(\bar{Y}_A, \bar{Y}_B) = 0$ by independence.
- 1 mark: correct final variance 25.

(b) **(4 marks)** Probability

- 1 mark: correct standard deviation $\sqrt{25} = 5$.
- 1 mark: correct standardisation to $z = 1.5$.
- 1 mark: correctly reads $\Phi(1.5) = 0.9332$ from the table.
- 1 mark: correct final probability 0.0668.

(c) **(3 marks)** Effect of quadrupling n

- 1 mark: correctly identifies $\text{Var}(\bar{Y}_A)$, $\text{Var}(\bar{Y}_B)$ each fall by a factor of 4.
- 1 mark: correct new variance of the difference, 6.25.
- 1 mark: correct new standard deviation 2.5, with correct link to it being half the original.

Question 4 (10 Marks)

A study of $n = 160$ hawker stalls in Singapore regresses average daily revenue (in SGD) on D_i, a dummy variable equal to 1 if the stall accepts e-payment and 0 if it is cash-only ($n_0 = 60$ cash-only, $n_1 = 100$ e-payment). The estimated regression is

$$\widehat{\text{Revenue}}_i = 520 + 85\, D_i, \qquad \text{SE}(\hat{\beta}_1) = 28.$$

(a) (3 marks) State the average daily revenue predicted by the regression for cash-only stalls and for e-payment stalls.

(b) (2 marks) Give an economic interpretation of $\hat{\beta}_1 = 85$.

(c) (3 marks) The platform's data team re-estimates the model to control for the number of hours each stall operates per day, Hours_i, assuming a *common slope* on Hours_i for both groups (i.e. the effect of an extra hour of operation is the same for cash-only and e-payment stalls):

$$\widehat{\text{Revenue}}_i = 150 + 85\, D_i + 45\, \text{Hours}_i.$$

Write the separate fitted lines for cash-only stalls ($D = 0$) and e-payment stalls ($D = 1$) as functions of Hours_i. State which group is the *base group*, and give an economic interpretation of the coefficient 85 on D_i in this extended model.

(d) (2 marks) Explain the *dummy variable trap*: why can the regression in part (c) *not* also include a second dummy $C_i = 1 - D_i$ (an indicator for cash-only) alongside D_i and the intercept? State the general rule for how many dummy variables to include for a categorical variable with g mutually exclusive groups.

Solution

Method 1: Substituting $D = 0$ and $D = 1$ into the fitted line

(a) Substitute each value of the dummy into $\widehat{\text{Revenue}} = 520 + 85\, D$. For cash-only stalls ($D = 0$):

$$\begin{aligned}\widehat{\text{Revenue}} &= 520 + 85(0) \\ &= 520.\end{aligned}$$

For e-payment stalls ($D = 1$):

$$\begin{aligned}\widehat{\text{Revenue}} &= 520 + 85(1) \\ &= 605.\end{aligned}$$

With a single binary regressor, the fitted value within each group equals that group's sample average (Lecture 5), so these are the average daily revenues of the 60 cash-only and 100 e-payment stalls. Therefore

Cash-only: SGD 520/day; E-payment: SGD 605/day.

(b) From part (a), the slope is the gap between the two predicted group averages:

$$\begin{aligned}\hat{\beta}_1 &= \widehat{\text{Revenue}}(D = 1) - \widehat{\text{Revenue}}(D = 0) \\ &= 605 - 520 \\ &= 85.\end{aligned}$$

Interpretation:

- Because D_i is binary, $\hat{\beta}_1 = 85$ is the difference in sample mean daily revenue between e-payment stalls and cash-only stalls: SGD 85 per day, i.e. $85/520 = 16.3\%$ above the cash-only average.
- It estimates the population difference $\beta_1 = E[\text{Revenue} \mid D = 1] - E[\text{Revenue} \mid D = 0]$.
- It is a comparison of group averages, not automatically the causal effect of adopting e-payment: stalls that adopt e-payment may differ in location or hours, which is a Least Squares Assumption 1 question.

Therefore

E-payment stalls earn, on average, SGD 85/day more than cash-only stalls.

(c) Substitute $D = 0$ into $\widehat{\text{Revenue}} = 150 + 85\,D + 45\,\text{Hours}$:

$$\begin{aligned}\widehat{\text{Revenue}} &= 150 + 85(0) + 45\,\text{Hours} \\ &= 150 + 45\,\text{Hours}.\end{aligned}$$

Substitute $D = 1$:

$$\begin{aligned}\widehat{\text{Revenue}} &= 150 + 85(1) + 45\,\text{Hours} \\ &= (150 + 85) + 45\,\text{Hours} \\ &= 235 + 45\,\text{Hours}.\end{aligned}$$

Interpretation:

- The base group is the group with $D_i = 0$, the cash-only stalls: its line uses the regression intercept 150 directly.
- Both lines have slope 45: in either group, one extra hour of operation per day is associated with SGD 45 more predicted daily revenue.
- The coefficient 85 on D_i is the vertical gap between the two parallel lines: holding hours of operation fixed, e-payment stalls are predicted to earn SGD 85 per day more than cash-only stalls. For example, at 8 hours the predictions are $150 + 45(8) = 510$ and $235 + 45(8) = 595$, a gap of 85.

Therefore

Cash-only: $\widehat{\text{Revenue}} = 150 + 45\,\text{Hours}$
E-payment: $\widehat{\text{Revenue}} = 235 + 45\,\text{Hours}$

(d) Explanation:

- Every stall is either e-payment or cash-only, so for every i

$$D_i + C_i = D_i + (1 - D_i) = 1.$$

- The intercept is the coefficient on a regressor equal to 1 for every stall, so $C_i = 1 - D_i$ is an exact linear function of the intercept regressor and D_i: the regressors are perfectly multicollinear.
- With perfect multicollinearity OLS cannot separate the intercept from the coefficient on C_i: many different coefficient vectors give the same fitted values, so there is no unique OLS solution (software drops one regressor).

- General rule: for a categorical variable with g mutually exclusive groups, include $g-1$ dummies together with the intercept; the omitted group is the base group, and each dummy's coefficient is measured relative to it. Here $g=2$, so one dummy (D_i).

Therefore

$$D_i + C_i = 1 \text{ for all } i \;\Rightarrow\; \text{exact collinearity with the intercept; use only } g-1 \text{ dummies for } g \text{ groups.}$$

Unofficial - QRS@NTU - qrsntu.org

Method 2: OLS formula with group means and coefficient matching

Let $\bar{Y}_0$ and $\bar{Y}_1$ be the sample mean revenues of the $n_0 = 60$ cash-only and $n_1 = 100$ e-payment stalls, $n = 160$, and $Y_i = \text{Revenue}_i$.

(a) The OLS formulas are $\hat{\beta}_1 = \sum_{i=1}^{n}(D_i - \bar{D})(Y_i - \bar{Y}) / \sum_{i=1}^{n}(D_i - \bar{D})^2$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{D}$. The share of e-payment stalls is

$$\bar{D} = \frac{n_1}{n} = \frac{100}{160} = 0.625, \qquad 1 - \bar{D} = \frac{n_0}{n} = 0.375.$$

The denominator splits by group ($D_i = 1$ gives $1 - \bar{D}$, $D_i = 0$ gives $-\bar{D}$):

$$\begin{aligned}\sum_{i=1}^{n}(D_i - \bar{D})^2 &= n_1(1 - \bar{D})^2 + n_0\bar{D}^2 \\ &= n_1\frac{n_0^2}{n^2} + n_0\frac{n_1^2}{n^2} = \frac{n_0 n_1(n_0 + n_1)}{n^2} = \frac{n_0 n_1}{n} \\ &= \frac{60 \times 100}{160} = 37.5.\end{aligned}$$

For the numerator, the $\bar{Y}$ term drops because $\sum_{i=1}^{n}(D_i - \bar{D}) = n_1 - n\bar{D} = 100 - 160(0.625) = 0$:

$$\begin{aligned}\sum_{i=1}^{n}(D_i - \bar{D})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}(D_i - \bar{D})Y_i - \bar{Y}\sum_{i=1}^{n}(D_i - \bar{D}) \\ &= \sum_{i:D_i=1}(1 - \bar{D})Y_i + \sum_{i:D_i=0}(-\bar{D})Y_i - 0 \\ &= \frac{n_0}{n}\, n_1\bar{Y}_1 - \frac{n_1}{n}\, n_0\bar{Y}_0 \\ &= \frac{n_0 n_1}{n}(\bar{Y}_1 - \bar{Y}_0).\end{aligned}$$

Dividing,

$$\hat{\beta}_1 = \frac{(n_0 n_1/n)(\bar{Y}_1 - \bar{Y}_0)}{n_0 n_1/n} = \bar{Y}_1 - \bar{Y}_0.$$

For the intercept, write the overall mean as a weighted average of the group means, $\bar{Y} = (n_0\bar{Y}_0 + n_1\bar{Y}_1)/n = \bar{Y}_0 + \bar{D}(\bar{Y}_1 - \bar{Y}_0)$:

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{D} \\ &= \bar{Y}_0 + \bar{D}(\bar{Y}_1 - \bar{Y}_0) - (\bar{Y}_1 - \bar{Y}_0)\bar{D} \\ &= \bar{Y}_0.\end{aligned}$$

Hence the group averages are read off the estimates:

$$\begin{aligned}\bar{Y}_0 &= \hat{\beta}_0 = 520, \\ \bar{Y}_1 &= \hat{\beta}_0 + \hat{\beta}_1 = 520 + 85 = 605.\end{aligned}$$

Check with the overall mean: $(60\times520+100\times605)/160 = (31200+60500)/160 = 573.125$, and $\hat{\beta}_0 + \hat{\beta}_1\bar{D} = 520 + 85(0.625) = 573.125$. Therefore

Cash-only: SGD 520/day; E-payment: SGD 605/day.

(b) From part (a), $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$, so

$$\begin{aligned}\hat{\beta}_1 &= 605 - 520 \\ &= 85,\end{aligned}$$

the e-payment group's average daily revenue minus the cash-only group's average, in SGD per day. Therefore

E-payment stalls earn, on average, SGD 85/day more than cash-only stalls.

(c) Start from two separate lines, one per group, with intercepts a_0, a_1 and slopes b_0, b_1:

$$\text{cash-only: } a_0 + b_0\,\text{Hours}, \qquad \text{e-payment: } a_1 + b_1\,\text{Hours}.$$

The common-slope assumption imposes $b_0 = b_1 = b$. Both lines are then reproduced by one equation with a dummy:

$$\widehat{\text{Revenue}} = a_0 + (a_1 - a_0)D + b\,\text{Hours},$$

since $D = 0$ returns $a_0 + b\,\text{Hours}$ and $D = 1$ returns $a_0 + (a_1 - a_0) + b\,\text{Hours} = a_1 + b\,\text{Hours}$. Match coefficients with $150 + 85\,D + 45\,\text{Hours}$:

$$\begin{aligned} a_0 &= 150, \\ a_1 - a_0 = 85 &\implies a_1 = 150 + 85 = 235, \\ b &= 45. \end{aligned}$$

The group whose intercept is the regression intercept, a_0, is the base group: cash-only stalls. The coefficient $85 = a_1 - a_0$ is the difference between the two intercepts, which, with equal slopes, is the revenue gap at every fixed number of hours. Therefore

Cash-only: $\widehat{\text{Revenue}} = 150 + 45\,\text{Hours}$
E-payment: $\widehat{\text{Revenue}} = 235 + 45\,\text{Hours}$

(d) Suppose C_i is added, with coefficients b_0, b_D, b_C, b_H. Substitute $C_i = 1 - D_i$:

$$\begin{aligned} b_0 + b_D D_i + b_C C_i + b_H \text{Hours}_i &= b_0 + b_D D_i + b_C(1 - D_i) + b_H \text{Hours}_i \\ &= (b_0 + b_C) + (b_D - b_C)D_i + b_H \text{Hours}_i. \end{aligned}$$

The fitted values, and hence the SSR, depend only on $b_0 + b_C$, $b_D - b_C$ and b_H. The fit of part (c) requires $b_0 + b_C = 150$, $b_D - b_C = 85$, $b_H = 45$, so for every real number c,

$$(b_0, b_D, b_C, b_H) = (150 - c,\ 85 + c,\ c,\ 45)$$

gives exactly the same fitted values. For instance,

$$\begin{aligned} c = 0: &\quad (b_0, b_D, b_C, b_H) = (150,\ 85,\ 0,\ 45), \\ c = 150: &\quad (b_0, b_D, b_C, b_H) = (0,\ 235,\ 150,\ 45), \\ c = -100: &\quad (b_0, b_D, b_C, b_H) = (250,\ -15,\ -100,\ 45). \end{aligned}$$

Infinitely many coefficient vectors minimise the SSR, so OLS has no unique solution. With g groups the g dummies sum to 1 for every observation, so the same problem arises whenever all g appear with the intercept; dropping one dummy (the base group) removes it. Therefore

$D_i + C_i = 1$ for all i $\Rightarrow$ exact collinearity with the intercept; use only $g - 1$ dummies for g groups.

Method 3: First-order conditions and partial effects

(a) OLS chooses (b_0, b_1) to minimise the SSR, which splits into the two groups because every D_i is 0 or 1:

$$\text{SSR}(b_0, b_1) = \sum_{i:D_i=0} (Y_i - b_0)^2 + \sum_{i:D_i=1} (Y_i - b_0 - b_1)^2.$$

Only the second sum contains b_1. Its first-order condition is

$$\begin{aligned} \frac{\partial \text{SSR}}{\partial b_1} = -2 \sum_{i:D_i=1} (Y_i - b_0 - b_1) &= 0 \\ n_1 \bar{Y}_1 - n_1 b_0 - n_1 b_1 &= 0 \\ b_0 + b_1 &= \bar{Y}_1. \end{aligned}$$

The first-order condition for b_0 is

$$\frac{\partial \text{SSR}}{\partial b_0} = -2 \sum_{i:D_i=0} (Y_i - b_0) - 2 \sum_{i:D_i=1} (Y_i - b_0 - b_1) = 0.$$

The second sum is zero by the b_1 condition, leaving

$$\begin{aligned} \sum_{i:D_i=0} (Y_i - b_0) &= 0 \\ n_0 \bar{Y}_0 - n_0 b_0 &= 0 \\ b_0 &= \bar{Y}_0. \end{aligned}$$

So the estimates satisfy $\hat{\beta}_0 = \bar{Y}_0$ and $\hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_1$, and

$$\begin{aligned} \bar{Y}_0 &= 520, \\ \bar{Y}_1 &= 520 + 85 = 605. \end{aligned}$$

Therefore

Cash-only: SGD 520/day; E-payment: SGD 605/day.

(b) Subtract the two first-order-condition results:

$$\begin{aligned} \hat{\beta}_1 &= (\hat{\beta}_0 + \hat{\beta}_1) - \hat{\beta}_0 \\ &= \bar{Y}_1 - \bar{Y}_0 \\ &= 605 - 520 \\ &= 85, \end{aligned}$$

the gap in average daily revenue between the two groups, in SGD per day. Therefore

E-payment stalls earn, on average, SGD 85/day more than cash-only stalls.

(c) Let $\widehat{R}(D, H) = 150 + 85D + 45H$, with H = Hours. The partial effect of hours is the derivative in H:

$$\frac{\partial \widehat{R}}{\partial H} = 45 \quad \text{for both } D = 0 \text{ and } D = 1,$$

so the two group lines have the same slope. The effect of switching D from 0 to 1 at fixed H is

$$\widehat{R}(1, H) - \widehat{R}(0, H) = (150 + 85 + 45H) - (150 + 45H)$$

$$= 85 \quad \text{for every } H.$$

Start from the base line ($D = 0$, cash-only), $\widehat{R}(0, H) = 150 + 45H$, and shift it up by 85 to get $\widehat{R}(1, H) = 235 + 45H$. At $H = 8$: $150 + 360 = 510$ and $235 + 360 = 595$, a gap of 85. The coefficient 85 is the predicted extra daily revenue of an e-payment stall over a cash-only stall operating the same number of hours. Therefore

$$\boxed{\begin{array}{c}\text{Cash-only: } \widehat{\text{Revenue}} = 150 + 45\,\text{Hours} \\ \text{E-payment: } \widehat{\text{Revenue}} = 235 + 45\,\text{Hours}\end{array}}$$

(d) With C_i included, let $r_i = Y_i - b_0 - b_D D_i - b_C C_i - b_H \text{Hours}_i$. The four first-order conditions (normal equations) are

$$\begin{aligned}\sum_{i=1}^{n} r_i &= 0 && (b_0),\\ \sum_{i=1}^{n} D_i r_i &= 0 && (b_D),\\ \sum_{i=1}^{n} C_i r_i &= 0 && (b_C),\\ \sum_{i=1}^{n} \text{Hours}_i\, r_i &= 0 && (b_H).\end{aligned}$$

Because $C_i = 1 - D_i$, the third condition is the first minus the second:

$$\sum_{i=1}^{n} C_i r_i = \sum_{i=1}^{n} (1 - D_i) r_i = \sum_{i=1}^{n} r_i - \sum_{i=1}^{n} D_i r_i.$$

In counts, $\sum_{i=1}^{n} 1 = 160 = \sum_{i=1}^{n} D_i + \sum_{i=1}^{n} C_i = 100 + 60$. So only three of the four equations are independent, there are four unknowns, and the system has infinitely many solutions: the OLS coefficients are not unique. Omitting one of the g dummies removes the redundant equation, which gives the rule of $g - 1$ dummies with an intercept. Therefore

$$\boxed{\begin{array}{c}D_i + C_i = 1 \text{ for all } i \;\Rightarrow\; \text{exact collinearity with the intercept;} \\ \text{use only } g - 1 \text{ dummies for } g \text{ groups.}\end{array}}$$

Marking scheme (10 marks)

(a) **(3 marks)** Group means

- 1 mark: correct method.
- 1 mark: correct cash-only mean (520).
- 1 mark: correct e-payment mean (605).

(b) **(2 marks)** Interpretation

- 1 mark: recognizes $\hat{\beta}_1$ as the difference in group means.
- 1 mark: correct direction/magnitude with units.

(c) **(3 marks)** Common-slope, different-intercepts model

- 1 mark: correct fitted line for cash-only stalls.
- 1 mark: correct fitted line for e-payment stalls, and correctly identifies cash-only as the base group.
- 1 mark: correct interpretation of the coefficient 85 holding Hours fixed (parallel-lines / vertical-shift interpretation).

(d) **(2 marks)** Dummy variable trap

- 1 mark: correctly explains the exact collinearity $D_i + C_i = 1$ with the intercept and why it breaks OLS.
- 1 mark: correctly states the general rule ($g - 1$ dummies for g groups, one omitted as base).

Unofficial – QRS@NTU – qrsntu.org

Question 5 (10 Marks)

(a) (4 marks) A ride-hailing platform in a large Southeast Asian city regresses a driver's weekly earnings on the number of hours the driver chose to be online (available for trips), using data from drivers who freely set their own hours. The platform's data team finds a strong positive coefficient and concludes that staying online longer *causes* higher weekly earnings on a per-hour basis. Explain, using Least Squares Assumption 1, why this conclusion may not be valid, giving a specific unobserved factor and the likely direction of the resulting bias.

(b) (3 marks) **Least Squares Assumption 1 counterexample.** Let $X \in \{0, 1\}$ be a binary regressor with $P(X = 0) = P(X = 1) = 0.5$, and let u be an error term with $E[u \mid X = 0] = -2$ and $E[u \mid X = 1] = 2$.

(i) Show that the *unconditional* mean $E[u] = 0$.

(ii) Explain why Least Squares Assumption 1 is nonetheless violated, and briefly describe (in words or with a simple sketch) what this would look like: two groups whose regression errors are centred at different, non-zero values even though the overall average error is zero.

(c) (3 marks) A public-health agency screens travellers for a disease outbreak using a rapid test, classifying each traveller as "positive" (quarantine) or "negative" (release) based on a biomarker threshold. Define, in this context, a Type I error and a Type II error, and explain the trade-off the agency faces if it lowers the biomarker threshold required to be classified "positive".

Solution

Method 1: Least Squares Assumption 1 as a conditional-mean condition

(a) Write the model as $\text{Earnings}_i = \beta_0 + \beta_1 \text{Hours}_i + u_i$, where β_1 is the causal effect of one more online hour on weekly earnings (SGD per hour) and u_i collects every other determinant of earnings.

- Least Squares Assumption 1 requires $E[u_i \mid \text{Hours}_i] = 0$: the omitted factors must have the same mean at every level of online hours. Only then does $\hat{\beta}_1$ estimate the causal β_1.
- A specific unobserved factor in u_i is driver skill: local knowledge of where and when demand surges and efficient routing, which raise weekly earnings for any given number of hours.
- Drivers set their own hours. Skilled drivers earn more per hour online, so staying online longer pays them more and they choose more hours. High-hours drivers are therefore more skilled on average, so $E[u_i \mid \text{Hours}_i]$ rises with Hours_i and $\text{Cov}(\text{Hours}_i, u_i) > 0$.
- The regression attributes to hours part of the earnings gap that is due to skill, so $\hat{\beta}_1$ overstates β_1: the bias is upward. For example, if an extra hour truly adds SGD 20 per week, but drivers who work 10 more hours are also skilled enough to earn SGD 50 more per week at the same hours, the slope picks up $20 + 50/10 = 25$ SGD per hour.
- The positive coefficient is therefore an association driven partly by self-selection, not a causal per-hour effect; random assignment of hours would remove the selection.

Therefore

$$\boxed{E[u \mid \text{Hours}] \neq 0 \text{ (self-selection on driver skill)} \Rightarrow \hat{\beta}_1 \text{ upward biased.}}$$

(b) (i) X takes only the values 0 and 1, so the law of total expectation averages the two conditional means with weights $P(X = 0)$ and $P(X = 1)$:

$$\begin{aligned} E[u] &= P(X = 0)E[u \mid X = 0] + P(X = 1)E[u \mid X = 1] \quad \text{(total expectation)} \\ &= 0.5(-2) + 0.5(2) \\ &= -1 + 1 \\ &= 0. \end{aligned}$$

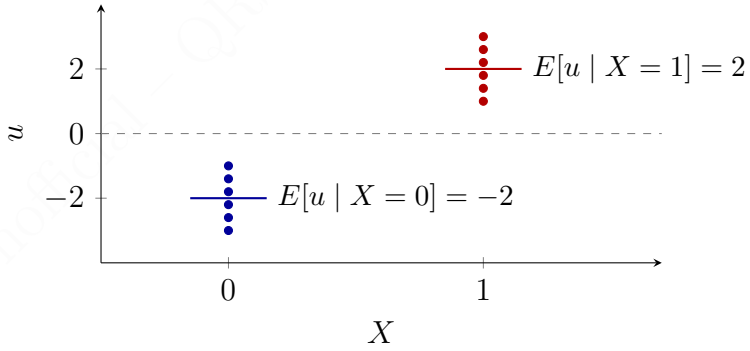
Therefore

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

(ii) Least Squares Assumption 1 requires $E[u \mid X = x] = 0$ for every value x that X takes. Check both values:

$$\begin{aligned} E[u \mid X = 0] &= -2 \neq 0, \\ E[u \mid X = 1] &= 2 \neq 0. \end{aligned}$$

The condition fails at both values, so Least Squares Assumption 1 is violated. The unconditional mean is zero only because the two group means -2 and $+2$ receive equal weights 0.5 and cancel. In a sketch, the errors of the $X = 0$ group are scattered around -2 and those of the $X = 1$ group around $+2$, while the overall average error is 0:



Therefore

$$\boxed{\begin{array}{c} E[u] = 0 \text{ does not imply } E[u \mid X] = 0 \text{ for all } X: \\ \text{Assumption 1 needs the conditional statement.} \end{array}}$$

(c) Take the null hypothesis H_0: the traveller is not infected. Classifying a traveller "positive" (quarantine) is rejecting H_0; classifying "negative" (release) is not rejecting it.

- A Type I error is rejecting a true H_0: a healthy traveller's biomarker exceeds the threshold, so the traveller is classified "positive" and quarantined unnecessarily (lost time, quarantine costs).
- A Type II error is not rejecting a false H_0: an infected traveller's biomarker falls below the threshold, so the traveller is classified "negative", released, and may spread the disease.

- Lowering the threshold makes the "positive" region larger, so more travellers are classified positive. Among healthy travellers, more are now quarantined, so the probability of a Type I error rises. Among infected travellers, fewer are released, so the probability of a Type II error falls.
- One threshold controls both error rates, so they cannot both be reduced by moving it; the agency trades more unnecessary quarantines for fewer missed cases, which is attractive during an outbreak when a missed case is costly.

Therefore

Lower threshold: $\downarrow$ Type II error (fewer missed cases), $\uparrow$ Type I error (more healthy travellers quarantined).

Unofficial QRS@NTU - qrsntu.org

Method 2: Omitted-variable-bias formula, covariance and a threshold model

(Part (a) uses the omitted-variable-bias formula, Stock and Watson Ch. 6 / beyond the quiz scope.)

(a) Suppose earnings depend on hours and skill, $\text{Earnings}_i = \beta_0 + \beta_1 \text{Hours}_i + \beta_2 \text{Skill}_i + v_i$ with $E[v_i \mid \text{Hours}_i, \text{Skill}_i] = 0$. Regressing on hours only leaves $u_i = \beta_2 \text{Skill}_i + v_i$ in the error. Write $H_i = \text{Hours}_i$ and $W_i = \text{Earnings}_i$. Since $W_i - \bar{W} = \beta_1(H_i - \bar{H}) + (u_i - \bar{u})$,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (H_i - \bar{H})(W_i - \bar{W})}{\sum_{i=1}^n (H_i - \bar{H})^2} \\ &= \frac{\beta_1 \sum_{i=1}^n (H_i - \bar{H})^2 + \sum_{i=1}^n (H_i - \bar{H})(u_i - \bar{u})}{\sum_{i=1}^n (H_i - \bar{H})^2} \\ &= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (H_i - \bar{H})(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^n (H_i - \bar{H})^2}.\end{aligned}$$

In large samples the sample moments converge to population moments, so

$$\begin{aligned}\hat{\beta}_1 &\xrightarrow{p} \beta_1 + \frac{\text{Cov}(H, u)}{\text{Var}(H)} \\ &= \beta_1 + \frac{\text{Cov}(H, \beta_2 \text{Skill} + v)}{\text{Var}(H)} \\ &= \beta_1 + \beta_2 \frac{\text{Cov}(H, \text{Skill})}{\text{Var}(H)},\end{aligned}$$

using $\text{Cov}(H, v) = 0$. Sign the bias term: $\beta_2 > 0$ because skill raises earnings at given hours, and $\text{Cov}(H, \text{Skill}) > 0$ because skilled drivers choose to stay online longer. The product is positive, so $\hat{\beta}_1$ is biased upward. For example, with $\beta_2 = 50$ SGD per week per unit of skill and $\text{Cov}(H, \text{Skill})/\text{Var}(H) = 0.1$ skill units per hour, the bias is $50 \times 0.1 = 5$ SGD per hour. The bias term is nonzero exactly because $\text{Cov}(H, u) \neq 0$. Least Squares Assumption 1 would give $\text{Cov}(H, u) = E\big[H\, E[u \mid H]\big] - E[H]\, E[u] = 0 - 0 = 0$, so $\text{Cov}(H, u) \neq 0$ implies $E[u \mid \text{Hours}] \neq 0$. Therefore

$$\boxed{E[u \mid \text{Hours}] \neq 0 \text{ (self-selection on driver skill)} \Rightarrow \hat{\beta}_1 \text{ upward biased.}}$$

(b) (i) Because $X \in \{0, 1\}$, the conditional mean of u is a linear function of X that matches both given values:

$$E[u \mid X] = -2 + 4X,$$

since $X = 0$ gives -2 and $X = 1$ gives $-2 + 4 = 2$. Also $E[X] = 0 \cdot 0.5 + 1 \cdot 0.5 = 0.5$. By the law of iterated expectations,

$$\begin{aligned}E[u] &= E\big[E[u \mid X]\big] \\ &= E[-2 + 4X] \\ &= -2 + 4E[X] \\ &= -2 + 4(0.5) \\ &= 0.\end{aligned}$$

Therefore

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

(ii) Least Squares Assumption 1 implies $\text{Cov}(X, u) = 0$, because $E[Xu] = E\big[X\,E[u \mid X]\big] = E[X \cdot 0] = 0$. Compute the covariance here, using $X^2 = X$ for a binary variable:

$$\begin{aligned}\text{Cov}(X, u) &= E[Xu] - E[X]E[u] \\ &= E\big[X\,E[u \mid X]\big] - 0.5 \cdot 0 \\ &= E[X(-2 + 4X)] \\ &= -2E[X] + 4E[X^2] \\ &= -2(0.5) + 4(0.5) \\ &= 1 \neq 0.\end{aligned}$$

A nonzero covariance is incompatible with Least Squares Assumption 1, so the assumption fails. To see what this does, let $Y = \beta_0 + \beta_1 X + u$. Then

$$\begin{aligned}E[Y \mid X] &= \beta_0 + \beta_1 X + (-2 + 4X) \\ &= (\beta_0 - 2) + (\beta_1 + 4)X,\end{aligned}$$

so the group means of Y are pushed down by 2 in the $X = 0$ group and up by 2 in the $X = 1$ group, and OLS (the difference of group means) estimates $\beta_1 + 4$ rather than β_1, even though the overall average error is zero. Therefore

> $E[u] = 0$ does not imply $E[u \mid X] = 0$ for all X:
> Assumption 1 needs the conditional statement.

(c) Let B be the biomarker and classify "positive" when $B > c$ for a threshold c. Let F_H and F_I be the cumulative distribution functions of B among healthy and infected travellers. With H_0: not infected,

$$\begin{aligned}P(\text{Type I}) &= \alpha(c) = P(B > c \mid \text{healthy}) = 1 - F_H(c), \\ P(\text{Type II}) &= \beta(c) = P(B \leq c \mid \text{infected}) = F_I(c).\end{aligned}$$

A Type I error quarantines a healthy traveller; a Type II error releases an infected one. Every cumulative distribution function is non-decreasing, so lowering the threshold from c to $c' < c$ gives

$$\begin{aligned}F_H(c') \leq F_H(c) &\implies \alpha(c') \geq \alpha(c), \\ F_I(c') \leq F_I(c) &\implies \beta(c') \leq \beta(c).\end{aligned}$$

Illustration: suppose $B \sim N(0, 1)$ for healthy and $B \sim N(3, 1)$ for infected travellers. At $c = 2$,

$$\begin{aligned}\alpha(2) &= 1 - \Phi(2) = 1 - 0.9772 = 0.0228, \\ \beta(2) &= \Phi\big((2 - 3)/1\big) = \Phi(-1) = 1 - 0.8413 = 0.1587.\end{aligned}$$

Lowering the threshold to $c = 1.5$,

$$\begin{aligned}\alpha(1.5) &= 1 - \Phi(1.5) = 1 - 0.9332 = 0.0668, \\ \beta(1.5) &= \Phi\big((1.5 - 3)/1\big) = \Phi(-1.5) = 1 - 0.9332 = 0.0668.\end{aligned}$$

The share of healthy travellers quarantined rises from 2.28% to 6.68% while the share of infected travellers released falls from 15.87% to 6.68%. Therefore

> Lower threshold: $\downarrow$ Type II error (fewer missed cases),
> $\uparrow$ Type I error (more healthy travellers quarantined).

Marking scheme (10 marks)

(a) **(4 marks)** Selection bias

- 1 mark: correctly states Assumption 1 as the relevant condition.
- 1 mark: identifies a specific, plausible confounding factor (e.g. driver skill/local knowledge).
- 1 mark: explains why the confounder correlates with hours chosen (self-selection).
- 1 mark: correct direction of bias (upward) with valid justification.

(b) **(3 marks)** Assumption 1 counterexample

- 1 mark: correct computation $E[u] = 0$ via the law of total expectation.
- 1 mark: correctly identifies that $E[u \mid X = 0] \neq 0$ and $E[u \mid X = 1] \neq 0$ violate Assumption 1.
- 1 mark: valid explanation/description that unconditional $E[u] = 0$ does not imply conditional mean independence.

(c) **(3 marks)** Type I/II trade-off

- 1 mark: correct Type I error definition in context.
- 1 mark: correct Type II error definition in context.
- 1 mark: correct description of the trade-off when the threshold is lowered.

Unofficial QRS@NTU material