# EC 003 Econometrics

Previous year questions, sorted by topic

All 31 questions from four final examinations (2022–23, 2023–24, December 2024 and December 2025), grouped into eleven topics. Every question is reproduced in full with its source and marks. A question with several parts stays in one piece under its main topic; *See also* notes point to parts that belong to another topic.

## PAPERS COVERED

| Tag | Examination | Course title on the paper | Marks, time |
|---|---|---|---|
| Dec 2025 | Final exam, 19 December 2025 | 003: Basic Econometrics | 70, 3 hours |
| Dec 2024 | Final exam, 20 December 2024 | 003: Basic Econometrics | 70, 3 hours |
| 2023–24 | Final exam, Semester I, 2023–24 | EC 003: Introductory Econometrics | 70, 3 hours |
| 2022–23 | Final exam, Semester I, 2022–23 | EC 003: Introductory Econometrics | 70, 3 hours |

## TOPIC MAP

| | Topic | Dec 2025 | Dec 2024 | 2023–24 | 2022–23 | Marks | Page |
|---|---|---|---|---|---|---|---|
| PART I · PROBABILITY AND STATISTICAL INFERENCE | | | | | | | |
| 1 | Sample spaces, events and $\sigma$-algebras | — | — | — | A1, A2 | 9 | 2 |
| 2 | Conditional probability and independence | — | — | A2, A4 | A3 | 19 | 2 |
| 3 | Random variables and distributions | A2 | A1, A3 | A1, A3 | — | 34 | 3 |
| 4 | Chebyshev and the law of large numbers | A1 | A2 | — | — | 21 | 4 |
| 5 | Estimation | B1 | B1 | — | — | 36 | 5 |
| 6 | Hypothesis testing | — | B2, B3 | C1 | C1 | 24 | 6 |
| PART II · ECONOMETRICS | | | | | | | |
| 7 | The linear regression model | — | — | — | B1, B2, C2 | 29 | 7 |
| 8 | Dummy variables and structural change | C2 | C2 | B2 | — | 44 | 8 |
| 9 | Multicollinearity | C1 | C1 | B1 | — | 28 | 10 |
| 10 | Heteroscedasticity | — | — | C2 | B3 | 19 | 11 |
| 11 | Autocorrelation | — | — | C3 | C3 | 17 | 12 |
| | **Total** | 5 Qs | 8 Qs | 9 Qs | 9 Qs | 280 | |

Cells give the part and question number on the original paper (A1 = Part A, Question 1). Marks are summed over all four papers, with each question counted once under its main topic.

## NOTES ON THE TEXT

- Wording, data and notation follow the papers. Obvious typos and small grammar slips are corrected (for example, "4nd level" → "4th level", "Schedule Caste" → "Scheduled Caste"), and Type I/Type II is spelled consistently.
- Handwritten notes on the scanned copies that are not part of the question have been left out.

† Dec 2025, Part A: the printed marks for Q1 and Q2 are struck out and corrected by hand to 14 and 9. With these the paper totals 70.

‡ Dec 2024, Part A, Q3: no marks are printed for this question. The 70-mark total leaves 7.

PART I

# Probability and Statistical Inference

## 1 Sample Spaces, Events and $\sigma$-Algebras

2 questions · 9 marks · 2022–23

**Q1.1** 2022–23 · Part A · Q1 **4 marks**

State whether each of the following statements is TRUE or FALSE:

(a) All events are not subsets of the sample space.
(b) All subsets of the sample space are events.
(c) All events are subsets of the sample space but all subsets are not events.
(d) All events are subsets of the sample space and all subsets are also events.

**Q1.2** 2022–23 · Part A · Q2 **5 marks**

What is the difference between an algebra and a $\sigma$-algebra (sigma algebra)? Provide a suitable example for each.

## 2 Conditional Probability, Independence and Bayes' Theorem

3 questions · 19 marks · 2023–24, 2022–23

**Q2.1** 2023–24 · Part A · Q2 **5 marks (3 + 2)**

Explain Simpson's paradox with the help of a hypothetical numerical example and a diagram. Give an intuitive explanation of Simpson's paradox.

**Q2.2** 2023–24 · Part A · Q4 **8 marks**

Define three events ($A$, $B$ and $C$) in such a way that they fulfill the following conditions:

(i) $A$ and $B$ are independent;
(ii) $B$ and $C$ are independent;
(iii) but $A$ and $C$ are not independent.

**Q2.3** 2022–23 · Part A · Q3 **6 marks (2 + 4)**

*Test for a disease.* Suppose that the test for a rare disease is 90 percent reliable in the following sense: if a person has the disease, there is a probability of 0.9 that the test will give a positive response; whereas, if a person does not have the disease, there is a probability of only 0.1 that the test will give a positive response. Data indicate that your chances of having the disease are only 1 in 10,000. If you learn that you had a positive response to the test, answer the following:

(a) What is the probability that you have the disease? (2 marks)
(b) Is it applicable in the case of Covid-19? Explain. (4 marks)

# 3 Random Variables, Distributions and Expectation

5 questions · 34 marks · Dec 2025, Dec 2024, 2023–24

**Q3.1** Dec 2025 · Part A · Q2 **9 marks**†

Suppose that a coin is tossed repeatedly until a head is obtained for the first time, and let $X$ denote the number of tosses that are required. Sketch the c.d.f. of $X$.

† Printed marks struck out and corrected by hand on the paper.

**Q3.2** Dec 2024 · Part A · Q1 **8 marks**

Consider the following continuous distributions: (i) Exponential distribution, (ii) Gamma distribution, (iii) Beta distribution, (iv) Normal distribution.

State which of the following exhibit the memoryless property and provide proof to support your answer (a proof for just one distribution is sufficient).

(a) i (b) ii (c) iii (d) iv
(e) i & ii (f) ii & iii (g) iii & iv (h) iv & i

**Q3.3** Dec 2024 · Part A · Q3 **7 marks**‡

Suppose that the c.d.f. $F$ of a random variable $X$ is as sketched below. Find each of the following probabilities:

(a) $\Pr(X = -1)$ (b) $\Pr(X < 0)$ (c) $\Pr(X = 1)$ (d) $\Pr(0 < X < 3)$
(e) $\Pr(1 < X \leq 2)$ (f) $\Pr(X > 5)$ (g) $\Pr(3 \leq X \leq 4)$

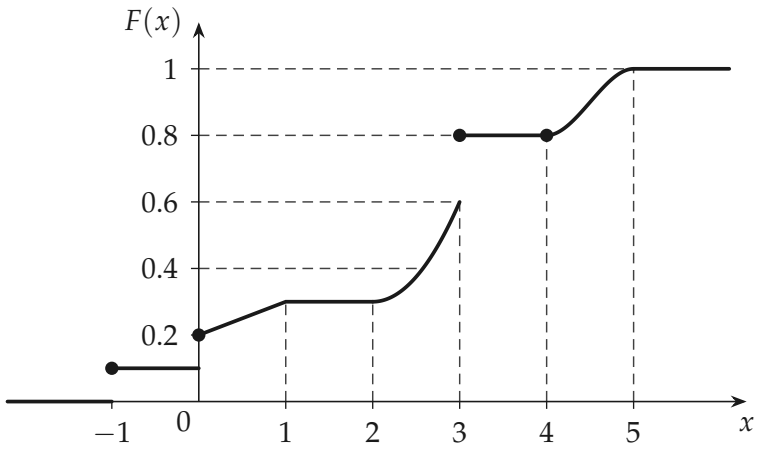



Figure redrawn from the question paper. ‡ Marks not printed on the paper; the 70-mark total leaves 7.

**Q3.4** 2023–24 · Part A · Q1 **5 marks**

Let $Y$ be the marks in EC003, $X_1$ a random vector of gender and $X_2$ a random vector of section. Find $E\big[\,E[Y \mid X_1, X_2] \mid X_1\big]$ using Table 1 and Table 2.

**Table 1**
Number of students by gender and section

| | Section A | Section B |
|---|---|---|
| Male | 39 | 50 |
| Female | 50 | 55 |
| Third gender | 1 | 5 |
| Total | 90 | 110 |

**Table 2**
Mean marks by gender and section

| | Section A | Section B |
|---|---|---|
| Male | 70 | 80 |
| Female | 75 | 85 |
| Third gender | 80 | 78 |

**Q3.5** 2023–24 · Part A · Q3 **5 marks**

A random variable $X$ has the normal distribution with mean $\mu$ and variance $\sigma^2$ ($-\infty < \mu < \infty$ and $\sigma > 0$) if $X$ has a continuous distribution with the following p.d.f.:

$$f(x \mid \mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \text{for } -\infty < x < \infty.$$

Prove that it is a valid p.d.f.

**See also.** 2022–23 B1(c)–(d), Q7.1: conditional versus unconditional mean and variance.

## 4 Chebyshev's Inequality and the Law of Large Numbers

2 questions · 21 marks · Dec 2025, Dec 2024

**Q4.1** Dec 2025 · Part A · Q1 **14 marks**†

Suppose that $X_1, \ldots, X_n$ form a random sample of size $n$ from a distribution for which the mean is 6.5 and the variance is 4. Determine how large the value of $n$ must be in order for the following relation to be satisfied:

$$\Pr\big(6 \le \bar{X}_n \le 7\big) \ge 0.8.$$

† Printed marks struck out and corrected by hand on the paper.

**Q4.2** Dec 2024 · Part A · Q2 **7 marks**

How does the law of large numbers justify the frequency interpretation of probability? Explain with the help of any hypothetical example.

**See also.** Dec 2024 B1(a), Q5.2: Chebyshev's bound for an interval of two standard deviations.

# 5 Estimation: Sampling Distributions, Sufficiency, MLE and Confidence Intervals

2 questions · 36 marks · Dec 2025, Dec 2024

**Q5.1** Dec 2025 · Part B · Q1 **24 marks**

A random sample of size $n$, $(X_1, \ldots, X_n)$, is drawn from a normal distribution with mean $\mu$ and variance equal to 1. State whether each of the following statements is TRUE or FALSE and provide a *brief* explanation for your answer. No answer should exceed half a page of your answer book. Please begin your answer with TRUE or FALSE and *then* proceed to the explanation. There are 2 marks for the correct answer in each case and 4 marks for the explanation. For full credit, your explanation must be brief and precise.

(a) The expectation of the sample mean $\bar{X}_n$ is independent of $n$.

(b) The sum of squared observations $\sum X_i^2$ is a sufficient statistic to estimate $\mu$.

(c) To ensure that the variance of $\bar{X}_n$ is no greater than one-tenth of the variance of $X$, we must have a sample size of at least 20.

(d) To test a null hypothesis of $\mu = 0$ against an alternative hypothesis of $\mu = 1$, it is not possible to find a test based on the statistic $\bar{X}_n$ which always leads to the correct decision (i.e. always rejects the null when it is false and accepts it when it is true).

**Q5.2** Dec 2024 · Part B · Q1 **12 marks**

State whether each of the following statements is TRUE or FALSE and provide a *one-sentence* proof if true or counter-example if false (no more than 15 words). Please begin your answer with TRUE or FALSE and *then* proceed to the explanation. The point of this exercise is for you to express yourself precisely and *in words*. Stay within the 15-word limit to get full credit.

(a) If a random variable $X$ has finite mean and variance, the probability that it takes values within an interval of 2 standard deviations from its mean is strictly greater than $\frac{1}{2}$.

(b) Given a random sample $(X_1, \ldots, X_n)$ from a distribution with finite mean and variance, the sum of squared observations $\sum X_i^2$ is always a sufficient statistic to estimate the variance of the distribution.

(c) To construct a 95% confidence interval for the mean of a normal distribution based on a random sample $(X_1, \ldots, X_n)$ from this distribution, we do not require prior knowledge of the variance of the distribution.

(d) If the maximum likelihood estimator for a parameter exists, its expectation is always the true value of the parameter.

**Pattern.** Both papers ask whether $\sum X_i^2$ is a sufficient statistic: for $\mu$ in Dec 2025, for the variance in Dec 2024.

**See also.** Dec 2024 B3(a), Q6.2: finding a maximum likelihood estimator.

# 6 Hypothesis Testing: Error Probabilities, Power Functions and Optimal Tests

4 questions · 24 marks · Dec 2024, 2023–24, 2022–23

**Q6.1** Dec 2024 · Part B · Q2 **5 marks**

Suppose that a single observation is sampled from the uniform distribution on the interval $[0, \theta]$, where the value of $\theta$ is unknown, and the following hypotheses are to be tested:

$$H_0 : \theta \leq 1 \qquad \text{and} \qquad H_1 : 1 < \theta \leq 2.$$

Consider a test procedure which rejects the null hypothesis whenever the value of the observation $x$ is greater than 1. Compare this with an alternative procedure which rejects the null hypothesis when $x \geq 1.5$. Which of these two test procedures do you prefer and why? Sketch the power function for these tests and mark the levels of Type I and Type II errors.

**Q6.2** Dec 2024 · Part B · Q3 **5 marks**

Let $X_1, X_2, \ldots, X_n$ represent a random sample from a distribution whose probability density function is given by

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \;\; 0 < \theta < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find the maximum likelihood estimator of $\theta$.

(b) Suppose you have a sample of size $n = 2$ and you would like to test the null hypothesis that $\theta = 1$ against the alternative $\theta = 2$. What is the test procedure that minimizes the sum of Type I and Type II errors?

**Q6.3** 2023–24 · Part C · Q1 **5 marks**

A single observation $X$ is taken from the uniform distribution on the interval $(\theta - 1, \; \theta + 1)$. Suppose that the following hypotheses are to be tested:

$$H_0 : \theta \leq 2, \qquad H_1 : \theta \geq 4.$$

Construct a test procedure such that both the Type I and Type II errors are zero.

**Q6.4** 2022–23 · Part C · Q1 **9 marks (5 + 4)**

Suppose that a single observation $X$ is drawn from the uniform distribution on the interval $[0, \theta]$, and that the following hypotheses are to be tested:

$$H_0 : \theta \geq 3, \qquad H_1 : \theta < 3.$$

Consider a test procedure such that the null is rejected when $X \leq 2$.

(a) Determine the power function of the test.

(b) Determine the size of the test.

**Pattern.** Three of the four questions test a hypothesis about $\theta$ from a single observation on a uniform distribution.

**See also.** Dec 2025 B1(d), Q5.1: a test that always reaches the correct decision; 2022–23 B1(b), Q7.1: the $p$-value versus the size of a test.

PART II
# Econometrics

## 7 The Linear Regression Model: OLS Properties, Inference and Specification

3 questions · 29 marks · 2022–23

**Q7.1** 2022–23 · Part B · Q1 **12 marks (6 × 2)**

State with reason whether the following statements are true, false, or uncertain. Be precise.

(a) The $t$ test of significance in two-variable regression requires that the sampling distributions of estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ follow the normal distribution.

(b) The $p$ value and the size of a test statistic mean the same thing.

(c) The conditional and unconditional means of a random variable are the same things.

(d) The conditional variance, $\text{var}(Y_i \mid X_i) = \sigma^2$, and the unconditional variance of $Y$, $\text{var}(Y) = \sigma_Y^2$, will be the same if $X$ had no influence on $Y$.

(e) The higher the value of $\sigma^2$, the larger is the variance of $\hat{\beta}_2$.

(f) Even though the disturbance term in the CLRM is not normally distributed, the OLS estimators are still unbiased.

**Q7.2** 2022–23 · Part B · Q2 **12 marks (4 + 3 + 3 + 2)**

(i) The linear regression function $Y_i = \alpha + \beta X_i$ has estimates $\hat{\alpha} = 2$, $\hat{\beta} = 0.4$, $\hat{\sigma}^2 = 4$ and $r_{xy} = 0.6$, considering $n = 100$. Test the hypothesis that $\beta = 0$. (4)

(ii) The equation to study the trade-off between time spent on sleeping and working, and other factors affecting sleep, is given as:

$$\begin{aligned} \widehat{sleep} = \underset{}{3638.25} &\quad \underset{(0.017)}{-0.148\, totwrk} \quad \underset{(5.88)}{-11.13\, educ} \quad \underset{(1.45)}{+2.20\, age} \\ n = 706, &\qquad R^2 = 0.113 \end{aligned}$$

where *sleep* and *totwrk* (total work) are measured in minutes per week, and *educ* and *age* are measured in years. Based on this equation, answer the following:

(i) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? (3)

(ii) Dropping *educ* and *age* from the equation gives

$$\begin{aligned} \widehat{sleep} = \underset{}{3586.38} &\quad \underset{(0.017)}{-0.151\, totwrk} \\ n = 706, &\qquad R^2 = 0.103 \end{aligned}$$

Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer. (3)

(iii) Does including *educ* and *age* in the model greatly affect the estimated trade-off between sleeping and working? (2)

**Q7.3** 2022–23 · Part C · Q2 **5 marks**

Suppose that the daily wages of a worker in the unorganized sector (*wages*) depend on two factors: the number of years of schooling of the worker (*school*) and the worker's ability (*ability*):

$$wages = \beta_0 + \beta_1\, school + \beta_2\, ability + u.$$

Assume that this equation satisfies the Gauss–Markov assumptions. If there are strong reasons to believe that *school* and *ability* are positively correlated, and that *ability* has a very strong impact on *wages*, what is the likely bias in the OLS estimate of $\beta_1$ obtained from the simple regression of *wages* on *school*?

**See also.** Dec 2024 C2(e)–(f), Q8.2: the number of observations from the $F$ statistic, and what a low $R^2$ does and does not mean.

## 8 Dummy Variables and Structural Change

3 questions · 44 marks · Dec 2025, Dec 2024, 2023–24

**Q8.1** Dec 2025 · Part C · Q2 **10 marks (3 + 3 + 2 + 2)**

To study the growth pattern of real per capita income in country W over the period 1980–2000, the following models are estimated.

$$\text{Model I:} \qquad \ln PCI_t = \gamma_0 + \gamma_1 t + e_t$$

$$\text{Model II:} \qquad \ln PCI_t = \delta_0 + \delta_1 t + \delta_2 D + \delta_3 (D \times t) + \varepsilon_t$$

where $PCI_t$ = real per capita income, $t$ = time trend, $D = 1$ for years 1990 onward and 0 otherwise, and ln denotes natural logarithm.

(a) Using Model I, calculate the average annual growth rate of real per capita income over the entire sample period.

(b) Are the growth rates statistically different in the pre-1990 and post-1990 periods? Explain. If different, compute the growth rates for both periods.

(c) Interpret the constant term in Model II.

(d) In Model III, the dummy variable is redefined as:

$$\text{Model III:} \qquad \ln PCI_t = \theta_0 + \theta_1 t + \theta_2 D_s + \theta_3 (D_s \times t) + \eta_t$$

where $D_s = 1$ for years before 1990 and 0 otherwise, while the dependent and other independent variables remain the same as in Model II.
Comment on Model III and compare it conceptually with Model II.

**Q8.2** Dec 2024 · Part C · Q2 **24 marks (6 × 4)**

Consider the regression output of health expenditure of adults in a private hospital on the following list of $X$ variables. (The unit of observation is a person.)

| **Health expenditure in private hospital (in '000 rupees)** | **Coefficient** | **Std. error** |
|---|---|---|
| $D_{\text{MALE}} = 1$ if male, 0 otherwise | 10 | 1 |
| Age | 20 | 5 |
| Age$^2$ | 5 | 2 |
| Education of the household head (years of education) | 44 | 1 |
| $D_{\text{SC}} = 1$ if person is from a Scheduled Caste household, 0 otherwise | −56 | 8 |
| $D_{\text{ST}} = 1$ if person is from a Scheduled Tribe household, 0 otherwise | −16 | 9 |
| $D_{\text{W2}} = 1$ if household has 2nd level of wealth status | 60 | 60 |
| $D_{\text{W3}} = 1$ if household has 3rd level of wealth status | 80 | 80 |
| $D_{\text{W4}} = 1$ if household has 4th level of wealth status | 96 | 8 |
| Constant | 132 | 56 |
| R-squared | 0.14 | |
| Adj. R-squared | 0.13 | |
| $F(9, 31420)$ | 200 | |
| Prob $> F$ | 0.02 | |

(95 percent significance cut-off is equal to 1.96.)

**Note:**

1. Households are classified as belonging to the 1st level of wealth status to the 4th level of wealth status. Households belonging to the 1st level are the poorest and the 4th level are the richest.
2. The average age of people in your sample is 45.
3. The average years of schooling for household heads is 5.

(a) Suppose the variables Age, Age$^2$ and Education of the household head were not in the regression. Interpret then the coefficient of the constant.

*For the rest of the questions, consider all variables in the table (so ignore (a)).*

(b) Does it matter that the patient is from a Scheduled Caste household? If yes, quantify the effect.

(c) Is the marginal effect in part (b) independent of the age of the person? What term would you have to add to the specification to allow for dependence of (b) on the age of the individual?

(d) What is the effect of wealth on health expenditure?

(e) How many observations is this regression based on? Explain from which formula you got this number.

(f) If someone says that since your $R^2$ is so low, the variables you have considered are not relevant, what will be your answer?

**Q8.3** 2023–24 · Part B · Q2 **10 marks (3 + 3 + 2 + 2)**

To study the rate of growth of population in country Y over the period 1970–1992, the following models are estimated.

$$\text{Model I:} \quad \begin{aligned} \widehat{\ln(Pop)}_t &= 4.73 + 0.025\,t \\ t &= (781.24) \quad (54.70) \end{aligned}$$

$$\text{Model II:} \quad \begin{aligned} \widehat{\ln(Pop)}_t &= 4.77 + 0.015\,t - 0.075\,D_t + 0.011\,D_t t \\ t &= (2477.93) \quad (34.02) \quad (-17.04) \quad (25.55) \end{aligned}$$

where Pop = population in millions, $t$ = trend variable, $D_t = 1$ for observations beginning in 1978 and 0 before 1978, and ln stands for natural logarithm.

(a) In Model I, calculate the rate of growth of country Y's population over the sample period.

(b) Are the population growth rates statistically different pre- and post-1978? Explain. If they are different, calculate the growth rates for 1972–1977 and 1978–1992.

(c) Interpret the constant term in Model II.

(d) In Model II a new variable $D_s$ is introduced as below:

$$\text{Model III:} \quad \widehat{\ln(Pop)}_t = \beta_0 + \beta_1 t - \beta_2 D_t - \beta_3 D_s + \beta_4 D_t t$$

where $D_s = 1$ if observations are before 1978 and $D_s = 0$ otherwise; the dependent and other independent variables in Model III remain the same as in Model II. Comment on Model III.

**Pattern.** Dec 2025 C2 and 2023–24 B2 ask the same four sub-questions on different data.

## 9 Multicollinearity

3 questions · 28 marks · Dec 2025, Dec 2024, 2023–24

**Q9.1** Dec 2025 · Part C · Q1 **13 marks (3 + 2 + 3 + 2 + 3)**

The following regression results are obtained using annual data for the agricultural sector of country Z.

$$\text{Regression (I):} \quad \ln Q_t = \alpha_0 + \alpha_1 \ln A_t + \alpha_2 \ln F_t + \alpha_3 t + u_t$$

$$\text{Regression (II):} \quad \ln Q_t = \beta_0 + \beta_1 \ln\left(\frac{A_t}{F_t}\right) + \beta_2 \ln F_t + \beta_3 t + v_t$$

where $Q_t$ = index of real agricultural output, $A_t$ = index of cultivated area, $F_t$ = index of fertilizer use, $t$ = time trend, $u_t$ and $v_t$ are error terms.

(a) Which of the above regressions is more likely to suffer from multicollinearity? Explain how you would identify it empirically.

(b) In the presence of multicollinearity, can the OLS estimators be considered BLUE? Why or why not?

(c) Does transforming Regression (I) into Regression (II) help in reducing multicollinearity? Explain.

(d) Discuss the possible remedies to solve the problem of multicollinearity if it exists in the above regressions.

(e) Can the goodness of fit of the above regressions be compared using $R^2$? If not, how can the models be made comparable?

**Q9.2** Dec 2024 · Part C · Q1 **2 marks**

"If an auxiliary regression shows that a particular $R_i^2$ is high, there is definite evidence of high collinearity." Is this statement true, false, or uncertain? Explain.

**Q9.3** 2023–24 · Part B · Q1 **13 marks (3 + 2 + 3 + 2 + 3)**

Following are the regression results obtained from the annual data of the manufacturing sector of country X.

$$\begin{aligned} \widehat{\log Y} = \ & 2.81 \quad - 0.53 \log K \quad + 0.91 \log L \quad + 0.047\, t \\ \text{se} = \ & (1.38) \quad (0.34) \quad (0.14) \quad (0.021) \\ & R^2 = 0.97, \qquad F = 189.8 \end{aligned} \tag{1}$$

where $Y$ = index of real output, $K$ = index of real capital input, $L$ = index of real labor input, and $t$ = time or trend.

Another regression is obtained using the same data:

$$\begin{aligned} \widehat{\log(Y/L)} = \ & -0.11 \quad + 0.11 \log(K/L) \quad + 0.006\, t \\ \text{se} = \ & (0.03) \quad (0.16) \quad (0.007) \\ & R^2 = 0.66, \qquad F = 19.4 \end{aligned} \tag{2}$$

(a) Identify which regression has multicollinearity. How would you identify it?

(b) Can we call our estimates BLUE in the presence of multicollinearity? Why or why not?

(c) Can multicollinearity be reduced by changing the form from either of the regressions to the other?

(d) What are the possible remedies to solve the problem of multicollinearity if it exists in the above regressions?

(e) Can we compare the $R^2$ values of the above regressions? If not, how can you make them comparable?

**Pattern.** Dec 2025 C1 and 2023–24 B1 ask the same five sub-questions on different data. Their part (e), comparing $R^2$ across models, returns in 2022–23 B3(c), Q10.2.

## 10 Heteroscedasticity

2 questions · 19 marks · 2023–24, 2022–23

**Q10.1** 2023–24 · Part C · Q2 **12 marks (7 + 5)**

Consider a regression of a company's sales ($Y_i$) on its expenditure on advertising ($X_i$) and a dummy variable ($D_i$) which takes a value 1 if the company belongs to the organised sector and 0 otherwise.

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$$

It is suspected that the variance of the error term might not be constant, i.e. $\text{Var}(u_i) = \sigma_i^2$.

(a) Explain how you would use White's test to test for heteroscedasticity.

(b) Outline the procedure to obtain White's robust standard errors for the estimated slope coefficients when $\sigma_i^2$ is unknown.

**Q10.2** 2022–23 · Part B · Q3 **7 marks**

In a regression of average wages ($W$, \$) on the number of employees ($N$) for a random sample of 30 firms, the following regression results were obtained:

$$\begin{aligned} \widehat{W} &= 7.5 \quad + 0.009\,N \\ t &= \text{n.a.} \quad (16.10) \qquad R^2 = 0.90 \end{aligned} \tag{1}$$

$$\begin{aligned} \widehat{W}/N &= 0.008 \quad + 7.8\,(1/N) \\ t &= (14.43) \quad (76.58) \qquad R^2 = 0.99 \end{aligned} \tag{2}$$

(a) How do you interpret the two regressions?

(b) Can you relate the slopes and intercepts of the two models?

(c) Can you compare the $R^2$ values of the two models? Why or why not?

## 11 Autocorrelation

2 questions · 17 marks · 2023–24, 2022–23

**Q11.1** 2023–24 · Part C · Q3 **7 marks**

Consider the model

$$Y_t = \beta_1 + \beta_2 X_t + u_t, \qquad \text{where } u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t,$$

that is, the error term follows an AR(2) scheme and $\varepsilon_t$ is a white noise error term. Explain the steps to obtain estimates of $\rho_1$ and $\rho_2$ using the Cochrane–Orcutt iterative procedure.

**Q11.2** 2022–23 · Part C · Q3 **10 marks**

Consider the model

$$Y_t = \beta_1 + \beta_2 X_t + u_t, \qquad \text{where } u_t = \rho_1 u_{t-1} + \varepsilon_t, \quad |\rho_1| < 1,$$

and $\varepsilon_t$ is a white noise error term. Explain the steps to obtain estimates of $\rho_1$ using the Cochrane–Orcutt iterative procedure.

**Pattern.** Both questions ask for the Cochrane–Orcutt iterative procedure: AR(2) errors in 2023–24, AR(1) errors in 2022–23.

— End of question bank —