

ĐỀ CƯƠNG ÔN TẬP

Kinh tế lượng

Dành cho sinh viên đại học khối kinh tế, tài chính

Econometrics — a bilingual study guide

Hồi quy hai biến · Hồi quy bội · Kiểm định và dự báo · Biến giả · Đa cộng tuyến · Phương sai sai số thay đổi · Tự tương quan · Chọn mô hình

Tính tay từng bước như bài thi · Đọc bảng kết quả EViews · Dữ liệu Việt Nam

Bài tập có lời giải · Tự luận có dàn ý · Trắc nghiệm có đáp án · Bảng tra xác suất

Song ngữ Việt – Anh · Vietnamese – English

Phiên bản 1.0 — Tháng 10 năm 2026

<https://globalmarket.vin>

M U C L U C

CONTENTS

Chương 1. Nhập môn kinh tế lượng và ôn tập xác suất thống kê — Chapter 1.

<i>Introduction to econometrics and a review of probability and statistics</i>	6
1.1. Kinh tế lượng là gì — <i>What is econometrics</i>	6
1.2. Quy trình nghiên cứu kinh tế lượng — <i>The econometric research process</i>	8
1.3. Dữ liệu trong kinh tế lượng — <i>Data in econometrics</i>	9
1.4. Tương quan và nhân quả — <i>Correlation and causation</i>	12
1.5. Ôn tập: biến ngẫu nhiên, kỳ vọng, phương sai, hiệp phương sai — <i>Review: random variables, expectation, variance, covariance</i>	14
1.6. Các phân phối xác suất dùng trong kinh tế lượng — <i>Probability distributions used in econometrics</i>	19
1.7. Ước lượng điểm và ước lượng khoảng — <i>Point and interval estimation</i>	22
1.8. Kiểm định giả thuyết thống kê — <i>Statistical hypothesis testing</i>	24
1.9. Phần mềm kinh tế lượng — <i>Econometric software</i>	27
1.10. Tóm tắt, so sánh và thuật ngữ — <i>Summary, comparisons and glossary</i>	29
1.11. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	31
1.12. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	34
1.13. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	36

Chương 2. Hồi quy hai biến: ước lượng bằng phương pháp bình phương nhỏ nhất — Chapter 2. Two-variable regression: estimation by ordinary least squares 39

2.1. Hàm hồi quy tổng thể và hàm hồi quy mẫu — <i>Population and sample regression functions</i>	39
2.2. Ý nghĩa của từ “tuyến tính” — <i>What “linear” means</i>	41
2.3. Phương pháp bình phương nhỏ nhất (OLS) — <i>The ordinary least squares (OLS) method</i>	42
2.4. Bài tính tay đầy đủ trên bộ A — <i>A full hand calculation on set A</i>	44
2.5. Ý nghĩa của các hệ số hồi quy — <i>Interpreting the regression coefficients</i>	47
2.6. Phân tích biến động: $TSS = ESS + RSS$, hệ số xác định R^2 — <i>Decomposing variation: $TSS = ESS + RSS$ and R^2</i>	48
2.7. Ước lượng phương sai của sai số ngẫu nhiên — <i>Estimating the error variance</i>	50
2.8. Giả thiết của mô hình hồi quy tuyến tính cổ điển và định lý Gauss–Markov — <i>Assumptions of the classical linear regression model and the Gauss–Markov theorem</i>	51
2.9. Phương sai và sai số chuẩn của β_1, β_2 — <i>Variances and standard errors of β_1 and β_2</i>	53
2.10. Tính chất của đường hồi quy mẫu — <i>Properties of the sample regression line</i>	55
2.11. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	59
2.12. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	62
2.13. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	64

Chương 3. Hồi quy hai biến: khoảng tin cậy, kiểm định và dự báo — Chapter 3. Two-variable regression: interval estimation, hypothesis testing and forecasting.....	66
3.1. Phân phối xác suất của các ước lượng OLS — <i>Sampling distributions of the OLS estimators</i>	66
3.2. Khoảng tin cậy của hệ số hồi quy và của phương sai sai số — <i>Confidence intervals for the coefficients and the disturbance variance</i>	68
3.3. Kiểm định giả thuyết về hệ số hồi quy — <i>Hypothesis tests on a regression coefficient</i>	70
3.4. Phân tích phương sai và kiểm định F — <i>Analysis of variance and the F-test</i>	73
3.5. Dự báo giá trị trung bình và giá trị cá biệt — <i>Forecasting the mean and an individual value</i>	74
3.6. Đổi đơn vị đo của X và Y — <i>Changing the units of X and Y</i>	76
3.7. Các dạng hàm hồi quy — <i>Functional forms</i>	77
3.8. Hồi quy qua gốc tọa độ — <i>Regression through the origin</i>	79
3.9. Đọc bảng kết quả EViews của hồi quy hai biến — <i>Reading the EViews output of a two-variable regression</i>	80
3.10. Bài thi mẫu dạng câu 1 đề UFM — <i>Model exam question in the style of UFM question 1</i>	82
3.11. Tóm tắt, so sánh khái niệm dễ nhầm và thuật ngữ — <i>Summary, commonly confused concepts and glossary</i>	85
3.12. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	88
3.13. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	91
3.14. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	93
Chương 4. Hồi quy bội — Chapter 4. Multiple regression	95
4.1. Mô hình hồi quy bội và ý nghĩa hệ số hồi quy riêng — <i>The multiple regression model and partial regression coefficients</i>	95
4.2. Dạng ma trận và ước lượng bình phương nhỏ nhất — <i>Matrix form and least-squares estimation</i>	98
4.3. Hệ số xác định R^2 và R^2 hiệu chỉnh — <i>R-squared and adjusted R-squared</i>	101
4.4. Kiểm định t từng hệ số và khoảng tin cậy — <i>t-tests on individual coefficients and confidence intervals</i>	103
4.5. Kiểm định F sự phù hợp của mô hình — <i>The F-test of overall significance</i>	104
4.6. Dự báo trong hồi quy bội — <i>Forecasting with multiple regression</i>	106
4.7. Hồi quy doanh thu Hạt Nâu theo giá, quảng cáo và diện tích — <i>Hạt Nâu revenue regressed on price, advertising and area</i>	107
4.8. Kiểm định Wald thu hẹp: bỏ một nhóm biến — <i>The Wald test for dropping a group of regressors</i>	110
4.9. Hàm sản xuất Cobb–Douglas — <i>The Cobb–Douglas production function</i>	112
4.10. Kiểm định ràng buộc tuyến tính giữa các hệ số — <i>Testing linear restrictions on coefficients</i>	114
4.11. Bài thi mẫu: bảng EViews bị khuyết ô — <i>Model exam question: an EViews table with missing cells</i>	116
4.12. Tóm tắt và thuật ngữ — <i>Summary and glossary</i>	119

4.13. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	122
4.14. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	125
4.15. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	127
Chương 5. Hồi quy với biến giả — Chapter 5. Regression with dummy variables	130
5.1. Biến định tính và cách mã hóa biến giả — <i>Qualitative variables and dummy coding</i>	130
5.2. Biến giả hệ số chặn — <i>Intercept dummies</i>	132
5.3. Bẫy biến giả và phạm trù gốc — <i>The dummy variable trap and the base category</i>	133
5.4. Biến định tính nhiều phạm trù: mô hình đầy đủ của chuỗi Hạt Nâu — <i>Multi-category variables: the full Hạt Nâu model</i>	134
5.5. Biến giả hệ số góc (biến tương tác) — <i>Slope dummies (interaction terms)</i>	138
5.6. So sánh hai hồi quy: kiểm định Chow và cách dùng biến giả — <i>Comparing two regressions: the Chow test and the dummy approach</i>	141
5.7. Biến giả mùa vụ — <i>Seasonal dummies</i>	145
5.8. Biến giả trong mô hình log — <i>Dummies in log models</i>	147
5.9. Biến phụ thuộc nhị phân: mô hình xác suất tuyến tính, logit, probit — <i>Binary dependent variables: linear probability, logit and probit models</i>	148
5.10. Tóm tắt và thuật ngữ — <i>Summary and glossary</i>	151
5.11. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	154
5.12. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	157
5.13. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	159
Chương 6. Đa cộng tuyến — Chapter 6. Multicollinearity	162
6.1. Bản chất của đa cộng tuyến — <i>The nature of multicollinearity</i>	162
6.2. Nguyên nhân của đa cộng tuyến — <i>Causes of multicollinearity</i>	165
6.3. Hậu quả của đa cộng tuyến — <i>Consequences of multicollinearity</i>	165
6.4. Phát hiện đa cộng tuyến — <i>Detecting multicollinearity</i>	168
6.5. Phân tích đa cộng tuyến trên bộ dữ liệu C — <i>Analysing multicollinearity in data set C</i>	172
6.6. Khắc phục đa cộng tuyến — <i>Remedies for multicollinearity</i>	177
6.7. Khi nào không cần khắc phục — <i>When no remedy is needed</i>	183
6.8. Tóm tắt, so sánh và thuật ngữ — <i>Summary, comparisons and glossary</i>	184
6.9. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	187
6.10. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	190
6.11. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	193
Chương 7. Phương sai sai số thay đổi — Chapter 7. Heteroscedasticity	195
7.1. Bản chất và nguyên nhân — <i>Nature and causes</i>	195
7.2. Hậu quả đối với ước lượng OLS — <i>Consequences for OLS estimation</i>	199
7.3. Phát hiện bằng đồ thị, kiểm định Park và Glejser — <i>Graphical detection, the Park and Glejser tests</i>	200
7.4. Kiểm định Goldfeld–Quandt — <i>The Goldfeld–Quandt test</i>	203
7.5. Kiểm định Breusch–Pagan–Godfrey và kiểm định White — <i>The Breusch–Pagan–Godfrey and White tests</i>	204

7.6. Khắc phục: bình phương nhỏ nhất có trọng số và biến đổi mô hình — <i>Remedies: weighted least squares and model transformations</i>	209
7.7. Sai số chuẩn vững White — <i>White robust standard errors</i>	212
7.8. Câu thi mẫu và quy trình xử lý — <i>Sample exam question and workflow</i>	214
7.9. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	218
7.10. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	222
7.11. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	224
Chương 8. Tự tương quan — Chapter 8. Autocorrelation	227
8.1. Bản chất và nguyên nhân của tự tương quan — <i>Nature and causes of autocorrelation</i>	227
8.2. Lược đồ tự hồi quy bậc nhất AR(1) và hậu quả — <i>The first-order autoregressive scheme AR(1) and its consequences</i>	229
8.3. Phát hiện bằng đồ thị: bộ số liệu Sao Bắc — <i>Graphical detection: the Sao Bắc data</i>	231
8.4. Kiểm định Durbin–Watson — <i>The Durbin–Watson test</i>	235
8.5. Kiểm định Breusch–Godfrey (LM) — <i>The Breusch–Godfrey (LM) test</i>	237
8.6. Durbin h và kiểm định đoạn mạch — <i>Durbin's h and the runs test</i>	240
8.7. Khắc phục khi biết ρ : bình phương nhỏ nhất tổng quát (GLS) — <i>Remedy with known ρ: generalised least squares (GLS)</i>	242
8.8. Ước lượng ρ và thủ tục Cochrane–Orcutt: ước lượng lại bộ Sao Bắc — <i>Estimating ρ and the Cochrane–Orcutt procedure: re-estimating the Sao Bắc data</i>	243
8.9. Sai số chuẩn Newey–West (HAC) và so sánh các cách xử lý — <i>Newey–West (HAC) standard errors and a comparison of remedies</i>	246
8.10. Câu thi mẫu, tóm tắt và thuật ngữ — <i>Sample exam question, summary and glossary</i>	248
8.11. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	253
8.12. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	256
8.13. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	258
Chương 9. Chọn mô hình, kiểm định chỉ định và ứng dụng với dữ liệu Việt Nam — Model selection, specification tests and an application to Vietnamese data	261
9.1. Thuộc tính của một mô hình tốt — <i>Properties of a good model</i>	261
9.2. Các sai lầm chỉ định — <i>Types of specification error</i>	262
9.3. Kiểm định chỉ định — <i>Specification tests</i>	265
9.4. Tiêu chí thông tin AIC, SC, HQ — <i>Information criteria AIC, SC, HQ</i>	269
9.5. Quy trình chọn mô hình đầy đủ trên bộ B — <i>A complete model-selection procedure on data set B</i>	271
9.6. Ứng dụng: hàm tiêu dùng Việt Nam với dữ liệu World Bank — <i>Application: a Vietnamese consumption function with World Bank data</i>	274
9.7. Viết tiểu luận, báo cáo kinh tế lượng — <i>Writing an econometrics term paper</i>	281
9.8. Tóm tắt, so sánh và thuật ngữ — <i>Summary, comparisons and glossary</i>	283
9.9. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	287
9.10. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	290
9.11. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	292

Phụ lục A. Đáp án câu hỏi trắc nghiệm — Appendix A. Answer key.....	295
Phụ lục B. Bảng tra xác suất — Appendix B. Statistical tables.....	300
B.1. Phân phối chuẩn tắc — B.1. Standard normal distribution.....	300
B.2. Phân phối Student t — B.2. Student's t distribution.....	300
B.3. Phân phối khi bình phương χ^2 — B.3. Chi-square distribution χ^2	301
B.4. Phân phối F, mức ý nghĩa 5% — B.4. F distribution, 5% level	303
B.5. Phân phối F, mức ý nghĩa 1% — B.5. F distribution, 1% level	304
B.6. Thống kê Durbin–Watson, mức ý nghĩa 5% — B.6. Durbin–Watson statistic, 5% level	306
Phụ lục C. Công thức kinh tế lượng thường dùng — Appendix C. Common econometric formulas	308
Phụ lục D. Thuật ngữ Việt – Anh — Appendix D. Vietnamese–English glossary.....	310
Tài liệu tham khảo — References.....	317

Chương 1. Nhập môn kinh tế lượng và ôn tập xác suất thống kê

— Chapter 1. Introduction to econometrics and a review of probability and statistics

Mục tiêu chương — Chapter objectives

- Nêu được kinh tế lượng là gì, các bước của một nghiên cứu kinh tế lượng và ba loại dữ liệu: dữ liệu chéo, chuỗi thời gian, dữ liệu bảng.
Explain what econometrics is, the steps of an econometric study and the three types of data: cross-sectional, time-series and panel.
- Phân biệt tương quan với nhân quả; nhận ra tương quan giả giữa hai chuỗi có xu thế.
Distinguish correlation from causation; recognise spurious correlation between trending series.
- Tính tay kỳ vọng, phương sai, hiệp phương sai, hệ số tương quan của biến ngẫu nhiên và của mẫu số liệu.
Compute by hand the expectation, variance, covariance and correlation of random variables and of sample data.
- Tra bảng phân phối chuẩn, χ^2 , t, F; lập khoảng tin cậy; thực hiện kiểm định giả thuyết hai phía, một phía và kết luận bằng giá trị tới hạn hoặc p-value.
Read normal, χ^2 , t and F tables; build confidence intervals; carry out two-sided and one-sided tests and conclude with critical values or p-values.
- Biết câu lệnh hồi quy cơ bản trên EViews, Stata, R và Python statsmodels.
Know the basic regression commands in EViews, Stata, R and Python statsmodels.

1.1. Kinh tế lượng là gì — What is econometrics

Kinh tế lượng (econometrics, nghĩa đen là “đo lường kinh tế”) dùng lý thuyết kinh tế, công cụ toán học và phương pháp thống kê để lượng hóa các quan hệ kinh tế từ số liệu thực tế: ước lượng các tham số, kiểm định giả thuyết kinh tế và dự báo. Lý thuyết nói “giá tăng thì lượng cầu giảm”; kinh tế lượng trả lời “giảm bao nhiêu, với độ tin cậy nào”.

Econometrics (literally “economic measurement”) uses economic theory, mathematics and statistical methods to quantify economic relationships from real data: estimating parameters, testing economic hypotheses and forecasting. Theory says “demand falls when price rises”; econometrics answers “by how much, and with what confidence”.

Quan hệ với kinh tế học, toán học và thống kê — Links with economics, mathematics and statistics

Môn học <i>Discipline</i>	Cung cấp cho kinh tế lượng <i>What it gives econometrics</i>	Ví dụ <i>Example</i>
Kinh tế học <i>Economics</i>	Giả thuyết, biến cần đưa vào, dấu kỳ vọng của hệ số <i>Hypotheses, variables to include, expected signs</i>	Luật cầu: hệ số của giá mang dấu âm <i>Law of demand: negative price coefficient</i>

Môn học <i>Discipline</i>	Cung cấp cho kinh tế lượng <i>What it gives econometrics</i>	Ví dụ <i>Example</i>
Toán kinh tế <i>Mathematical economics</i>	Dạng hàm biểu diễn quan hệ (tuyến tính, mũ, log) <i>Functional form of the relationship (linear, power, log)</i>	Cobb–Douglas dạng log: $\ln Q = \ln A + \alpha \ln K + \beta \ln L$ <i>Cobb–Douglas in logs: $\ln Q = \ln A + \alpha \ln K + \beta \ln L$</i>
Thống kê kinh tế <i>Economic statistics</i>	Thu thập, xử lý, trình bày số liệu <i>Collecting, processing and presenting data</i>	Chỉ số giá tiêu dùng, GDP theo giá so sánh <i>Consumer price index, GDP at constant prices</i>
Thống kê toán <i>Mathematical statistics</i>	Ước lượng, kiểm định, phân phối xác suất <i>Estimation, testing, probability distributions</i>	Kiểm định t, khoảng tin cậy <i>t-test, confidence interval</i>
Kinh tế lượng <i>Econometrics</i>	Kết hợp cả bốn để có con số kiểm chứng được <i>Combines all four into testable numbers</i>	Giá tăng 1 nghìn đồng, lượng bán giảm 2,11 nghìn gói <i>A VND 1 thousand price rise cuts sales by 2.11 thousand packs</i>

Mô hình toán và mô hình kinh tế lượng — *Mathematical model and econometric model*

Mô hình toán là quan hệ tất định: biết X thì biết đúng Y . Số liệu thực tế không như vậy: cùng một mức giá, các tháng khác nhau vẫn có lượng bán khác nhau. Mô hình kinh tế lượng thêm **sai số ngẫu nhiên** u để gom mọi yếu tố không có trong mô hình.

*A mathematical model is deterministic: knowing X gives Y exactly. Real data do not behave like that: at the same price, different months still show different sales. An econometric model adds a **disturbance** u that collects every factor not in the model.*

$$Y = \beta_1 + \beta_2 X \quad (1.1)$$

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (1.2)$$

Y là biến phụ thuộc (biến được giải thích), X là biến độc lập (biến giải thích), β_1 là hệ số chặn, β_2 là hệ số góc, chỉ số i đánh dấu quan sát thứ i . β_1, β_2 là **tham số** của tổng thể, không quan sát được; kinh tế lượng ước lượng chúng bằng $\hat{\beta}_1, \hat{\beta}_2$ tính từ mẫu.

*Y is the dependent (explained) variable, X the independent (explanatory) variable, β_1 the intercept, β_2 the slope, and the subscript i marks observation i . β_1 and β_2 are population **parameters** and cannot be observed; econometrics estimates them by $\hat{\beta}_1, \hat{\beta}_2$ computed from a sample.*

PHẢI THUỘC / MEMORISE — Bốn nguồn gốc của sai số ngẫu nhiên u : (1) **biến bị bỏ sót** (thu nhập, thời tiết, giá đối thủ); (2) **sai số đo lường** của Y ; (3) **hành vi con người có tính ngẫu nhiên**; (4) **dạng hàm** chỉ là xấp xỉ của quan hệ thật. Ba mục tiêu của kinh tế lượng: **ước lượng** tham số, **kiểm định** giả thuyết, **dự báo** và phân tích chính sách.

*Four sources of the disturbance u : (1) **omitted variables** (income, weather, rivals' prices); (2) **measurement error in Y** ; (3) **inherent randomness of human behaviour**; (4) **the functional form is***

only an approximation of the true relationship. Three aims of econometrics: **estimate** parameters, **test** hypotheses, **forecast** and analyse policy.

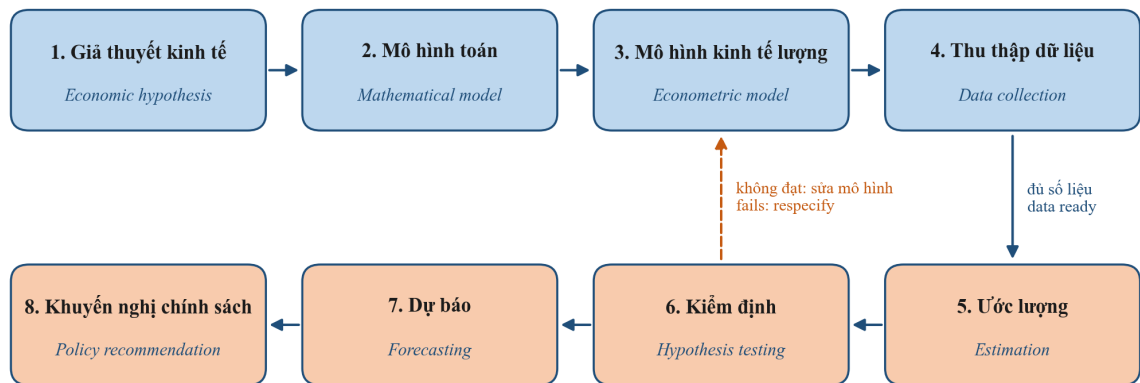
VÍ DỤ / EXAMPLE — Công ty TNHH Thực phẩm Mây Xanh bán sữa chua uống theo lốc 4 chai. Gọi Y là lượng bán (nghìn lốc/tháng), X là giá bán (nghìn đồng/lốc). Mô hình toán $Y = \beta_1 + \beta_2 X$ nói mọi tháng có cùng giá thì cùng lượng bán. Mô hình kinh tế lượng $Y_i = \beta_1 + \beta_2 X_i + u_i$ cho phép tháng 7/2025 bán 48 nghìn lốc ở giá 30 nghìn đồng, cao hơn tháng 6 (42 nghìn lốc ở giá 28 nghìn đồng), vì u_i chứa các yếu tố khác như nghỉ hè, khuyến mại của đối thủ. Kỳ vọng theo luật cầu: $\beta_2 < 0$.

Mây Xanh Food Co. Ltd sells drinking yoghurt in packs of 4 bottles. Let Y be sales (thousand packs per month) and X the price (VND thousand per pack). The mathematical model $Y = \beta_1 + \beta_2 X$ says every month with the same price has the same sales. The econometric model $Y_i = \beta_1 + \beta_2 X_i + u_i$ allows July 2025 to sell 48 thousand packs at VND 30 thousand, more than June (42 thousand packs at VND 28 thousand), because u_i holds other factors such as the summer holiday or a rival's promotion. Expected by the law of demand: $\beta_2 < 0$.

1.2. Quy trình nghiên cứu kinh tế lượng — The econometric research process

Một nghiên cứu kinh tế lượng đi qua tám bước. Bước 6 là chỗ rẽ: nếu mô hình vi phạm giả thiết hoặc hệ số trái dấu kỳ vọng mà không giải thích được, phải quay lại bước 3 để sửa mô hình chứ không đi tiếp sang dự báo.

An econometric study goes through eight steps. Step 6 is the branch point: if the model violates its assumptions or a coefficient has an unexplained wrong sign, go back to step 3 and respecify rather than move on to forecasting.



Hình 1.1. Quy trình tám bước của một nghiên cứu kinh tế lượng

Figure 1.1. The eight-step process of an econometric study

Bước Step	Việc phải làm Task	Ví dụ Mây Xanh Mây Xanh example
1. Giả thuyết <i>1. Hypothesis</i>	Phát biểu từ lý thuyết, có dấu kỳ vọng <i>State it from theory, with expected signs</i>	Giá tăng thì lượng bán giảm <i>Higher price, lower sales</i>
2. Mô hình toán <i>2. Mathematical model</i>	Chọn biến và dạng hàm <i>Choose variables and functional form</i>	$Y = \beta_1 + \beta_2 X$
3. Mô hình kinh tế lượng <i>3. Econometric model</i>	Thêm sai số u , nêu giả thiết về u <i>Add u, state assumptions on u</i>	$Y_i = \beta_1 + \beta_2 X_i + u_i$, kỳ vọng $\beta_2 < 0$ <i>$Y_i = \beta_1 + \beta_2 X_i + u_i$, expect $\beta_2 < 0$</i>
4. Dữ liệu <i>4. Data</i>	Thu thập, làm sạch, ghi đơn vị và nguồn <i>Collect, clean, record units and sources</i>	10 tháng, 01–10/2025 (bộ A) <i>10 months, 01–10/2025 (data set A)</i>
5. Ước lượng <i>5. Estimation</i>	Tính β bằng OLS trên phần mềm <i>Compute β by OLS in software</i>	$\beta_2 = -2,1127$ (Chương 2) <i>$\beta_2 = -2.1127$ (Chapter 2)</i>
6. Kiểm định <i>6. Testing</i>	Kiểm định hệ số, sự phù hợp, các giả thiết <i>Test coefficients, fit, assumptions</i>	$H_0: \beta_2 = 0$, $H_1: \beta_2 < 0$ (Chương 3) <i>$H_0: \beta_2 = 0$, $H_1: \beta_2 < 0$ (Chapter 3)</i>
7. Dự báo <i>7. Forecasting</i>	Dự báo điểm và khoảng <i>Point and interval forecasts</i>	Lượng bán khi giá 36 nghìn đồng <i>Sales at a price of VND 36 thousand</i>
8. Khuyến nghị <i>8. Recommendation</i>	Diễn giải bằng ngôn ngữ kinh tế, đề xuất <i>Interpret in economic terms, recommend</i>	Có nên tăng giá hay không <i>Whether to raise the price</i>

LỖI HAY GẶP / COMMON MISTAKE — Bỏ qua bước 1–3: đưa mọi biến có sẵn vào phần mềm rồi giữ lại biến nào có p-value nhỏ. Bài làm như vậy không nêu được dấu kỳ vọng nên không đánh giá được hệ số trái dấu, và dễ giữ lại các tương quan ngẫu nhiên. Đề thi thường cho điểm riêng ý “nêu kỳ vọng dấu của hệ số và giải thích theo lý thuyết”.
Skipping steps 1–3: putting every available variable into the software and keeping those with small p-values. Such work cannot state expected signs, so it cannot judge a wrong-signed coefficient, and it tends to keep chance correlations. Exams often give separate marks for “state the expected sign of the coefficient and justify it from theory”.

1.3. Dữ liệu trong kinh tế lượng — Data in econometrics

Ba loại dữ liệu — Three types of data

Dữ liệu chéo quan sát nhiều đơn vị (cá nhân, hộ, doanh nghiệp, tỉnh) tại cùng một thời điểm; chỉ số i . **Dữ liệu chuỗi thời gian** quan sát một đơn vị qua nhiều thời kỳ; chỉ số t .

thứ tự quan sát có ý nghĩa. **Dữ liệu bảng** quan sát cùng một nhóm đơn vị qua nhiều thời kỳ; chỉ số kép i, t . Dữ liệu chéo gộp (pooled cross-section) là các mẫu chéo ở nhiều năm nhưng không theo dõi cùng một đơn vị.

Cross-sectional data observe many units (individuals, households, firms, provinces) at one point in time; index i . Time-series data observe one unit over many periods; index t , and the order matters. Panel data follow the same units over several periods; double index i, t . A pooled cross-section stacks cross-sections from several years without tracking the same units.

Tiêu chí <i>Criterion</i>	Dữ liệu chéo <i>Cross-section</i>	Chuỗi thời gian <i>Time series</i>	Dữ liệu bảng <i>Panel</i>
Đơn vị $\times$ thời gian <i>Units $\times$ time</i>	Nhiều $\times$ một <i>Many $\times$ one</i>	Một $\times$ nhiều <i>One $\times$ many</i>	Nhiều $\times$ nhiều <i>Many $\times$ many</i>
Ký hiệu <i>Notation</i>	$Y_i, i = 1, \dots, n$	$Y_t, t = 1, \dots, T$	Y_{it}
Ví dụ trong sách <i>Example in this book</i>	Bộ B: 36 cửa hàng Hạt Nâu, tháng 9/2026 <i>Set B: 36 Hạt Nâu shops, September 2026</i>	Bộ E: Sao Bắc, 40 quý 2016Q1–2025Q4 <i>Set E: Sao Bắc, 40 quarters 2016Q1–2025Q4</i>	36 cửa hàng $\times$ 12 tháng = 432 quan sát <i>36 shops $\times$ 12 months = 432 observations</i>
Ví dụ Việt Nam <i>Vietnamese example</i>	Khảo sát mức sống dân cư một năm <i>One round of the household living standards survey</i>	GDP, CPI theo quý <i>Quarterly GDP, CPI</i>	Báo cáo tài chính các công ty niêm yết qua nhiều năm <i>Listed firms' financial statements over years</i>
Vấn đề hay gặp <i>Typical problem</i>	Phương sai sai số thay đổi (Chương 7) <i>Heteroscedasticity (Chapter 7)</i>	Tự tương quan, xu thế (Chương 8) <i>Autocorrelation, trends (Chapter 8)</i>	Cả hai, cộng thêm khác biệt riêng của từng đơn vị <i>Both, plus unit-specific effects</i>

Thang đo của biến — *Measurement scales*

Thang đo <i>Scale</i>	Đặc điểm <i>Feature</i>	Ví dụ <i>Example</i>	Dùng trong hồi quy <i>Use in regression</i>
Định danh <i>Nominal</i>	Chỉ phân loại, không thứ tự <i>Categories only, no order</i>	Khu vực, ngành <i>Region, industry</i>	Biến giả 0/1 (Chương 5) <i>Dummy variables (Chapter 5)</i>
Thứ bậc <i>Ordinal</i>	Có thứ tự, khoảng cách không đều <i>Ordered, unequal gaps</i>	Mức hài lòng 1–5 <i>Satisfaction 1–5</i>	Biến giả hoặc mô hình thứ bậc <i>Dummies or ordered models</i>

Thang đo <i>Scale</i>	Đặc điểm <i>Feature</i>	Ví dụ <i>Example</i>	Dùng trong hồi quy <i>Use in regression</i>
Khoảng <i>Interval</i>	Hiệu số có nghĩa, không có số 0 tuyệt đối <i>Differences meaningful, no true zero</i>	Nhiệt độ °C, năm <i>Temperature in °C, calendar year</i>	Dùng trực tiếp, không lấy log <i>Used directly, no logs</i>
Tỉ lệ <i>Ratio</i>	Có số 0 tuyệt đối, tỉ số có nghĩa <i>True zero, ratios meaningful</i>	Doanh thu, giá, thu nhập <i>Revenue, price, income</i>	Dùng trực tiếp hoặc log <i>Used directly or in logs</i>

Nguồn dữ liệu Việt Nam — Vietnamese data sources

Cơ quan <i>Agency</i>	Loại dữ liệu chính <i>Main data</i>	Tần suất <i>Frequency</i>
Cục Thống kê (Bộ Tài chính) <i>National Statistics Office (Ministry of Finance)</i>	Niên giám thống kê, GDP, CPI, sản xuất công nghiệp, khảo sát mức sống dân cư, điều tra lao động – việc làm, điều tra doanh nghiệp <i>Statistical yearbook, GDP, CPI, industrial output, household living standards survey, labour force survey, enterprise census</i>	Tháng, quý, năm <i>Monthly, quarterly, annual</i>
Ngân hàng Nhà nước Việt Nam <i>State Bank of Vietnam</i>	Lãi suất, tỷ giá, tăng trưởng tín dụng, tổng phương tiện thanh toán <i>Interest rates, exchange rates, credit growth, money supply</i>	Ngày, tháng <i>Daily, monthly</i>
Cục Hải quan (Bộ Tài chính) <i>Customs Department (Ministry of Finance)</i>	Kim ngạch xuất khẩu, nhập khẩu theo mặt hàng, thị trường <i>Exports and imports by product and partner</i>	Tháng <i>Monthly</i>
Bộ Tài chính <i>Ministry of Finance</i>	Thu, chi ngân sách nhà nước, nợ công <i>State budget revenue and spending, public debt</i>	Tháng, năm <i>Monthly, annual</i>
Ủy ban Chứng khoán Nhà nước, các sở giao dịch chứng khoán <i>State Securities Commission, stock exchanges</i>	Giá, khối lượng giao dịch cổ phiếu; báo cáo tài chính doanh nghiệp niêm yết <i>Share prices and volumes; listed firms' financial statements</i>	Ngày, quý <i>Daily, quarterly</i>
Ngân hàng Thế giới (World Bank WDI) <i>World Bank (WDI)</i>	Chuỗi vĩ mô so sánh được giữa các nước, tải qua API <i>Cross-country comparable macro series, downloadable via API</i>	Năm <i>Annual</i>

Trước khi đưa số liệu vào mô hình cần ghi rõ: đơn vị đo; giá hiện hành hay giá so sánh (năm gốc nào); tần suất; thời điểm thay đổi phương pháp tính hoặc thay đổi địa giới hành chính. Sau sắp xếp đơn vị hành chính ngày 1/7/2025 (34 tỉnh, thành phố), chuỗi số liệu cấp tỉnh trước và sau mốc này phải được ghép lại theo địa giới mới trước khi so sánh.

Before data go into a model, record: units; current or constant prices (which base year); frequency; dates of methodological or boundary changes. After the administrative reorganisation of 1 July 2025 (34 provinces and cities), provincial series before and after that date must be re-aggregated to the new boundaries before comparison.

LỖI HAY GẶP / COMMON MISTAKE — Trộn GDP giá hiện hành với tiêu dùng giá so sánh trong cùng một mô hình; hoặc ghép chuỗi từ hai nguồn có năm gốc khác nhau. Hệ số ước lượng khi đó lẫn biến động giá vào quan hệ thực. Luôn ghi năm gốc vào tên biến (ví dụ: GDP giá so sánh 2010).

Mixing GDP at current prices with consumption at constant prices in one model, or splicing series from two sources with different base years. The estimated coefficient then picks up price movements, not just the real relationship. Always put the base year in the variable name (e.g. GDP at 2010 prices).

1.4. Tương quan và nhân quả — Correlation and causation

Hồi quy cho biết Y thay đổi bao nhiêu khi X thay đổi một đơn vị **trong số liệu**. Nói X **gây ra** sự thay đổi của Y là khẳng định mạnh hơn: cần lý thuyết kinh tế, cần loại trừ biến thứ ba và nhân quả ngược. Một hệ số hồi quy có ý nghĩa thống kê chưa đủ để kết luận nhân quả.

A regression tells how much Y changes when X changes by one unit in the data. Saying X causes the change in Y is a stronger claim: it needs economic theory and the exclusion of third variables and reverse causality. A statistically significant regression coefficient is not enough to conclude causation.

Tiêu chí <i>Criterion</i>	Tương quan <i>Correlation</i>	Nhân quả <i>Causation</i>
Nội dung <i>Meaning</i>	Hai biến cùng biến động <i>Two variables move together</i>	Thay đổi X làm Y thay đổi <i>Changing X changes Y</i>
Chiều <i>Direction</i>	Đối xứng: $r(X, Y) = r(Y, X)$ <i>Symmetric: $r(X, Y) = r(Y, X)$</i>	Có chiều: X tác động lên Y <i>Directional: X acts on Y</i>
Đo bằng <i>Measured by</i>	Hệ số tương quan r <i>Correlation coefficient r</i>	Hệ số hồi quy khi đã kiểm soát các biến khác, thực nghiệm <i>Regression coefficient with controls, experiments</i>
Đủ để kết luận khi <i>Sufficient when</i>	Có số liệu <i>Data are available</i>	Có lý thuyết, đúng thứ tự thời gian, loại được biến thứ ba <i>Theory, correct timing, third variables ruled out</i>

Ba trường hợp tương quan không phải nhân quả — Three cases of correlation without causation

- **Biến thứ ba.** Doanh số kem và số người đi bơi cùng tăng vào mùa hè; cả hai do nhiệt độ quyết định. Hồi quy doanh số kem theo số người đi bơi cho hệ số dương nhưng không có ý nghĩa nhân quả.

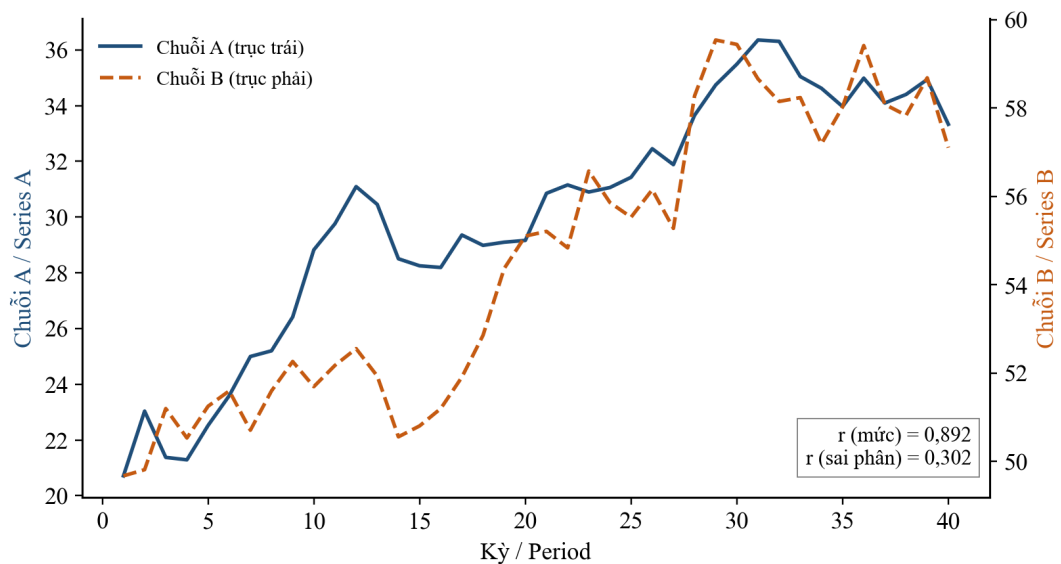
Third variable. Ice-cream sales and swimmer numbers both rise in summer; both are driven by temperature. Regressing ice-cream sales on swimmers gives a positive coefficient with no causal meaning.

- **Nhân quả ngược.** Doanh nghiệp đặt ngân sách quảng cáo bằng một tỉ lệ cố định của doanh thu dự kiến. Tương quan dương giữa quảng cáo và doanh thu khi đó một phần là do doanh thu quyết định quảng cáo.

Reverse causality. A firm sets its advertising budget as a fixed share of expected revenue. The positive correlation between advertising and revenue is then partly revenue driving advertising.

- **Tương quan giả giữa hai chuỗi có xu thế.** Hai chuỗi thời gian cùng tăng theo thời gian sẽ có r gần 1 dù không liên quan gì tới nhau. Hình 1.2 vẽ hai chuỗi mô phỏng độc lập, mỗi chuỗi là một bước ngẫu nhiên (giá trị kỳ này = giá trị kỳ trước + nhiễu).

Spurious correlation between trending series. Two time series that both drift upward will show r close to 1 even when unrelated. Figure 1.2 plots two independent simulated series, each a random walk (this period's value = last period's value + noise).



Hình 1.2. Hai bước ngẫu nhiên độc lập vẫn có hệ số tương quan cao (chuỗi mô phỏng)

Figure 1.2. Two independent random walks still show a high correlation (simulated series)

Hai chuỗi được sinh độc lập nên tương quan tổng thể bằng 0, nhưng trên 40 kỳ $r = 0,892$. Lấy sai phân (mức thay đổi từng kỳ) để bỏ xu thế thì r chỉ còn 0,302. Với chuỗi thời gian, r cao giữa hai biến ở dạng mức chưa nói lên điều gì; Chương 8 và Chương 9 trở lại vấn đề này.

The two series were generated independently, so the population correlation is zero, yet over 40 periods $r = 0.892$. Taking first differences (period-to-period changes) removes the trend and r drops to 0.302. With time series, a high r between two variables in levels says nothing on its own; Chapters 8 and 9 return to this issue.

LỖI HAY GẶP / COMMON MISTAKE — Viết “ $r = 0,95$ nên X là nguyên nhân làm tăng Y” hoặc “ R^2 cao nên mô hình đúng”. Hệ số tương quan chỉ đo mức độ cùng biến động tuyến tính; kết luận về tác động phải dựa vào hệ số hồi quy trong mô hình được xây dựng từ lý

thuyết, đã kiểm soát các biến liên quan.

Writing “ $r = 0.95$, so X causes Y to rise” or “ R^2 is high, so the model is right”. The correlation coefficient only measures linear co-movement; conclusions about effects must rest on regression coefficients in a theory-based model that controls for the relevant variables.

1.5. Ôn tập: biến ngẫu nhiên, kỳ vọng, phương sai, hiệp phương sai — Review: random variables, expectation, variance, covariance

Biến ngẫu nhiên và phân phối xác suất — Random variables and probability distributions

Biến ngẫu nhiên nhận giá trị bằng số tùy theo kết quả của một phép thử. Biến **rời rạc** nhận hữu hạn hoặc đếm được giá trị (số đơn hàng, số khách); biến **liên tục** nhận mọi giá trị trong một khoảng (thời gian giao hàng, lợi suất). Biến rời rạc mô tả bằng bảng phân phối xác suất $p_i = P(X = x_i)$ với $\sum p_i = 1$; biến liên tục mô tả bằng hàm mật độ $f(x)$ với diện tích dưới đường mật độ bằng 1.

A random variable takes numerical values depending on the outcome of an experiment. A discrete variable takes finitely or countably many values (orders, customers); a continuous variable takes any value in an interval (delivery time, return). A discrete variable is described by its probability table $p_i = P(X = x_i)$ with $\sum p_i = 1$; a continuous one by a density $f(x)$ whose total area equals 1.

$$E(X) = \mu = \sum x_i p_i \quad (1.3)$$

$$\text{Var}(X) = \sigma^2 = E[(X - \mu)^2] = E(X^2) - [E(X)]^2 \quad (1.4)$$

$$E(aX + b) = aE(X) + b; \text{Var}(aX + b) = a^2 \text{Var}(X) \quad (1.5)$$

Kỳ vọng là giá trị trung bình theo xác suất; phương sai đo độ phân tán quanh kỳ vọng, đơn vị là bình phương đơn vị của X ; độ lệch chuẩn $\sigma = \sqrt{\text{Var}(X)}$ cùng đơn vị với X . Cộng thêm hằng số b dời kỳ vọng nhưng không làm đổi phương sai.

The expectation is the probability-weighted mean; the variance measures spread around it in squared units of X ; the standard deviation $\sigma = \sqrt{\text{Var}(X)}$ is in the units of X . Adding a constant b shifts the mean but leaves the variance unchanged.

Bảng 1.1. Số đơn hàng bị trả lại trong một ngày — Table 1.1. Returned orders per day

x_i (đơn) x_i (orders)	0	1	2	3	4	Tổng Total
p_i	0,30 <i>0.30</i>	0,35 <i>0.35</i>	0,20 <i>0.20</i>	0,10 <i>0.10</i>	0,05 <i>0.05</i>	1
$x_i p_i$	0,00 <i>0.00</i>	0,35 <i>0.35</i>	0,40 <i>0.40</i>	0,30 <i>0.30</i>	0,20 <i>0.20</i>	1,25 <i>1.25</i>
$x_i^2 p_i$	0,00 <i>0.00</i>	0,35 <i>0.35</i>	0,80 <i>0.80</i>	0,90 <i>0.90</i>	0,80 <i>0.80</i>	2,85 <i>2.85</i>

VÍ DỤ / EXAMPLE — Kho thương mại điện tử của Công ty Thương mại Lam Sơn ghi nhận số đơn hàng bị trả lại mỗi ngày theo Bảng 1.1 (dữ kiện đề bài). Mỗi đơn trả lại tốn 35 nghìn

đồng phí vận chuyển ngược, chi phí cố định xử lý hàng trả là 50 nghìn đồng/ngày. Tính kỳ vọng, độ lệch chuẩn của số đơn trả lại và của chi phí $C = 50 + 35X$.

The e-commerce warehouse of Lam Son Trading records returned orders per day as in Table 1.1 (given data). Each return costs VND 35 thousand in reverse shipping, and the fixed cost of handling returns is VND 50 thousand a day. Find the expectation and standard deviation of returned orders and of the cost $C = 50 + 35X$.

1. $E(X) = \sum x_i p_i = 1,25$ đơn/ngày.
 $E(X) = \sum x_i p_i = 1.25$ orders per day.
2. $E(X^2) = \sum x_i^2 p_i = 2,85$; $\text{Var}(X) = 2,85 - 1,25^2 = 2,85 - 1,5625 = 1,2875$; $\sigma = \sqrt{1,2875} = 1,1347$ đơn.
 $E(X^2) = \sum x_i^2 p_i = 2.85$; $\text{Var}(X) = 2.85 - 1.25^2 = 2.85 - 1.5625 = 1.2875$; $\sigma = \sqrt{1.2875} = 1.1347$ orders.
3. Theo (1.5): $E(C) = 50 + 35 \times 1,25 = 93,75$ nghìn đồng; $\text{Var}(C) = 35^2 \times 1,2875 = 1.577,1875$; $\sigma(C) = 35 \times 1,1347 = 39,71$ nghìn đồng. Chi phí cố định 50 không ảnh hưởng tới độ lệch chuẩn.
By (1.5): $E(C) = 50 + 35 \times 1.25 = \text{VND } 93.75$ thousand; $\text{Var}(C) = 35^2 \times 1.2875 = 1,577.1875$; $\sigma(C) = 35 \times 1.1347 = \text{VND } 39.71$ thousand. The fixed cost of 50 does not affect the standard deviation.

Hiệp phương sai và hệ số tương quan của hai biến ngẫu nhiên — *Covariance and correlation of two random variables*

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y) \quad (1.6)$$

$$\rho_{XY} = \text{Cov}(X, Y) / (\sigma_X \sigma_Y), \quad -1 \leq \rho_{XY} \leq 1 \quad (1.7)$$

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \cdot \text{Cov}(X, Y) \quad (1.8)$$

$\text{Cov} > 0$: hai biến có xu hướng cùng lớn hơn hoặc cùng nhỏ hơn kỳ vọng của chúng; $\text{Cov} < 0$: ngược chiều. Cov phụ thuộc đơn vị đo nên không so sánh được giữa các cặp biến; ρ chuẩn hóa về khoảng $[-1; 1]$, không có đơn vị. $\rho = \pm 1$ khi Y là hàm tuyến tính chính xác của X . Hai biến độc lập thì $\text{Cov} = 0$, nhưng $\text{Cov} = 0$ không suy ra độc lập: nếu X nhận $-1, 0, 1$ với xác suất bằng nhau và $Y = X^2$ thì $\text{Cov}(X, Y) = E(X^3) - E(X)E(X^2) = 0$ dù Y được xác định duy nhất bởi X . ρ chỉ đo quan hệ **tuyến tính**.

*$\text{Cov} > 0$: the two variables tend to be above or below their means together; $\text{Cov} < 0$: in opposite directions. Covariance depends on units and cannot be compared across pairs; ρ rescales it to $[-1; 1]$ and is unit-free. $\rho = \pm 1$ when Y is an exact linear function of X . Independence implies $\text{Cov} = 0$, but $\text{Cov} = 0$ does not imply independence: if X takes $-1, 0, 1$ with equal probability and $Y = X^2$, then $\text{Cov}(X, Y) = E(X^3) - E(X)E(X^2) = 0$ although Y is fully determined by X . ρ measures only **linear** association.*

Bảng 1.2. Lợi suất hai dự án theo trạng thái nền kinh tế (%/năm) — Table 1.2.
Returns of two projects by state of the economy (% a year)

Trạng thái <i>State</i>	Xác suất <i>Probability</i>	Dự án X <i>Project X</i>	Dự án Y <i>Project Y</i>
Suy giảm <i>Downturn</i>	0,2 <i>0.2</i>	-4	2
Bình thường <i>Normal</i>	0,5 <i>0.5</i>	8	6
Tăng trưởng <i>Expansion</i>	0,3 <i>0.3</i>	16	7

VÍ DỤ / EXAMPLE — Quỹ đầu tư Bến Thành cân nhắc hai dự án có lợi suất theo Bảng 1.2 (dữ kiện đề bài). Tính kỳ vọng, phương sai, hiệp phương sai, hệ số tương quan và phương sai của danh mục chia đều vốn 50% – 50%.

Bến Thành Investment Fund considers two projects with the returns in Table 1.2 (given data). Compute the expectations, variances, covariance, correlation and the variance of a 50%-50% portfolio.

- $E(X) = 0,2 \times (-4) + 0,5 \times 8 + 0,3 \times 16 = 8,0\%$; $E(Y) = 0,2 \times 2 + 0,5 \times 6 + 0,3 \times 7 = 5,5\%$.
 $E(X) = 0.2 \times (-4) + 0.5 \times 8 + 0.3 \times 16 = 8.0\%$; $E(Y) = 0.2 \times 2 + 0.5 \times 6 + 0.3 \times 7 = 5.5\%$.
- Độ lệch so với kỳ vọng: $X - E(X) = -12,0; 0,0; 8,0$; $Y - E(Y) = -3,5; 0,5; 1,5$.
 $Var(X) = \sum p(X - E(X))^2 = 48,00$; $Var(Y) = 3,25$.
Deviations from the means: $X - E(X) = -12.0; 0.0; 8.0$; $Y - E(Y) = -3.5; 0.5; 1.5$. $Var(X) = \sum p(X - E(X))^2 = 48.00$; $Var(Y) = 3.25$.
- $Cov(X, Y) = \sum p(X - E(X))(Y - E(Y)) = 12,00$; $\rho = 12,00 / \sqrt{(48,00 \times 3,25)} = \mathbf{0,9608}$: hai dự án biến động cùng chiều, khá chặt.
 $Cov(X, Y) = \sum p(X - E(X))(Y - E(Y)) = 12.00$; $\rho = 12.00 / \sqrt{(48.00 \times 3.25)} = \mathbf{0.9608}$: the projects move together fairly closely.
- Danh mục $P = 0,5X + 0,5Y$: $E(P) = 6,75\%$; theo (1.8) $Var(P) = 0,25 \times 48,00 + 0,25 \times 3,25 + 2 \times 0,25 \times 12,00 = \mathbf{18,8125}$, $\sigma(P) = 4,337\%$. Nếu quên số hạng hiệp phương sai sẽ ra 12,8125, thấp hơn rủi ro thật.
 $Portfolio P = 0.5X + 0.5Y$: $E(P) = 6.75\%$; by (1.8) $Var(P) = 0.25 \times 48.00 + 0.25 \times 3.25 + 2 \times 0.25 \times 12.00 = \mathbf{18.8125}$, $\sigma(P) = 4.337\%$. Forgetting the covariance term gives 12.8125, understating the true risk.

Kỳ vọng có điều kiện — Conditional expectation

$E(Y | X = x)$ là giá trị trung bình của Y trong nhóm có $X = x$. Khi X thay đổi, $E(Y | X)$ vẽ thành một đường; đường đó là **hàm hồi quy tổng thể** (PRF) của Chương 2. Ví dụ: lượng bán trung bình của mọi tháng có giá 25 nghìn đồng/lốc là $E(Y | X = 25)$. Bài tập 2 cuối chương tính $E(Y | X)$ từ một bảng phân phối đồng thời.

$E(Y | X = x)$ is the mean of Y in the group with $X = x$. As X varies, $E(Y | X)$ traces a line; that line is the **population regression function (PRF)** of Chapter 2. Example: average sales over all months priced at VND 25 thousand per pack is $E(Y | X = 25)$. Exercise 2 at the end of the chapter computes $E(Y | X)$ from a joint distribution table.

$$E(Y | X = x) = \sum_j y_j \cdot P(Y = y_j | X = x), \quad P(Y = y | X = x) = P(X = x, Y = y) / P(X = x) \quad (1.9)$$

Thống kê mẫu: trung bình, phương sai, hiệp phương sai, hệ số tương quan —
Sample statistics: mean, variance, covariance, correlation

Trong thực tế ta không biết phân phối tổng thể mà chỉ có mẫu n quan sát. Đặt $x_i = X_i - \bar{X}$ và $y_i = Y_i - \bar{Y}$ (chữ thường là độ lệch so với trung bình; ký hiệu này dùng suốt cuốn sách). Các thống kê mẫu chia cho $n - 1$ để là ước lượng không chệch của tham số tổng thể.

In practice the population distribution is unknown and only a sample of n observations is available. Let $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$ (lower case denotes deviations from the mean; this notation is used throughout the book). Sample statistics divide by $n - 1$ so that they are unbiased estimators of the population parameters.

$$\bar{X} = \sum X_i / n; \quad S_x^2 = \sum x_i^2 / (n - 1) \quad (1.10)$$

$$S_{xy} = \sum x_i y_i / (n - 1); \quad r = \sum x_i y_i / \sqrt{(\sum x_i^2 \cdot \sum y_i^2)} = S_{xy} / (S_x S_y) \quad (1.11)$$

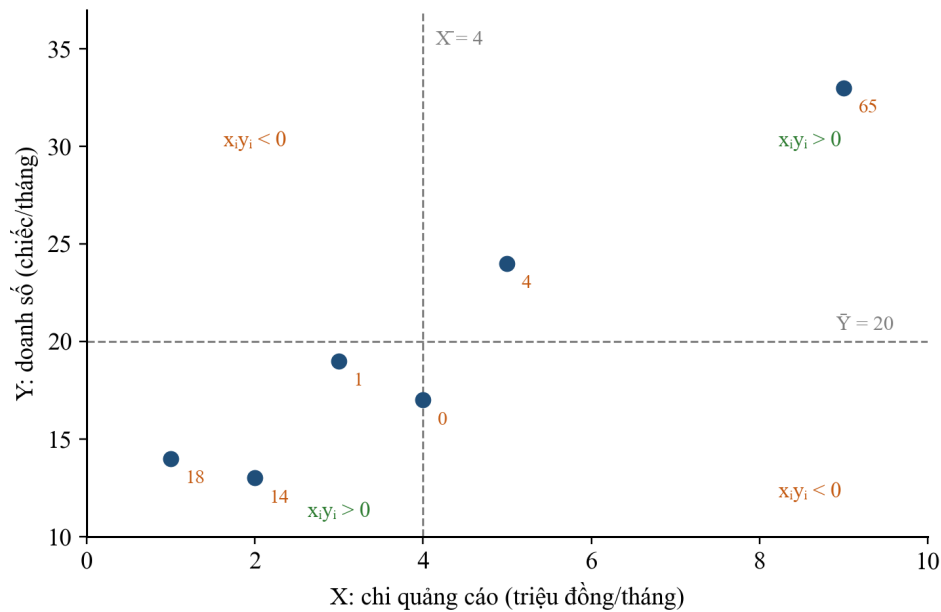
Bảng 1.3. Bài tính tay: chi quảng cáo và doanh số của 6 cửa hàng — *Table 1.3. Hand calculation: advertising and sales of 6 shops*

Cửa hàng Shop	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	1	14	-3	-6	18	9	36
2	2	13	-2	-7	14	4	49
3	3	19	-1	-1	1	1	1
4	4	17	0	-3	0	0	9
5	5	24	1	4	4	1	16
6	9	33	5	13	65	25	169
Tổng Sum	24	120	0	0	102	40	280

VÍ DỤ / EXAMPLE — Chuỗi cửa hàng điện thoại Ánh Dương có 6 cửa hàng; X là chi quảng cáo (triệu đồng/tháng), Y là doanh số (chiếc/tháng), số liệu ở Bảng 1.3 (dữ kiện đề bài). Tính hiệp phương sai mẫu và hệ số tương quan.

The Ánh Dương phone retailer has 6 shops; X is advertising spend (VND million per month), Y is unit sales (phones per month), data in Table 1.3 (given data). Compute the sample covariance and the correlation coefficient.

1. $\bar{X} = 24/6 = 4$; $\bar{Y} = 120/6 = 20$. Lập cột x_i, y_i ; kiểm $\sum x_i = \sum y_i = 0$.
 $\bar{X} = 24/6 = 4$; $\bar{Y} = 120/6 = 20$. Build the x_i, y_i columns; check $\sum x_i = \sum y_i = 0$.
2. $\sum x_i y_i = 102$; $\sum x_i^2 = 40$; $\sum y_i^2 = 280$. $S_{xy} = 102/5 = 20,4$; $S_x = \sqrt{(40/5)} = 2,8284$; $S_y = \sqrt{(280/5)} = 7,4833$.
 $\sum x_i y_i = 102$; $\sum x_i^2 = 40$; $\sum y_i^2 = 280$. $S_{xy} = 102/5 = 20,4$; $S_x = \sqrt{(40/5)} = 2.8284$; $S_y = \sqrt{(280/5)} = 7.4833$.
3. $r = 102/\sqrt{(40 \times 280)} = 102/105,8301 = \mathbf{0,9638}$. Quan hệ tuyến tính cùng chiều, chặt: cửa hàng chi quảng cáo nhiều hơn có doanh số cao hơn.
 $r = 102/\sqrt{(40 \times 280)} = 102/105.8301 = \mathbf{0.9638}$. A strong positive linear relationship: shops that advertise more sell more.



Hình 1.3. Dấu của tích $x_i y_i$ theo bốn góc phần tư quanh điểm $(\bar{X}, \bar{Y})$ — số cạnh mỗi điểm là $x_i y_i$
Figure 1.3. Sign of $x_i y_i$ in the four quadrants around $(\bar{X}, \bar{Y})$ — the number beside each point is $x_i y_i$

Điểm nằm ở góc trên phải và dưới trái góp tích $x_i y_i$ dương; góc trên trái và dưới phải góp tích âm. $\sum x_i y_i$ dương nghĩa là phần lớn khối lượng nằm ở hai góc cùng dấu. Cửa hàng 6 ($X = 9$) góp tích lớn nhất, 65 trên tổng 102: một quan sát xa trung bình có thể chi phối r .

Points in the upper-right and lower-left quadrants contribute positive $x_i y_i$; upper-left and lower-right contribute negative products. A positive $\sum x_i y_i$ means most of the weight lies in the same-sign quadrants. Shop 6 ($X = 9$) contributes the largest product, 65 out of 102: one observation far from the mean can dominate r .

Kiểm định ý nghĩa của hệ số tương quan — Testing the significance of a correlation

$$t = r\sqrt{(n-2) / \sqrt{(1-r^2)}} \sim t(n-2) \text{ khi } H_0: \rho = 0 \quad (1.12)$$

VÍ DỤ / EXAMPLE — Với Bảng 1.3: $t = 0,9638 \times \sqrt{4/\sqrt{(1-0,9289)}} = 7,231$. Tra $t_{0,025(4)} = 2,776$. $|t| > 2,776$ nên bác bỏ $H_0: \rho = 0$ ở mức 5%. Trên bộ A của Mây Xanh (giá và lượng bán, 10 tháng), $r = -0,9622$, $r^2 = 0,9258$: đúng bằng $R^2 = 0,9258$ của hồi quy hai biến ở Chương 2.

For Table 1.3: $t = 0.9638 \times \sqrt{4/\sqrt{1 - 0.9289}} = 7.231$. Table value $t_{0.025}^{(4)} = 2.776$. $|t| > 2.776$, so reject $H_0: \rho = 0$ at 5%. On data set A of *Mây Xanh* (price and sales, 10 months), $r = -0.9622$ and $r^2 = 0.9258$: exactly the $R^2 = 0.9258$ of the two-variable regression in Chapter 2.

LỖI HAY GẶP / COMMON MISTAKE — Chia cho n thay vì $n - 1$ khi tính phương sai, hiệp phương sai mẫu (máy tính bỏ túi có hai phím σ_n và σ_{n-1}). Hệ số r không bị ảnh hưởng vì tử và mẫu cùng chia, nhưng S_x , S_y , S_{xy} sai. Lỗi thứ hai: viết $\text{Var}(X - Y) = \text{Var}(X) - \text{Var}(Y)$. Đúng theo (1.8) với $a = 1$, $b = -1$: $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$; phương sai không bao giờ âm.

Dividing by n instead of $n - 1$ for the sample variance or covariance (calculators have separate σ_n and σ_{n-1} keys). r is unaffected because numerator and denominator are divided alike, but S_x , S_y and S_{xy} are wrong. A second slip: writing $\text{Var}(X - Y) = \text{Var}(X) - \text{Var}(Y)$. By (1.8) with $a = 1$, $b = -1$: $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$; a variance is never negative.

1.6. Các phân phối xác suất dùng trong kinh tế lượng — Probability distributions used in econometrics

Phân phối chuẩn — The normal distribution

$X \sim N(\mu, \sigma^2)$ có đường mật độ hình chuông, đối xứng quanh μ ; khoảng $\mu \pm 1,96\sigma$ chứa 95% xác suất. Chuẩn hóa đưa mọi phân phối chuẩn về phân phối chuẩn tắc $N(0, 1)$. Tổ hợp tuyến tính của các biến chuẩn độc lập vẫn có phân phối chuẩn. **Định lý giới hạn trung tâm**: trung bình mẫu của n quan sát độc lập, cùng phân phối (phương sai hữu hạn) có phân phối xấp xỉ chuẩn khi n đủ lớn, bất kể phân phối gốc.

*$X \sim N(\mu, \sigma^2)$ has a bell-shaped density symmetric about μ ; the interval $\mu \pm 1.96\sigma$ holds 95% of the probability. Standardisation turns any normal variable into the standard normal $N(0, 1)$. A linear combination of independent normal variables is again normal. **Central limit theorem**: the mean of n independent, identically distributed observations (finite variance) is approximately normal for large n , whatever the parent distribution.*

$$Z = (X - \mu) / \sigma \sim N(0, 1) \quad (1.13)$$

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = (\bar{X} - \mu) / (\sigma/\sqrt{n}) \sim N(0, 1) \quad (1.14)$$

Phân phối χ^2 , Student t và Fisher F — The χ^2 , Student t and Fisher F distributions

Ba phân phối sau đều dựng từ biến chuẩn tắc. Cho $Z_1, \dots, Z_k$ độc lập, cùng phân phối $N(0, 1)$; $V \sim \chi^2(k)$ và Z độc lập với V ; $V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$ độc lập. Tham số k (k_1, k_2) gọi là **bậc tự do**.

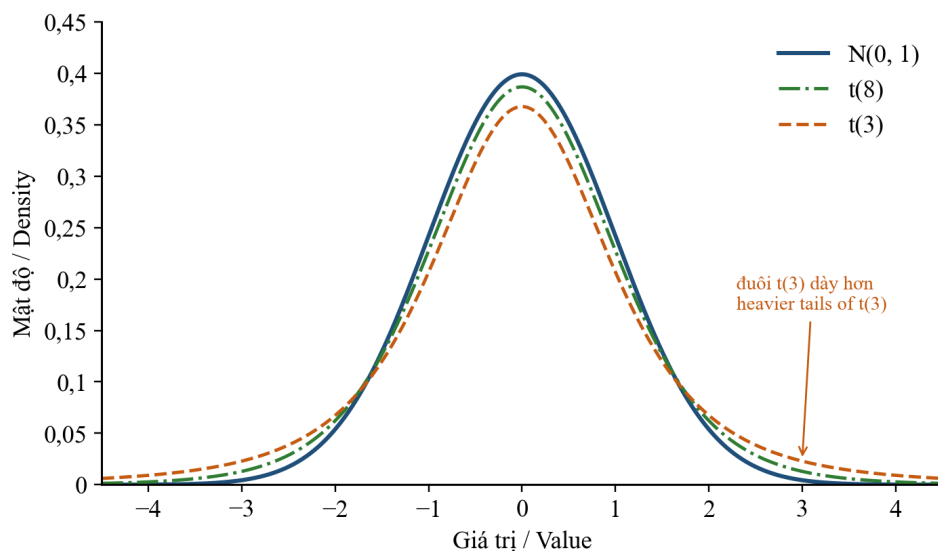
*The next three distributions are all built from standard normals. Let $Z_1, \dots, Z_k$ be independent $N(0, 1)$; $V \sim \chi^2(k)$ with Z independent of V ; $V_1 \sim \chi^2(k_1)$ and $V_2 \sim \chi^2(k_2)$ independent. The parameter k (k_1, k_2) is called the **degrees of freedom**.*

$$\chi^2(k) = Z_1^2 + Z_2^2 + \dots + Z_k^2 \quad (1.15)$$

$$t(k) = Z / \sqrt{V/k} \quad (1.16)$$

$$F(k_1, k_2) = (V_1/k_1) / (V_2/k_2) \quad (1.17)$$

Phân phối <i>Distribution</i>	Hình dạng <i>Shape</i>	Kỳ vọng <i>Mean</i>	Dùng ở đâu trong kinh tế lượng <i>Use in econometrics</i>
$N(0, 1)$	Đối xứng quanh 0 <i>Symmetric about 0</i>	0	Mẫu lớn, kiểm định z, Jarque–Bera dựa trên giả thiết chuẩn <i>Large samples, z-tests, normality checks</i>
$\chi^2(k)$	Chỉ nhận giá trị dương, lệch phải <i>Positive only, right-skewed</i>	k	Khoảng tin cậy σ^2 , kiểm định White, BG, Jarque–Bera <i>CI for σ^2, White, BG, Jarque–Bera tests</i>
$t(k)$	Đối xứng, đuôi dày hơn $N(0, 1)$ <i>Symmetric, heavier tails than $N(0, 1)$</i>	0 ($k > 1$)	Kiểm định và khoảng tin cậy từng hệ số hồi quy <i>Tests and CIs for each coefficient</i>
$F(k_1, k_2)$	Chỉ nhận giá trị dương, lệch phải <i>Positive only, right-skewed</i>	$k_2/(k_2 - 2)$	Kiểm định sự phù hợp, Wald, Chow <i>Overall fit, Wald, Chow tests</i>



Hình 1.4. Mật độ chuẩn tắc và mật độ Student t: bậc tự do càng nhỏ, đuôi càng dày
Figure 1.4. Standard normal and Student t densities: fewer degrees of freedom, heavier tails

Đuôi dày hơn nên giá trị tới hạn của t lớn hơn của z: hai phía mức 5%, $z = 1,96$ nhưng $t_{0,025}^{(10)} = 2,228$. Khi bậc tự do tăng, t tiến về $N(0, 1)$; từ khoảng 30 bậc tự do trở lên hai phân phối gần như trùng nhau.

Heavier tails make the t critical value larger than z: two-sided at 5%, $z = 1.96$ but $t_{0.025}^{(10)} = 2.228$. As the degrees of freedom grow, t approaches $N(0, 1)$; from about 30 degrees of freedom the two are almost identical.

Quan hệ giữa các phân phối — Relationships between the distributions

Quan hệ Relationship	Kiểm bằng bảng tra (mức 5%) Check with tables (5% level)
$Z^2 \sim \chi^2(1)$	$1,96^2 = 3,8416 \approx \chi^2_{0,05}(1) = 3,841$ $1.96^2 = 3.8416 \approx \chi^2_{0.05}(1) = 3.841$
$T \sim t(k) \Rightarrow T^2 \sim F(1, k)$	$2,228^2 = 4,964 \approx F_{0,05}(1; 10) = 4,965$ $2.228^2 = 4.964 \approx F_{0.05}(1, 10) = 4.965$
$t(k)$ tiến về $N(0, 1)$ khi k lớn $t(k)$ tends to $N(0, 1)$ as k grows	$t_{0,025}(10) = 2,228$; $t_{0,025}(120) = 1,980$; $z = 1,96$ $t_{0.025}(10) = 2.228$; $t_{0.025}(120) = 1.980$; $z = 1.96$
$k_1 \cdot F(k_1, k_2)$ tiến về $\chi^2(k_1)$ khi k_2 lớn $k_1 \cdot F(k_1, k_2)$ tends to $\chi^2(k_1)$ as k_2 grows	$3 \times F_{0,05}(3; 1000) = 3 \times 2,614 = 7,842 \approx \chi^2_{0,05}(3) = 7,815$ $3 \times F_{0.05}(3, 1000) = 3 \times 2.614 = 7.842 \approx \chi^2_{0.05}(3) = 7.815$

Quan hệ $T^2 = F(1, k)$ giải thích vì sao trong hồi quy hai biến thống kê F của kiểm định sự phù hợp đúng bằng bình phương thống kê t của hệ số góc (Chương 3).

The relation $T^2 = F(1, k)$ explains why, in a two-variable regression, the F statistic for overall fit equals the square of the slope's t statistic (Chapter 3).

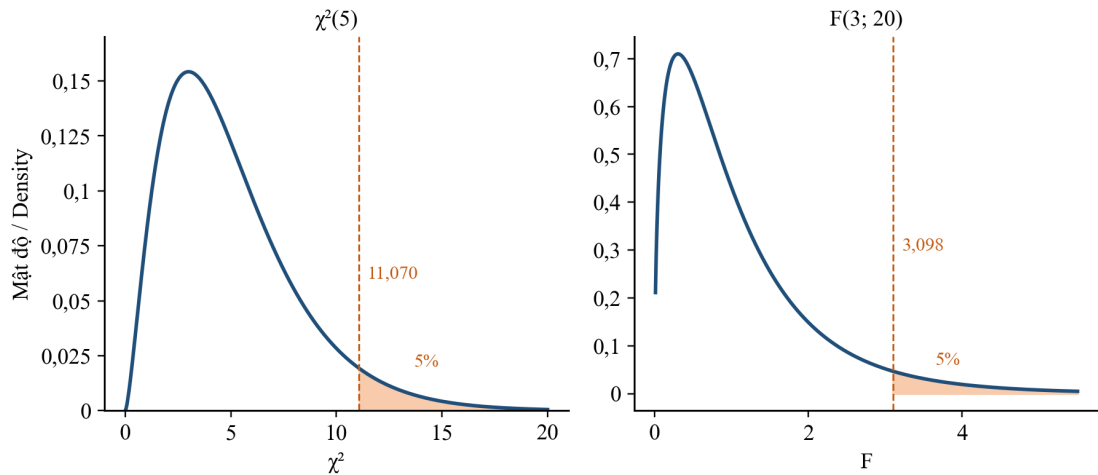
Cách tra bảng — Reading the tables

Bảng ở Phụ lục B cho **giá trị tới hạn đuôi phải**: số c sao cho xác suất phân phối vượt c bằng α . Kiểm định hai phía mức α thì tra $\alpha/2$. Đầu dòng là bậc tự do, đầu cột là α . Bảng F có hai bậc tự do: k_1 của tử số (đầu cột), k_2 của mẫu số (đầu dòng); đổi chỗ hai bậc tự do sẽ ra số khác.

*The tables in Appendix B give **right-tail critical values**: the number c such that the probability of exceeding c is α . A two-sided test at level α uses $\alpha/2$. Rows are degrees of freedom, columns are α . The F table has two degrees of freedom: k_1 for the numerator (columns) and k_2 for the denominator (rows); swapping them gives a different number.*

Cần tìm Needed	Ký hiệu Notation	Giá trị Value	Dùng khi Used for
t , hai phía 5%, $df = 15$ t , two-sided 5%, $df = 15$	$t_{0,025}(15)$ $t_{0.025}(15)$	2,131 2.131	Khoảng tin cậy 95%, $H1: \mu \neq \mu_0$ 95% CI, $H1: \mu \neq \mu_0$
t , một phía 5%, $df = 15$ t , one-sided 5%, $df = 15$	$t_{0,05}(15)$ $t_{0.05}(15)$	1,753 1.753	$H1: \mu > \mu_0$ hoặc $\mu < \mu_0$ $H1: \mu > \mu_0$ or $\mu < \mu_0$
χ^2 đuôi phải 2,5%, $df = 15$ χ^2 upper 2.5%, $df = 15$	$\chi^2_{0,025}(15)$ $\chi^2_{0.025}(15)$	27,488 27.488	Cận dưới khoảng tin cậy σ^2 Lower limit of CI for σ^2
χ^2 đuôi phải 97,5%, $df = 15$ χ^2 upper 97.5%, $df = 15$	$\chi^2_{0,975}(15)$ $\chi^2_{0.975}(15)$	6,262 6.262	Cận trên khoảng tin cậy σ^2 Upper limit of CI for σ^2
χ^2 mức 5%, $df = 5$ χ^2 at 5%, $df = 5$	$\chi^2_{0,05}(5)$ $\chi^2_{0.05}(5)$	11,070 11.070	Kiểm định White, BG (Chương 7, 8) White, BG tests (Chapters 7, 8)

Cần tìm <i>Needed</i>	Ký hiệu <i>Notation</i>	Giá trị <i>Value</i>	Dùng khi <i>Used for</i>
F mức 5%, (3; 20) <i>F at 5%, (3, 20)</i>	$F_{0,05}(3; 20)$ <i>$F_{0.05}(3, 20)$</i>	3,098 <i>3.098</i>	Kiểm định F hồi quy bội (Chương 4) <i>F-test in multiple regression (Chapter 4)</i>
z hai phía 5% <i>z two-sided 5%</i>	$Z_{0,025}$ <i>$Z_{0.025}$</i>	1,96 <i>1.96</i>	σ đã biết hoặc mẫu lớn <i>σ known or large sample</i>
z một phía 5% <i>z one-sided 5%</i>	$Z_{0,05}$ <i>$Z_{0.05}$</i>	1,645 <i>1.645</i>	Kiểm định z một phía <i>One-sided z-test</i>



Hình 1.5. Phân phối $\chi^2(5)$ và $F(3; 20)$: vùng tô là 5% đuôi phải, bên phải giá trị tới hạn
Figure 1.5. The $\chi^2(5)$ and $F(3, 20)$ distributions: the shaded area is the 5% right tail beyond the critical value

LỖI HAY GẶP / COMMON MISTAKE — Ba lỗi tra bảng hay gặp: (1) kiểm định hai phía mà tra cột α thay vì $\alpha/2$ (dùng 1,753 thay cho 2,131 khi $df = 15$); (2) dùng bậc tự do n thay vì $n - 1$ (với trung bình) hoặc $n - k$ (với hồi quy k tham số); (3) với F , đổi chỗ bậc tự do tử số và mẫu số.

Three common table errors: (1) using column α instead of $\alpha/2$ in a two-sided test (1.753 instead of 2.131 for $df = 15$); (2) using n degrees of freedom instead of $n - 1$ (for a mean) or $n - k$ (for a regression with k parameters); (3) for F , swapping numerator and denominator degrees of freedom.

1.7. Ước lượng điểm và ước lượng khoảng — Point and interval estimation

Tham số (μ , σ^2 , β_2) là con số cố định của tổng thể, không biết. **Ước lượng** (estimator) là một công thức tính từ mẫu, ví dụ $\bar{X} = \sum X_i/n$; nó là biến ngẫu nhiên vì mẫu thay đổi thì giá trị thay đổi. **Giá trị ước lượng** (estimate) là con số cụ thể khi thế một mẫu vào công thức, ví dụ $\bar{X} = 42$ phút. Ba tính chất dùng để so sánh các ước lượng:

A **parameter** (μ , σ^2 , β_2) is a fixed, unknown population number. An **estimator** is a formula computed from the sample, e.g. $\bar{X} = \sum X_i/n$; it is a random variable since its value changes from sample to sample. An **estimate** is the specific number obtained from one sample, e.g. $\bar{X} = 42$ minutes. Three properties are used to compare estimators:

Không chệch: $E(\hat{\theta}) = \theta$; Hiệu quả: $\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta})$; Vững: $\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta$
(1.18)

Tính chất <i>Property</i>	Định nghĩa <i>Definition</i>	Ví dụ <i>Example</i>
Không chệch <i>Unbiased</i>	Trung bình của θ qua mọi mẫu bằng θ <i>The mean of θ over all samples equals θ</i>	$E(\bar{X}) = \mu$; S^2 chia $n - 1$ không chệch, chia n bị chệch xuống <i>$E(\bar{X}) = \mu$; S^2 with $n - 1$ is unbiased, with n biased downward</i>
Hiệu quả <i>Efficient</i>	Không chệch và có phương sai nhỏ nhất trong một lớp ước lượng <i>Unbiased with the smallest variance in a class</i>	Gauss–Markov: OLS là BLUE (Chương 2) <i>Gauss–Markov: OLS is BLUE (Chapter 2)</i>
Vững <i>Consistent</i>	θ hội tụ theo xác suất về θ khi n tăng <i>θ converges in probability to θ as n grows</i>	$\text{Var}(\bar{X}) = \sigma^2/n$ tiến về 0 <i>$\text{Var}(\bar{X}) = \sigma^2/n$ tends to 0</i>

VÍ DỤ / EXAMPLE — So sánh hai ước lượng của μ từ mẫu n quan sát độc lập: $\bar{X} = \Sigma X_i/n$ và $\tilde{\mu} = (X_1 + X_n)/2$ (trung bình của quan sát đầu và cuối). Cả hai không chệch vì $E(\tilde{\mu}) = (\mu + \mu)/2 = \mu$. Nhưng $\text{Var}(\bar{X}) = \sigma^2/n$ còn $\text{Var}(\tilde{\mu}) = \sigma^2/2$. Với $n = 16$, $\text{Var}(\bar{X}) = \sigma^2/16$, chỉ bằng 1/8 $\text{Var}(\tilde{\mu})$: $\bar{X}$ hiệu quả hơn. Khi n tăng, $\text{Var}(\bar{X})$ tiến về 0 nên $\bar{X}$ vững; $\text{Var}(\tilde{\mu})$ luôn bằng $\sigma^2/2$ nên $\tilde{\mu}$ không vững dù không chệch.

Compare two estimators of μ from n independent observations: $\bar{X} = \Sigma X_i/n$ and $\tilde{\mu} = (X_1 + X_n)/2$ (the mean of the first and last observations). Both are unbiased since $E(\tilde{\mu}) = (\mu + \mu)/2 = \mu$. But $\text{Var}(\bar{X}) = \sigma^2/n$ while $\text{Var}(\tilde{\mu}) = \sigma^2/2$. With $n = 16$, $\text{Var}(\bar{X}) = \sigma^2/16$, only 1/8 of $\text{Var}(\tilde{\mu})$: $\bar{X}$ is more efficient. As n grows $\text{Var}(\bar{X})$ tends to 0, so $\bar{X}$ is consistent; $\text{Var}(\tilde{\mu})$ stays at $\sigma^2/2$, so $\tilde{\mu}$ is not consistent although unbiased.

Khoảng tin cậy — Confidence intervals

Khoảng tin cậy $(1 - \alpha)$ là khoảng $(a; b)$ tính từ mẫu sao cho, nếu lặp lại việc lấy mẫu nhiều lần, tỉ lệ $(1 - \alpha)$ số khoảng chứa tham số thật. Với một mẫu cụ thể, khoảng đã tính hoặc chứa hoặc không chứa μ ; “95%” là độ tin cậy của phương pháp.

A $(1 - \alpha)$ confidence interval is an interval $(a; b)$ computed from the sample such that, over repeated sampling, a share $(1 - \alpha)$ of the intervals contain the true parameter. For one given sample the interval either contains μ or not; “95%” is the reliability of the method.

$$\sigma \text{ đã biết: } \bar{X} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{X} + z_{\alpha/2} \cdot \sigma / \sqrt{n} \quad (1.19)$$

$$\sigma \text{ chưa biết: } \bar{X} - t_{\alpha/2}^{(n-1)} \cdot s / \sqrt{n} < \mu < \bar{X} + t_{\alpha/2}^{(n-1)} \cdot s / \sqrt{n} \quad (1.20)$$

$$(n - 1)s^2 / \chi^2_{\alpha/2}(n-1) < \sigma^2 < (n - 1)s^2 / \chi^2_{1-\alpha/2}(n-1) \quad (1.21)$$

VÍ DỤ / EXAMPLE — **σ đã biết.** Gói cà phê ghi khối lượng 500 g; máy đóng gói có $\sigma = 6$ g theo thông số kỹ thuật. Cân 36 gói được $\bar{X} = 498$ g (dữ kiện đề bài). $se = \sigma/\sqrt{n} = 6/\sqrt{36} = 1$; $z_{0,025} = 1,96$. Khoảng tin cậy 95%: $498 \pm 1,96 \times 1 = (496,04; 499,96)$ g. Khoảng không chứa 500: khối lượng trung bình thấp hơn mức ghi trên bao bì.

σ known. A coffee pack is labelled 500 g; the packing machine has $\sigma = 6$ g by specification. Weighing 36 packs gives $\bar{X} = 498$ g (given data). $se = \sigma/\sqrt{n} = 6/\sqrt{36} = 1$; $z_{0.025} = 1.96$. 95% confidence interval: $498 \pm 1.96 \times 1 = (496.04; 499.96)$ g. The interval excludes 500: the mean weight is below the label.

VÍ DỤ / EXAMPLE — σ chưa biết. Dịch vụ Giao Nhanh Phương Nam đo 16 đơn nội thành: $\bar{X} = 42$ phút, $s = 6$ phút (dữ kiện đề bài, giả sử thời gian giao có phân phối chuẩn). $se = 6/\sqrt{16} = 1.5$; $t_{0.025}(15) = 2.131$. Khoảng tin cậy 95% cho μ : $42 \pm 2.131 \times 1.5 = (38.8035; 45.1965)$ phút. Cho σ^2 : $(n-1)s^2 = 15 \times 36 = 540$; $\chi^2_{0.025}(15) = 27.488$, $\chi^2_{0.975}(15) = 6.262$; khoảng (19,6449; 86,2344) phút².

σ unknown. Giao Nhanh Phương Nam measures 16 inner-city deliveries: $\bar{X} = 42$ minutes, $s = 6$ minutes (given data, delivery time assumed normal). $se = 6/\sqrt{16} = 1.5$; $t_{0.025}(15) = 2.131$. 95% confidence interval for μ : $42 \pm 2.131 \times 1.5 = (38.8035; 45.1965)$ minutes. For σ^2 : $(n-1)s^2 = 15 \times 36 = 540$; $\chi^2_{0.025}(15) = 27.488$, $\chi^2_{0.975}(15) = 6.262$; interval (19.6449; 86.2344) minutes².

Độ rộng khoảng tin cậy tăng khi độ tin cậy tăng (z , t lớn hơn), khi độ phân tán s lớn hơn, và giảm khi cỡ mẫu n tăng (theo $\sqrt{n}$: muốn khoảng hẹp đi một nửa phải tăng cỡ mẫu lên gấp bốn).

The interval widens with a higher confidence level (larger z or t) and a larger spread s , and narrows as n grows (by $\sqrt{n}$: halving the width requires four times the sample size).

1.8. Kiểm định giả thuyết thống kê — Statistical hypothesis testing

Giả thuyết không H_0 là phát biểu về tham số được giữ cho tới khi dữ liệu cho bằng chứng ngược lại; luôn chứa dấu bằng. **Giả thuyết đối H_1** là điều người nghiên cứu muốn chứng minh. Kiểm định là quy tắc dựa vào mẫu để quyết định bác bỏ hay chưa đủ cơ sở bác bỏ H_0 .

*The **null hypothesis** H_0 is a statement about a parameter that stands until the data give evidence against it; it always contains the equality. The **alternative hypothesis** H_1 is what the researcher wants to show. A test is a sample-based rule for deciding to reject or not to reject H_0 .*

Loại kiểm định Test type	Cặp giả thuyết Hypotheses	Bác bỏ H_0 khi Reject H_0 when	Ví dụ câu hỏi Example question
Hai phía Two-sided	$H_0: \mu = \mu_0; H_1: \mu \neq \mu_0$	$ t > t(\alpha/2; df)$	Thời gian giao trung bình có khác 46 phút không? <i>Does mean delivery time differ from 46 minutes?</i>
Một phía phải Right-tailed	$H_0: \mu \leq \mu_0; H_1: \mu > \mu_0$	$t > t(\alpha; df)$	Có giao chậm hơn cam kết 40 phút không? <i>Is delivery slower than the 40-minute promise?</i>
Một phía trái Left-tailed	$H_0: \mu \geq \mu_0; H_1: \mu < \mu_0$	$t < -t(\alpha; df)$	Giảm giá có làm lượng bán giảm không? ($H_1: \beta_2 < 0$) <i>Does price cut sales? ($H_1: \beta_2 < 0$)</i>

Sai lầm loại I, loại II và mức ý nghĩa — *Type I and Type II errors and the significance level*

Quyết định <i>Decision</i>	H0 đúng <i>H0 true</i>	H0 sai <i>H0 false</i>
Bác bỏ H0 <i>Reject H0</i>	Sai lầm loại I , xác suất α <i>Type I error, probability α</i>	Đúng, xác suất $1 - \beta$ (lực kiểm định) <i>Correct, probability $1 - \beta$ (power)</i>
Chưa đủ cơ sở bác bỏ H0 <i>Fail to reject H0</i>	Đúng, xác suất $1 - \alpha$ <i>Correct, probability $1 - \alpha$</i>	Sai lầm loại II , xác suất β <i>Type II error, probability β</i>

Mức ý nghĩa $\alpha = P(\text{bác bỏ } H_0 \mid H_0 \text{ đúng})$ do người kiểm định chọn trước, thường 5%; 1% khi hậu quả của sai lầm loại I nặng, 10% với mẫu nhỏ. Với cỡ mẫu cố định, giảm α làm β tăng; muốn giảm cả hai phải tăng cỡ mẫu.

The significance level $\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$ is chosen in advance, usually 5%; 1% when a Type I error is costly, 10% for small samples. For a fixed sample size, lowering α raises β ; reducing both requires a larger sample.

Thủ tục kiểm định năm bước — *The five-step testing procedure*

1. Đặt H_0 , H_1 theo câu hỏi (điều cần chứng minh đặt ở H_1); xác định hai phía hay một phía.
State H_0 and H_1 from the question (the claim to be shown goes in H_1); decide two-sided or one-sided.
2. Chọn mức ý nghĩa α và thống kê kiểm định cùng phân phối của nó khi H_0 đúng (z , t , χ^2 , F).
Choose α and the test statistic with its distribution under H_0 (z , t , χ^2 , F).
3. Tính giá trị quan sát của thống kê kiểm định từ mẫu.
Compute the observed value of the test statistic from the sample.
4. Tra giá trị tới hạn (miền bác bỏ) hoặc tính p-value.
Look up the critical value (rejection region) or compute the p-value.
5. Kết luận bằng lời theo ngữ cảnh: “bác bỏ H_0 , ...” hoặc “chưa đủ cơ sở bác bỏ H_0 , ...”.

Conclude in words in context: “reject H_0 , ...” or “fail to reject H_0 , ...”.

$$t = (\bar{X} - \mu_0) / (s/\sqrt{n}) \sim t(n-1) \text{ khi } H_0 \text{ đúng} \quad (1.22)$$

$$\text{p-value hai phía} = 2 \cdot P(T > |t_{qs}|); \text{ một phía phải} = P(T > t_{qs}) \quad (1.23)$$

p-value là xác suất, khi H_0 đúng, thu được giá trị thống kê kiểm định cực đoan bằng hoặc hơn giá trị quan sát. p-value nhỏ nghĩa là dữ liệu khó xảy ra nếu H_0 đúng. Quy tắc: $\text{p-value} < \alpha$ thì bác bỏ H_0 . Phần mềm in sẵn p-value hai phía ở cột Prob.; kiểm định một phía thì chia đôi khi dấu của t đúng chiều H_1 .

*The **p-value** is the probability, if H_0 is true, of a test statistic at least as extreme as the observed value. A small p-value means the data would be unlikely under H_0 . Rule: reject H_0 when p-value*

$< \alpha$. Software prints the two-sided p -value in the Prob. column; for a one-sided test, halve it when t has the sign of H_1 .

VÍ DỤ / EXAMPLE — Kiểm định z. Tiếp ví dụ gói cà phê: $H_0: \mu = 500$; $H_1: \mu \neq 500$. $z = (498 - 500)/1 = -2,00$; $|z| = 2,00 > 1,96$ nên bác bỏ H_0 ; $p\text{-value} = 2 \cdot P(Z > 2,00) = 0,0455 < 0,05$. Kết luận: ở mức 5%, khối lượng trung bình khác mức ghi trên bao bì. Kết luận trùng với khoảng tin cậy ở mức 1.7 vì 500 nằm ngoài khoảng.

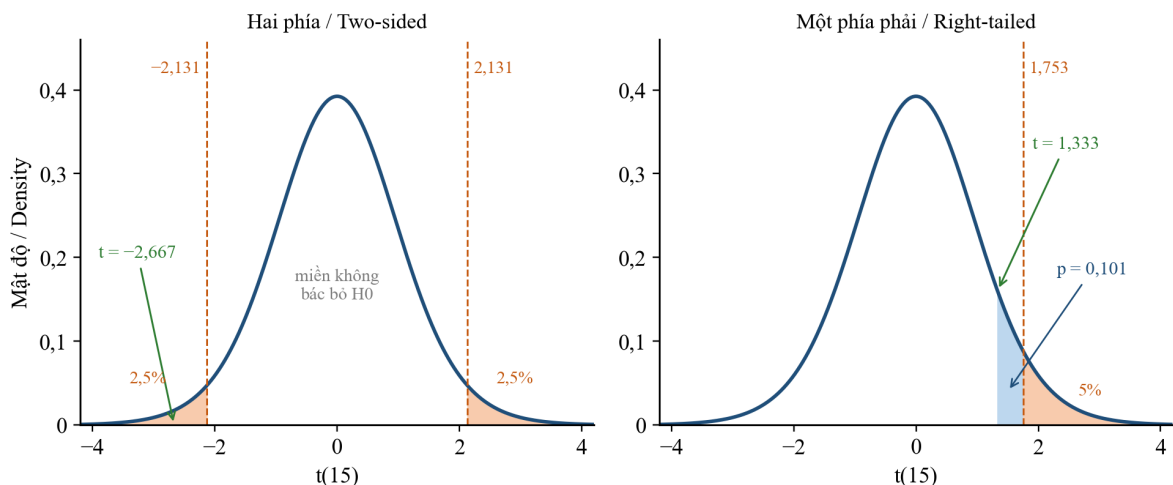
z-test. Continuing the coffee example: $H_0: \mu = 500$; $H_1: \mu \neq 500$. $z = (498 - 500)/1 = -2.00$; $|z| = 2.00 > 1.96$, so reject H_0 ; $p\text{-value} = 2 \cdot P(Z > 2.00) = 0.0455 < 0.05$. Conclusion: at 5%, the mean weight differs from the label. This matches the confidence interval in section 1.7, since 500 lies outside it.

VÍ DỤ / EXAMPLE — Kiểm định t hai phía. Báo cáo năm trước của Giao Nhanh Phương Nam ghi thời gian giao trung bình 46 phút. Năm nay có khác không? $H_0: \mu = 46$; $H_1: \mu \neq 46$. $t = (42 - 46)/1,5 = -2,667$; $|t| = 2,667 > t_{0,025}(15) = 2,131$ nên **bác bỏ H_0** ; $p\text{-value} = 0,0176$. Ở mức 5%, thời gian giao trung bình năm nay khác 46 phút (thấp hơn).

Two-sided t-test. Last year's report of Giao Nhanh Phương Nam gave a mean delivery time of 46 minutes. Is this year different? $H_0: \mu = 46$; $H_1: \mu \neq 46$. $t = (42 - 46)/1.5 = -2.667$; $|t| = 2.667 > t_{0,025}(15) = 2.131$, so **reject H_0** ; $p\text{-value} = 0.0176$. At 5%, this year's mean delivery time differs from 46 minutes (it is lower).

VÍ DỤ / EXAMPLE — Kiểm định t một phía. Công ty cam kết thời gian giao trung bình không quá 40 phút. Có bằng chứng giao chậm hơn cam kết không? $H_0: \mu \leq 40$; $H_1: \mu > 40$. $t = (42 - 40)/1,5 = 1,333 < t_{0,05}(15) = 1,753$ nên **chưa đủ cơ sở bác bỏ H_0** ; $p\text{-value} = P(T > 1,333) = 0,1012 > 0,05$. Mẫu 16 đơn chưa đủ bằng chứng để nói công ty vi phạm cam kết.

One-sided t-test. The company promises a mean delivery time of at most 40 minutes. Is there evidence of slower delivery? $H_0: \mu \leq 40$; $H_1: \mu > 40$. $t = (42 - 40)/1.5 = 1.333 < t_{0,05}(15) = 1.753$, so **we fail to reject H_0** ; $p\text{-value} = P(T > 1.333) = 0.1012 > 0.05$. A sample of 16 deliveries is not enough evidence that the company breaks its promise.



Hình 1.6. Miền bác bỏ (vùng cam) của kiểm định t với 15 bậc tự do: trái là hai phía, phải là một phía; vùng xanh nhạt là p -value

Figure 1.6. Rejection regions (orange) of a t -test with 15 degrees of freedom: two-sided on the left, one-sided on the right; the light-blue area is the p -value

Cách kết luận <i>Approach</i>	Bác bỏ H_0 (hai phía) khi <i>Reject H_0 (two-sided) when</i>	Ưu điểm <i>Advantage</i>
Miền bác bỏ <i>Rejection region</i>	$ t_{\text{quan sát}} > t(\alpha/2; df)$ $ observed\ t > t(\alpha/2; df)$	Làm được bằng bảng tra trong phòng thi <i>Works with printed tables in the exam</i>
Khoảng tin cậy <i>Confidence interval</i>	μ_0 nằm ngoài khoảng tin cậy $(1 - \alpha)$ μ_0 lies outside the $(1 - \alpha)$ interval	Thấy luôn các giá trị μ phù hợp với dữ liệu <i>Shows all values of μ consistent with the data</i>
p-value	p-value $< \alpha$	Đọc thẳng từ phần mềm, cho biết mức độ bằng chứng <i>Read directly from software, shows strength of evidence</i>

LỖI HAY GẶP / COMMON MISTAKE — Viết “chấp nhận H_0 ” hoặc “ H_0 đúng”. Không bác bỏ H_0 chỉ nghĩa là dữ liệu chưa đủ bằng chứng; ví dụ một phía ở trên có khoảng tin cậy 95% (38,80; 45,20) chứa cả 40 lẫn 45. Đề thi trừ điểm câu kết luận “chấp nhận H_0 ”.

Writing “accept H_0 ” or “ H_0 is true”. Not rejecting H_0 only means the data lack evidence; the one-sided example above has a 95% interval (38.80; 45.20) containing both 40 and 45. Exams deduct marks for the conclusion “accept H_0 ”.

LỖI HAY GẶP / COMMON MISTAKE — Hiểu p-value là “xác suất H_0 đúng” hoặc “xác suất kết luận sai”. p-value được tính khi giả sử H_0 đúng, nên không thể là xác suất của chính H_0 . Lỗi thứ hai: kiểm định một phía nhưng so p-value hai phía phần mềm in ra với α , hoặc ngược lại.

Reading the p-value as “the probability that H_0 is true” or “the probability of a wrong conclusion”. The p-value is computed assuming H_0 is true, so it cannot be the probability of H_0 itself. A second slip: running a one-sided test but comparing the software’s two-sided p-value with α , or vice versa.

1.9. Phần mềm kinh tế lượng — Econometric software

Đề thi UFM cho bảng kết quả theo khuôn EViews; khi làm tiểu luận, sinh viên có thể dùng bất kỳ phần mềm nào trong bảng dưới. Các cột hệ số (Coefficient), sai số chuẩn (Std. Error), thống kê t (t-Statistic), p-value (Prob.) và R^2 có cùng ý nghĩa ở mọi phần mềm; chỉ tên gọi và cách sắp xếp khác nhau.

UFM exams give output in EViews format; for coursework students may use any package in the table below. The coefficient, standard error, t-statistic, p-value (Prob.) and R^2 columns mean the same in every package; only labels and layout differ.

Phần mềm <i>Software</i>	Giấy phép <i>Licence</i>	Cách dùng <i>Interface</i>	Thế mạnh <i>Strength</i>
EViews	Thương mại, có bản sinh viên <i>Commercial, student edition</i>	Menu và câu lệnh <i>Menus and commands</i>	Chuỗi thời gian, bảng kết quả giống đề thi <i>Time series, output matches the exam</i>

Phần mềm <i>Software</i>	Giấy phép <i>Licence</i>	Cách dùng <i>Interface</i>	Thế mạnh <i>Strength</i>
Stata	Thương mại <i>Commercial</i>	Câu lệnh, tệp do <i>Commands, do-files</i>	Dữ liệu bảng, vi mô, sai số chuẩn vững <i>Panel and micro data, robust errors</i>
R	Miễn phí, mã nguồn mở <i>Free, open source</i>	Câu lệnh, RStudio <i>Commands, RStudio</i>	Gói lmtest, sandwich, đồ thị <i>lmtest and sandwich packages, graphics</i>
Python (statsmodels)	Miễn phí, mã nguồn mở <i>Free, open source</i>	Câu lệnh, Jupyter <i>Code, Jupyter</i>	Kết hợp xử lý dữ liệu (pandas), tải dữ liệu qua API <i>Combines with pandas, API data downloads</i>

Câu lệnh hồi quy Y theo X (có hệ số chặn), kèm kiểm định White và Breusch–Godfrey bậc 2 ở mỗi phần mềm. Tên biến viết thường; dữ liệu đã nạp sẵn.

Commands to regress Y on X (with an intercept), plus the White test and the order-2 Breusch–Godfrey test, in each package. Variable names are lower case; the data are already loaded.

```
' EViews
equation eq1.ls y c x
eq1.hettest(type=white)
eq1.auto(2)
```

```
// Stata
regress y x
estat imtest, white
estat bgodfrey, lags(2)
```

```
# R
m <- lm(y ~ x, data = d)
summary(m)
lmtest::bptest(m, ~ x + I(x^2), data = d)
lmtest::bptest(m, order = 2)
```

```
# Python statsmodels
import statsmodels.formula.api as smf
from statsmodels.stats.diagnostic import het_white, acorr_breusch_godfrey
m = smf.ols('y ~ x', data=d).fit()
print(m.summary())
print(het_white(m.resid, m.model.exog))
print(acorr_breusch_godfrey(m, nlags=2))
```

Trong EViews, chữ c là hệ số chặn và phải ghi rõ; Stata, R, Python tự thêm hệ số chặn. Stata yêu cầu khai báo biến thời gian (tsset) trước khi chạy estat bgodfrey. Bảng kết quả sách dùng (từ Chương 3) in theo khuôn EViews, số thập phân giữ dấu chấm như máy in.

In EViews, c is the intercept and must be written; Stata, R and Python add it automatically. Stata needs the time variable declared (tsset) before estat bgodfrey. The output tables in this book (from Chapter 3) follow the EViews layout, with decimal points as the software prints them.

LỖI HAY GẶP / COMMON MISTAKE — Quên chữ c trong lệnh EViews (ls y x): mô hình bị ép đi qua gốc tọa độ, R^2 và mọi hệ số thay đổi. Lỗi tương tự trong Python khi dùng `sm.OLS(y, x)` mà không thêm `sm.add_constant(x)`.

Leaving out c in the EViews command ($ls\ y\ x$): the model is forced through the origin and R^2 and every coefficient change. The same slip in Python is `sm.OLS(y, x)` without `sm.add_constant(x)`.

1.10. Tóm tắt, so sánh và thuật ngữ — Summary, comparisons and glossary

PHẢI THUỘC / MEMORISE — Mô hình kinh tế lượng $Y_i = \beta_1 + \beta_2 X_i + u_i$; u gom biến bỏ sót, sai số đo, yếu tố ngẫu nhiên, sai dạng hàm. $E(aX + b) = aE(X) + b$; $Var(aX + b) = a^2 Var(X)$; $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$. $Cov = E(XY) - E(X)E(Y)$; $\rho = Cov/(\sigma_X \cdot \sigma_Y)$. Mẫu: $S_{xy} = \sum x_i y_i / (n - 1)$; $r = \sum x_i y_i / \sqrt{(\sum x_i^2 \sum y_i^2)}$; kiểm định $\rho = 0$: $t = r\sqrt{(n - 2) / (1 - r^2)}$, $df = n - 2$. Khoảng tin cậy μ : $\bar{X} \pm z \cdot \sigma / \sqrt{n}$ (σ biết), $\bar{X} \pm t(n - 1) \cdot s / \sqrt{n}$ (σ chưa biết); σ^2 : dùng $\chi^2(n - 1)$. Kiểm định: H_0 chứa dấu bằng, hai phía tra $\alpha/2$, một phía tra α ; $p\text{-value} < \alpha$ thì bác bỏ H_0 ; kết luận “chưa đủ cơ sở bác bỏ H_0 ”, không viết “chấp nhận H_0 ”. $T^2 \sim F(1, k)$; $Z^2 \sim \chi^2(1)$.

Econometric model $Y_i = \beta_1 + \beta_2 X_i + u_i$; u collects omitted variables, measurement error, randomness and functional-form error. $E(aX + b) = aE(X) + b$; $Var(aX + b) = a^2 Var(X)$; $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$. $Cov = E(XY) - E(X)E(Y)$; $\rho = Cov/(\sigma_X \cdot \sigma_Y)$. Sample: $S_{xy} = \sum x_i y_i / (n - 1)$; $r = \sum x_i y_i / \sqrt{(\sum x_i^2 \sum y_i^2)}$; test of $\rho = 0$: $t = r\sqrt{(n - 2) / (1 - r^2)}$, $df = n - 2$. CI for μ : $\bar{X} \pm z \cdot \sigma / \sqrt{n}$ (σ known), $\bar{X} \pm t(n - 1) \cdot s / \sqrt{n}$ (σ unknown); for σ^2 use $\chi^2(n - 1)$. Testing: H_0 holds the equality, two-sided uses $\alpha/2$, one-sided uses α ; reject H_0 when $p\text{-value} < \alpha$; write “fail to reject H_0 ”, never “accept H_0 ”. $T^2 \sim F(1, k)$; $Z^2 \sim \chi^2(1)$.

Các cặp khái niệm dễ nhầm — Easily confused pairs

Khái niệm 1 <i>Concept 1</i>	Khái niệm 2 <i>Concept 2</i>	Phân biệt <i>Distinction</i>
Tổng thể <i>Population</i>	Mẫu <i>Sample</i>	Tham số (μ, β) thuộc tổng thể; thống kê ($\bar{X}, \beta$) tính từ mẫu <i>Parameters (μ, β) belong to the population; statistics ($\bar{X}, \beta$) come from the sample</i>
Ước lượng (estimator) <i>Estimator</i>	Giá trị ước lượng (estimate) <i>Estimate</i>	Công thức, là biến ngẫu nhiên; con số cụ thể từ một mẫu <i>A formula and a random variable; a number from one sample</i>
Hiệp phương sai <i>Covariance</i>	Hệ số tương quan <i>Correlation</i>	Cov có đơn vị, không giới hạn; ρ không đơn vị, trong $[-1; 1]$ <i>Cov has units and no bounds; ρ is unit-free, in $[-1; 1]$</i>
Tương quan <i>Correlation</i>	Nhân quả <i>Causation</i>	Cùng biến động; tác động có chiều, cần lý thuyết <i>Co-movement; directional effect needing theory</i>
Không chệch <i>Unbiased</i>	Vững <i>Consistent</i>	Đúng trung bình với mọi n ; tiến về θ khi n lớn <i>Right on average for every n; tends to θ for large n</i>
Sai lầm loại I <i>Type I error</i>	Sai lầm loại II <i>Type II error</i>	Bác bỏ H_0 đúng (α); không bác bỏ H_0 sai (β) <i>Rejecting a true H_0 (α); not rejecting a false H_0 (β)</i>

Khái niệm 1 <i>Concept 1</i>	Khái niệm 2 <i>Concept 2</i>	Phân biệt <i>Distinction</i>
Kiểm định hai phía <i>Two-sided test</i>	Kiểm định một phía <i>One-sided test</i>	Tra $\alpha/2$, $H_1: \neq$; tra α , $H_1: >$ hoặc $<$ <i>Use $\alpha/2$, $H_1: \neq$; use α, $H_1: >$ or $<$</i>
z (σ đã biết) <i>z (σ known)</i>	t (σ chưa biết) <i>t (σ unknown)</i>	Dùng σ của tổng thể; dùng s của mẫu, bậc tự do $n - 1$ <i>Uses population σ; uses sample s, $n - 1$ df</i>
Mức ý nghĩa α <i>Significance level α</i>	p-value	Chọn trước khi kiểm định; tính từ mẫu <i>Chosen before testing; computed from the sample</i>

Thuật ngữ Việt – Anh — Vietnamese–English glossary

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Kinh tế lượng	Econometrics	Đo lường quan hệ kinh tế bằng số liệu <i>Measuring economic relationships with data</i>
Biến phụ thuộc	Dependent variable	Biến được giải thích, Y <i>The explained variable, Y</i>
Biến độc lập	Explanatory variable	Biến giải thích, X <i>The regressor, X</i>
Sai số ngẫu nhiên	Disturbance (error term)	Phần Y không giải thích được bởi X , u <i>Part of Y not explained by X, u</i>
Dữ liệu chéo	Cross-sectional data	Nhiều đơn vị, một thời điểm <i>Many units, one date</i>
Chuỗi thời gian	Time-series data	Một đơn vị, nhiều thời kỳ <i>One unit, many periods</i>
Dữ liệu bảng	Panel data	Cùng đơn vị theo dõi nhiều kỳ <i>Same units over several periods</i>
Biến ngẫu nhiên	Random variable	Giá trị phụ thuộc kết quả phép thử <i>Value depends on chance</i>
Kỳ vọng	Expected value	Trung bình theo xác suất <i>Probability-weighted mean</i>
Phương sai	Variance	Trung bình bình phương độ lệch <i>Mean squared deviation</i>
Hiệp phương sai	Covariance	Mức cùng biến động của hai biến <i>Joint variability of two variables</i>
Hệ số tương quan	Correlation coefficient	Hiệp phương sai chuẩn hóa, trong $[-1; 1]$ <i>Standardised covariance in $[-1; 1]$</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Tương quan giả	Spurious correlation	Tương quan không có cơ sở nhân quả <i>Correlation without a causal basis</i>
Kỳ vọng có điều kiện	Conditional expectation	Trung bình của Y khi biết X <i>Mean of Y given X</i>
Bậc tự do	Degrees of freedom	Số thông tin độc lập còn lại <i>Number of independent pieces left</i>
Ước lượng không chệch	Unbiased estimator	$E(\hat{\theta}) = \theta$
Ước lượng hiệu quả	Efficient estimator	Không chệch, phương sai nhỏ nhất <i>Unbiased, minimum variance</i>
Ước lượng vững	Consistent estimator	Hội tụ về θ khi n tăng <i>Converges to θ as n grows</i>
Khoảng tin cậy	Confidence interval	Khoảng chứa tham số với độ tin cậy cho trước <i>Interval covering the parameter at a set confidence</i>
Giả thuyết không	Null hypothesis	H_0 , chứa dấu bằng <i>H_0, contains the equality</i>
Giả thuyết đối	Alternative hypothesis	H_1 , điều cần chứng minh <i>H_1, the claim to be shown</i>
Mức ý nghĩa	Significance level	Xác suất sai lầm loại I, α <i>Probability of a Type I error, α</i>
Giá trị tới hạn	Critical value	Ranh giới miền bác bỏ <i>Boundary of the rejection region</i>
Giá trị p	p-value	Xác suất kết quả cực đoan như quan sát khi H_0 đúng <i>Probability of a result as extreme, given H_0</i>

1.11. Bài tập vận dụng có lời giải — *Applied exercises with solutions*

Bài 1. Hiệp phương sai, hệ số tương quan và kiểm định ý nghĩa của r — *Exercise 1. Covariance, correlation and testing the significance of r*

Chuỗi nhà thuốc Thiện Tâm có 8 cửa hàng. Dữ kiện đề bài: X là số nhân viên đứng quầy, Y là doanh thu bình quân (triệu đồng/ngày). X = 2; 3; 4; 5; 5; 6; 7; 8. Y = 20; 26; 27; 30; 35; 33; 40; 45. (a) Tính $\bar{X}$, $\bar{Y}$, $\sum x_i y_i$, $\sum x_i^2$, $\sum y_i^2$ ($x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$). (b) Tính hiệp phương sai mẫu, độ lệch chuẩn mẫu của X và Y. (c) Tính hệ số tương quan r và nhận xét. (d) Với mức ý nghĩa 5%, kiểm định giả thuyết hệ số tương quan tổng thể bằng 0. (e) Có thể kết luận tăng thêm nhân viên làm tăng doanh thu không?

The Thiện Tâm pharmacy chain has 8 shops. Given data: X is the number of counter staff, Y is average revenue (VND million per day). $X = 2; 3; 4; 5; 5; 6; 7; 8$. $Y = 20; 26; 27; 30; 35; 33; 40; 45$. (a) Compute $\bar{X}$, $\bar{Y}$, $\Sigma x_i y_i$, Σx_i^2 , Σy_i^2 ($x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$). (b) Compute the sample covariance and the sample standard deviations of X and Y . (c) Compute the correlation coefficient r and comment. (d) At the 5% significance level, test whether the population correlation is zero. (e) Can we conclude that adding staff raises revenue?

1. $\bar{X} = 40/8 = 5$; $\bar{Y} = 256/8 = 32$. Độ lệch $x_i = -3; -2; -1; 0; 0; 1; 2; 3$; $y_i = -12; -6; -5; -2; 3; 1; 8; 13$.
 $\bar{X} = 40/8 = 5$; $\bar{Y} = 256/8 = 32$. Deviations $x_i = -3; -2; -1; 0; 0; 1; 2; 3$; $y_i = -12; -6; -5; -2; 3; 1; 8; 13$.
2. $\Sigma x_i y_i = 109$; $\Sigma x_i^2 = 28$; $\Sigma y_i^2 = 452$. Kiểm: $\Sigma x_i = 0$ và $\Sigma y_i = 0$.
 $\Sigma x_i y_i = 109$; $\Sigma x_i^2 = 28$; $\Sigma y_i^2 = 452$. Check: $\Sigma x_i = 0$ and $\Sigma y_i = 0$.
3. $S_{xy} = \Sigma x_i y_i / (n - 1) = 109/7 = 15,5714$; $S_x = \sqrt{(28/7)} = 2,0000$ người; $S_y = \sqrt{(452/7)} = 8,0356$ triệu đồng. Hiệp phương sai dương: cửa hàng đông nhân viên hơn mức trung bình thường có doanh thu trên mức trung bình.
 $S_{xy} = \Sigma x_i y_i / (n - 1) = 109/7 = 15,5714$; $S_x = \sqrt{(28/7)} = 2,0000$ staff; $S_y = \sqrt{(452/7)} = \text{VND } 8,0356 \text{ million}$. The covariance is positive: shops with above-average staff tend to have above-average revenue.
4. $r = \Sigma x_i y_i / \sqrt{(\Sigma x_i^2 \cdot \Sigma y_i^2)} = 109 / \sqrt{(28 \times 452)} = \mathbf{0,9689}$. Kiểm bằng cách khác: $r = S_{xy} / (S_x \cdot S_y) = 15,5714 / (2,0000 \times 8,0356) = 0,9689$. Quan hệ tuyến tính cùng chiều, chặt.
 $r = \Sigma x_i y_i / \sqrt{(\Sigma x_i^2 \cdot \Sigma y_i^2)} = 109 / \sqrt{(28 \times 452)} = \mathbf{0,9689}$. Check another way: $r = S_{xy} / (S_x \cdot S_y) = 15,5714 / (2,0000 \times 8,0356) = 0,9689$. A strong positive linear relationship.
5. $H_0: \rho = 0$; $H_1: \rho \neq 0$. $t = r\sqrt{(n - 2) / (1 - r^2)} = 0,9689 \times \sqrt{6 / (1 - 0,9388)} = 9,591$. Tra bảng $t_{0,025}^{(6)} = 2,447$. $|t| = 9,591 > 2,447$ nên **bác bỏ H_0** : hệ số tương quan tổng thể khác 0 ở mức 5% ($p\text{-value} = 0,0001$).
 $H_0: \rho = 0$; $H_1: \rho \neq 0$. $t = r\sqrt{(n - 2) / (1 - r^2)} = 0,9689 \times \sqrt{6 / (1 - 0,9388)} = 9,591$. From the table $t_{0,025}^{(6)} = 2,447$. $|t| = 9,591 > 2,447$, so **reject H_0** : the population correlation differs from zero at 5% ($p\text{-value} = 0,0001$).
6. Không kết luận được quan hệ nhân quả. Cửa hàng ở vị trí đông khách vừa có doanh thu cao vừa được bố trí nhiều nhân viên (nhân quả ngược, biến bị bỏ sót: vị trí, diện tích). Muốn đo tác động của nhân viên phải đưa thêm các biến đó vào mô hình hồi quy bội (Chương 4).
No causal conclusion can be drawn. Shops in busy locations both earn more and are given more staff (reverse causality and omitted variables: location, floor area). Measuring the effect of staff requires adding those variables to a multiple regression (Chapter 4).

Bài 2. Phân phối đồng thời, hiệp phương sai và kỳ vọng có điều kiện — Exercise 2. Joint distribution, covariance and conditional expectation

Một chuỗi cửa hàng mỹ phẩm theo dõi khách hàng thành viên. X là số lần khách quay lại trong tháng (0, 1, 2); $Y = 1$ nếu khách mua kèm sản phẩm chăm sóc da, $Y = 0$ nếu không. Dữ kiện đề bài, xác suất đồng thời: $P(X = 0, Y = 0) = 0,25$; $P(1, 0) = 0,15$; $P(2, 0) = 0,10$; $P(0, 1) = 0,05$; $P(1, 1) = 0,20$; $P(2, 1) = 0,25$. (a) Tìm phân phối biên của X và Y . (b) Tính

$E(X)$, $E(Y)$, $\text{Var}(X)$, $\text{Var}(Y)$. (c) Tính $\text{Cov}(X, Y)$ và ρ . (d) X và Y có độc lập không? (e) Tính $E(Y | X = x)$ với $x = 0, 1, 2$ và nêu ý nghĩa.

A cosmetics chain tracks its member customers. X is the number of return visits in a month (0, 1, 2); $Y = 1$ if the customer also buys a skincare product, $Y = 0$ otherwise. Given joint probabilities: $P(X = 0, Y = 0) = 0.25$; $P(1, 0) = 0.15$; $P(2, 0) = 0.10$; $P(0, 1) = 0.05$; $P(1, 1) = 0.20$; $P(2, 1) = 0.25$. (a) Find the marginal distributions of X and Y . (b) Compute $E(X)$, $E(Y)$, $\text{Var}(X)$, $\text{Var}(Y)$. (c) Compute $\text{Cov}(X, Y)$ and ρ . (d) Are X and Y independent? (e) Compute $E(Y | X = x)$ for $x = 0, 1, 2$ and interpret.

- Cộng theo cột: $P(X = 0) = 0,30$; $P(X = 1) = 0,35$; $P(X = 2) = 0,35$. Cộng theo dòng: $P(Y = 0) = 0,50$; $P(Y = 1) = 0,50$. Tổng mỗi phân phối bằng 1.
Sum down the columns: $P(X = 0) = 0.30$; $P(X = 1) = 0.35$; $P(X = 2) = 0.35$. Sum across the rows: $P(Y = 0) = 0.50$; $P(Y = 1) = 0.50$. Each distribution sums to 1.
- $E(X) = 0 \times 0,30 + 1 \times 0,35 + 2 \times 0,35 = 1,05$; $E(X^2) = 1,75$; $\text{Var}(X) = 1,75 - 1,05^2 = 0,6475$. $E(Y) = 0,50$; $\text{Var}(Y) = 0,50 - 0,50^2 = 0,25$.
 $E(X) = 0 \times 0.30 + 1 \times 0.35 + 2 \times 0.35 = 1.05$; $E(X^2) = 1.75$; $\text{Var}(X) = 1.75 - 1.05^2 = 0.6475$. $E(Y) = 0.50$; $\text{Var}(Y) = 0.50 - 0.50^2 = 0.25$.
- $E(XY)$ chỉ gồm các ô $Y = 1$: $1 \times 0,20 + 2 \times 0,25 = 0,70$. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0,70 - 1,05 \times 0,50 = 0,1750$. $\rho = 0,1750 / \sqrt{(0,6475 \times 0,25)} = \mathbf{0,4350}$.
 $E(XY)$ involves only the $Y = 1$ cells: $1 \times 0.20 + 2 \times 0.25 = 0.70$. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0.70 - 1.05 \times 0.50 = 0.1750$. $\rho = 0.1750 / \sqrt{(0.6475 \times 0.25)} = \mathbf{0.4350}$.
- Độc lập đòi hỏi $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ với MỌI ô. Ô (0, 0): $0,25 \neq 0,30 \times 0,50 = 0,15$. Vậy X, Y không độc lập ($\text{Cov} \neq 0$ cũng đủ để kết luận).
Independence requires $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for EVERY cell. Cell (0, 0): $0.25 \neq 0.30 \times 0.50 = 0.15$. So X and Y are not independent ($\text{Cov} \neq 0$ is enough to conclude).
- $E(Y | X = x) = P(Y = 1, X = x) / P(X = x)$: $x = 0$: $0,05 / 0,30 = 0,1667$; $x = 1$: $0,20 / 0,35 = 0,5714$; $x = 2$: $0,25 / 0,35 = 0,7143$. Khách quay lại càng nhiều thì xác suất mua kèm càng cao. Dãy giá trị $E(Y | X)$ theo X chính là hàm hồi quy tổng thể của Chương 2.
 $E(Y | X = x) = P(Y = 1, X = x) / P(X = x)$: $x = 0$: $0.05 / 0.30 = 0.1667$; $x = 1$: $0.20 / 0.35 = 0.5714$; $x = 2$: $0.25 / 0.35 = 0.7143$. The more often a customer returns, the higher the probability of an add-on purchase. The values of $E(Y | X)$ as a function of X are the population regression function of Chapter 2.

Bài 3. Khoảng tin cậy và kiểm định về trung bình, phương sai — Exercise 3.

Confidence intervals and tests for a mean and a variance

Ngân hàng TMCP Đông Hải cam kết duyệt hồ sơ vay tiêu dùng trong 5 ngày làm việc. Kiểm tra ngẫu nhiên 25 hồ sơ: thời gian duyệt trung bình $\bar{X} = 5,6$ ngày, độ lệch chuẩn mẫu $s = 1,5$ ngày (dữ kiện đề bài, giả sử thời gian duyệt có phân phối chuẩn). (a) Ước lượng khoảng tin cậy 95% cho thời gian duyệt trung bình μ . (b) Với $\alpha = 5\%$, kiểm định $H_0: \mu = 5$ với $H_1: \mu \neq 5$. (c) Với $\alpha = 5\%$, có thể kết luận ngân hàng duyệt chậm hơn cam kết không? (d) Tính p-value của hai kiểm định. (e) Ước lượng khoảng tin cậy 95% cho phương sai σ^2 .

Đông Hải Commercial Bank promises to approve consumer-loan files within 5 working days. A random check of 25 files gives a mean approval time $\bar{X} = 5.6$ days and a sample standard

deviation $s = 1.5$ days (given data; approval time is assumed normal). (a) Construct a 95% confidence interval for the mean time μ . (b) At $\alpha = 5\%$, test $H_0: \mu = 5$ against $H_1: \mu \neq 5$. (c) At $\alpha = 5\%$, can we conclude that the bank is slower than promised? (d) Find the p-values of both tests. (e) Construct a 95% confidence interval for the variance σ^2 .

1. σ chưa biết, $n = 25$: dùng phân phối t với $n - 1 = 24$ bậc tự do. $se(\bar{X}) = s/\sqrt{n} = 1,5/\sqrt{25} = 0,30$. Tra $t_{0,025}(24) = 2,064$.
 σ is unknown, $n = 25$: use the t distribution with $n - 1 = 24$ degrees of freedom. $se(\bar{X}) = s/\sqrt{n} = 1.5/\sqrt{25} = 0.30$. Table value $t_{0.025}(24) = 2.064$.
2. Khoảng tin cậy 95%: $5,6 \pm 2,064 \times 0,30 = 5,6 \pm 0,6192$, tức **(4,9808; 6,2192) ngày**.
*95% confidence interval: $5.6 \pm 2.064 \times 0.30 = 5.6 \pm 0.6192$, i.e. **(4.9808; 6.2192) days**.*
3. (b) $t = (\bar{X} - \mu_0)/se = (5,6 - 5)/0,30 = 2,000$. Hai phía: $|t| = 2,000 < 2,064$ nên **chưa đủ cơ sở bác bỏ H_0** . Khớp với (a): 5 nằm trong khoảng tin cậy.
*(b) $t = (\bar{X} - \mu_0)/se = (5.6 - 5)/0.30 = 2.000$. Two-sided: $|t| = 2.000 < 2.064$, so **we fail to reject H_0** . Consistent with (a): 5 lies inside the interval.*
4. (c) $H_0: \mu \leq 5$; $H_1: \mu > 5$ (điều cần chứng minh đặt ở H_1). Tra $t_{0,05}(24) = 1,711$. $t = 2,000 > 1,711$ nên **bác bỏ H_0** : ở mức 5%, ngân hàng duyệt chậm hơn cam kết.
*(c) $H_0: \mu \leq 5$; $H_1: \mu > 5$ (the claim to be shown goes in H_1). Table value $t_{0.05}(24) = 1.711$. $t = 2.000 > 1.711$, so **reject H_0** : at 5%, the bank is slower than promised.*
5. (d) p-value một phía = $P(T > 2,000) = 0,0285 < 0,05$; p-value hai phía = $2 \times 0,0285 = 0,0569 > 0,05$. Cùng một mẫu, kiểm định một phía bác bỏ còn hai phía thì không phải chọn H_1 theo câu hỏi TRƯỚC khi nhìn số liệu.
(d) One-sided p-value = $P(T > 2.000) = 0.0285 < 0.05$; two-sided p-value = $2 \times 0.0285 = 0.0569 > 0.05$. With the same sample the one-sided test rejects and the two-sided test does not: H_1 must be chosen from the question BEFORE looking at the data.
6. (e) $(n - 1)s^2 = 24 \times 2,25 = 54$. Tra $\chi^2_{0,025}(24) = 39,364$; $\chi^2_{0,975}(24) = 12,401$. Khoảng tin cậy 95% cho σ^2 : $(54/39,364; 54/12,401) = \mathbf{(1,3718; 4,3545)}$ ngày². Khoảng không đối xứng quanh $s^2 = 2,25$ vì phân phối χ^2 lệch phải.
(e) $(n - 1)s^2 = 24 \times 2.25 = 54$. Table values $\chi^2_{0.025}(24) = 39.364$; $\chi^2_{0.975}(24) = 12.401$. 95% confidence interval for σ^2 : $(54/39.364; 54/12.401) = \mathbf{(1.3718; 4.3545)}$ days². The interval is not symmetric about $s^2 = 2.25$ because the χ^2 distribution is right-skewed.

1.12. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines

Câu 1. Trình bày các bước của một nghiên cứu kinh tế lượng. Minh họa bằng câu hỏi: giá bán có ảnh hưởng tới lượng bán sữa chua uống của một doanh nghiệp không?

Question 1. Describe the steps of an econometric study. Illustrate with the question: does the selling price affect a firm's sales of drinking yoghurt?

- Nêu giả thuyết từ lý thuyết kinh tế: theo luật cầu, giá tăng thì lượng mua giảm, các yếu tố khác không đổi.

State the hypothesis from economic theory: by the law of demand, quantity bought falls when price rises, other things equal.

- Lập mô hình toán $Y = \beta_1 + \beta_2 X$, rồi mô hình kinh tế lượng $Y = \beta_1 + \beta_2 X + u$; kỳ vọng dấu $\beta_2 < 0$; u gom các yếu tố không đưa vào mô hình.
Write the mathematical model $Y = \beta_1 + \beta_2 X$, then the econometric model $Y = \beta_1 + \beta_2 X + u$; expected sign $\beta_2 < 0$; u collects the factors left out.
- Thu thập dữ liệu: giá và lượng bán theo tháng (chuỗi thời gian) hoặc theo cửa hàng (dữ liệu chéo); ghi đơn vị đo, nguồn.
Collect data: price and sales by month (time series) or by shop (cross-section); record units and sources.
- Ước lượng bằng OLS được β_1, β_2 ; kiểm định $H_0: \beta_2 = 0$ với $H_1: \beta_2 < 0$; kiểm tra các giả thiết của mô hình (phương sai sai số thay đổi, tự tương quan, dạng hàm).
Estimate by OLS to get β_1, β_2 ; test $H_0: \beta_2 = 0$ against $H_1: \beta_2 < 0$; check the model assumptions (heteroscedasticity, autocorrelation, functional form).
- Dự báo lượng bán ở mức giá dự kiến; đề xuất chính sách giá. Nếu mô hình không đạt ở bước kiểm định, quay lại bước lập mô hình (thêm biến, đổi dạng hàm).
Forecast sales at a planned price; recommend a pricing policy. If the model fails the checks, return to the specification step (add variables, change the functional form).
- Kết quả tham khảo trên bộ A (Chương 2): $\beta_2 = -2,1127$, tức giá tăng 1 nghìn đồng/lốc thì lượng bán giảm khoảng 2,11 nghìn lốc/tháng.
Reference result on data set A (Chapter 2): $\beta_2 = -2.1127$, i.e. a price rise of VND 1 thousand per pack lowers sales by about 2.11 thousand packs a month.

Câu 2. Phân biệt tương quan và nhân quả. Nêu các nguyên nhân khiến hai biến có tương quan mạnh mà không có quan hệ nhân quả, và cách kinh tế lượng xử lý từng trường hợp.

Question 2. *Distinguish correlation from causation. Give the reasons two variables can be strongly correlated without a causal link, and how econometrics deals with each case.*

- Tương quan đo mức độ cùng biến động tuyến tính giữa hai biến, đối xứng ($r(X, Y) = r(Y, X)$); nhân quả là thay đổi X làm Y thay đổi, có chiều.
Correlation measures how far two variables move together linearly and is symmetric ($r(X, Y) = r(Y, X)$); causation means changing X changes Y , and has a direction.
- Biến thứ ba (biến bị bỏ sót) tác động lên cả X và Y . Xử lý: đưa biến đó vào hồi quy bội; nếu bỏ sót sẽ có thiên lệch do bỏ sót biến (Chương 9).
A third (omitted) variable drives both X and Y . Remedy: include it in a multiple regression; leaving it out causes omitted variable bias (Chapter 9).
- Nhân quả ngược: Y tác động lại X (doanh nghiệp đặt ngân sách quảng cáo theo tỉ lệ doanh thu). Xử lý: dùng biến trễ, biến công cụ, thiết kế thực nghiệm.
Reverse causality: Y affects X (firms set advertising budgets as a share of sales). Remedy: lags, instrumental variables, experimental designs.
- Hai chuỗi thời gian cùng có xu thế (tương quan giả). Xử lý: lấy sai phân, đưa biến xu thế, kiểm tra tính dừng.
Two trending time series (spurious correlation). Remedy: take differences, add a trend variable, test for stationarity.

- Ngẫu nhiên do mẫu nhỏ. Xử lý: kiểm định ý nghĩa của r hoặc của hệ số hồi quy, tăng cỡ mẫu.
Chance in a small sample. Remedy: test the significance of r or of the regression coefficient, enlarge the sample.

Câu 3. Thế nào là sai lầm loại I, sai lầm loại II, mức ý nghĩa và p-value? Vì sao khi không bác bỏ H_0 phải kết luận “chưa đủ cơ sở bác bỏ H_0 ” mà không viết “chấp nhận H_0 ”?

Question 3. What are a Type I error, a Type II error, the significance level and the p-value? Why must a non-rejection be stated as “we fail to reject H_0 ” rather than “we accept H_0 ”?

- Sai lầm loại I: bác bỏ H_0 khi H_0 đúng; xác suất của nó là mức ý nghĩa α , do người kiểm định chọn (1%, 5%, 10%).
Type I error: rejecting H_0 when it is true; its probability is the significance level α , chosen by the tester (1%, 5%, 10%).
- Sai lầm loại II: không bác bỏ H_0 khi H_0 sai; xác suất β ; $1 - \beta$ là lực kiểm định. Với cỡ mẫu cố định, giảm α thì β tăng.
Type II error: failing to reject H_0 when it is false; probability β ; $1 - \beta$ is the power. For a fixed sample size, lowering α raises β .
- p-value: xác suất (khi H_0 đúng) thu được giá trị thống kê kiểm định ít nhất là cực đoan như giá trị quan sát. Quy tắc: p-value $< \alpha$ thì bác bỏ H_0 .
p-value: the probability, if H_0 is true, of a test statistic at least as extreme as the one observed. Rule: reject H_0 when p-value $< \alpha$.
- Không bác bỏ H_0 chỉ có nghĩa dữ liệu chưa đủ bằng chứng chống H_0 ; xác suất mắc sai lầm loại II β thường không biết và có thể lớn, nên không khẳng định H_0 đúng.
Not rejecting H_0 only means the data lack evidence against it; the Type II error probability β is usually unknown and may be large, so H_0 cannot be declared true.
- Ví dụ: bài 3, kiểm định hai phía có p-value = 0,0569 $>$ 0,05 nhưng khoảng tin cậy (4,98; 6,22) chứa cả 5 lẫn 6: dữ liệu không phân biệt được hai giá trị đó.
Example: in exercise 3 the two-sided test has p-value = 0.0569 $>$ 0.05, yet the confidence interval (4.98; 6.22) contains both 5 and 6: the data cannot tell them apart.

1.13. Câu hỏi trắc nghiệm — Multiple-choice questions

1. Doanh thu tháng 9/2026 của 36 cửa hàng cà phê thuộc cùng một chuỗi là loại dữ liệu nào?
September 2026 revenue of 36 coffee shops in one chain is which type of data?
 - A. Dữ liệu chuỗi thời gian / Time-series data
 - B. Dữ liệu chéo / Cross-sectional data
 - C. Dữ liệu bảng / Panel data
 - D. Dữ liệu định danh / Nominal data

2. Biến ngẫu nhiên X có $E(X) = 5$, $Var(X) = 4$. $Var(3X - 2)$ bằng bao nhiêu?
A random variable X has $E(X) = 5$ and $Var(X) = 4$. What is $Var(3X - 2)$?
 - A. 10
 - B. 12
 - C. 34
 - D. 36

3. $Cov(X, Y) = -6$, $Var(X) = 9$, $Var(Y) = 16$. Hệ số tương quan ρ bằng:
 $Cov(X, Y) = -6$, $Var(X) = 9$, $Var(Y) = 16$. The correlation coefficient ρ is:
 - A. -0,5 / -0.5
 - B. -0,0417 / -0.0417
 - C. 0,5 / 0.5
 - D. -2

4. Kiểm định t hai phía, mức ý nghĩa 5%, bậc tự do 10. Giá trị tới hạn bằng:
A two-sided t-test at the 5% level with 10 degrees of freedom. The critical value is:
 - A. 1,812 / 1.812
 - B. 1,96 / 1.96
 - C. 2,228 / 2.228
 - D. 2,764 / 2.764

5. Sai lầm loại I là:
A Type I error is:
 - A. Không bác bỏ H_0 khi H_0 sai / Failing to reject H_0 when H_0 is false
 - B. Bác bỏ H_1 khi H_1 đúng / Rejecting H_1 when H_1 is true
 - C. Bác bỏ H_0 khi H_0 đúng / Rejecting H_0 when H_0 is true
 - D. Không bác bỏ H_0 khi H_0 đúng / Failing to reject H_0 when H_0 is true

6. Kiểm định $H_0: \beta_2 = 0$ cho p-value = 0,032. Kết luận đúng là:
A test of $H_0: \beta_2 = 0$ gives p-value = 0.032. The correct conclusion is:
 - A. Bác bỏ H_0 ở mức 5%, chưa đủ cơ sở bác bỏ H_0 ở mức 1% / Reject H_0 at 5% but not at 1%
 - B. Chấp nhận H_0 ở mức 5% / Accept H_0 at 5%
 - C. Bác bỏ H_0 ở cả mức 1% và 5% / Reject H_0 at both 1% and 5%
 - D. Xác suất H_0 đúng là 3,2% / The probability that H_0 is true is 3.2%

7. Nếu T có phân phối Student với k bậc tự do thì T^2 có phân phối:
If T follows Student's t distribution with k degrees of freedom, then T^2 follows:
 - A. $\chi^2(k)$
 - B. $N(0, 1)$
 - C. $F(k, 1)$
 - D. $F(1, k)$

8. Ước lượng θ của tham số θ được gọi là vững nếu:

An estimator θ of θ is called consistent if:

- A.** $E(\theta) = \theta$ với mọi cỡ mẫu / *$E(\theta) = \theta$ for every sample size*
- B.** θ hội tụ theo xác suất về θ khi n tăng vô hạn / *θ converges in probability to θ as n grows without bound*
- C.** θ có phương sai nhỏ nhất trong các ước lượng tuyến tính / *θ has the smallest variance among linear estimators*
- D.** θ có phân phối chuẩn / *θ is normally distributed*

Chương 2. Hồi quy hai biến: ước lượng bằng phương pháp bình phương nhỏ nhất — Chapter 2. Two-variable regression: estimation by ordinary least squares

Mục tiêu chương — Chapter objectives

- Phân biệt hàm hồi quy tổng thể với hàm hồi quy mẫu, sai số ngẫu nhiên với phần dư; nhận biết mô hình tuyến tính theo tham số.
Distinguish the population from the sample regression function and the disturbance from the residual; recognise models that are linear in parameters.
- Tính tay β_1 , β_2 bằng cả dạng độ lệch và dạng trung bình, trình bày bảng tính như bài thi.
Compute β_1 and β_2 by hand in both the deviation form and the means form, laying out the worksheet as in an exam.
- Diễn giải hệ số theo đúng đơn vị; tính TSS, ESS, RSS, R^2 , r và σ^2 .
Interpret coefficients in the correct units; compute TSS, ESS, RSS, R^2 , r and σ^2 .
- Nêu các giả thiết của mô hình hồi quy tuyến tính cổ điển, định lý Gauss–Markov và tính sai số chuẩn của β_1 , β_2 .
State the assumptions of the classical linear regression model and the Gauss–Markov theorem, and compute the standard errors of β_1 and β_2 .
- Dùng các tính chất của đường hồi quy mẫu để kiểm tra bảng tính và tìm ô bị khuyết.
Use the properties of the sample regression line to check a worksheet and fill missing cells.

PHẢI THUỘC / MEMORISE — Dữ liệu dùng trong chương (bộ A, dùng lại ở Chương 3). Công ty TNHH Thực phẩm Mây Xanh bán sữa chua uống theo lốc 4 chai; X là giá bán (nghìn đồng/lốc), Y là lượng bán (nghìn lốc/tháng), 10 tháng 01–10/2025. X : 20; 22; 24; 25; 26; 28; 30; 32; 34; 35. Y : 62; 58; 53; 50; 46; 42; 48; 37; 32; 27. Các bài tập cuối chương dùng bộ số nhỏ riêng, ghi rõ là dữ kiện đề bài.

Data used (set A, reused in Chapter 3). Mây Xanh Food Co., Ltd sells drinking yoghurt in packs of four bottles; X is price (VND thousand a pack), Y is sales (thousand packs a month), ten months from January to October 2025. X : 20, 22, 24, 25, 26, 28, 30, 32, 34, 35. Y : 62, 58, 53, 50, 46, 42, 48, 37, 32, 27. The end-of-chapter exercises use their own small data sets, marked as problem data.

2.1. Hàm hồi quy tổng thể và hàm hồi quy mẫu — Population and sample regression functions

Phân tích hồi quy nghiên cứu sự phụ thuộc của một biến (biến phụ thuộc Y , còn gọi là biến được giải thích) vào một hay nhiều biến khác (biến độc lập X , biến giải thích), nhằm ước lượng và dự báo giá trị trung bình của Y khi biết X . Hồi quy hai biến (hồi quy đơn) có đúng một biến giải thích. Ở mỗi mức X có nhiều giá trị Y khác nhau: cùng giá 25 nghìn đồng/lốc, lượng bán tháng này khác tháng khác vì thời tiết, đối thủ, mùa vụ. Tập hợp các giá trị Y ứng với một mức X tạo thành phân phối có điều kiện của Y ; trung bình của phân phối đó là kỳ vọng có điều kiện $E(Y|X)$.

Regression analysis studies how one variable (the dependent variable Y , or explained variable) depends on one or more other variables (independent or explanatory variables X), in order to estimate and forecast the mean value of Y given X . Two-variable (simple) regression has exactly one explanatory variable. Each level of X comes with many values of Y : at the same price of VND 25 thousand a pack, sales differ from month to month because of weather, rivals and season. The values of Y at a given X form the conditional distribution of Y ; its mean is the conditional expectation $E(Y|X)$.

Hàm hồi quy tổng thể (PRF) nối các kỳ vọng có điều kiện đó. Khi giả định quan hệ tuyến tính:

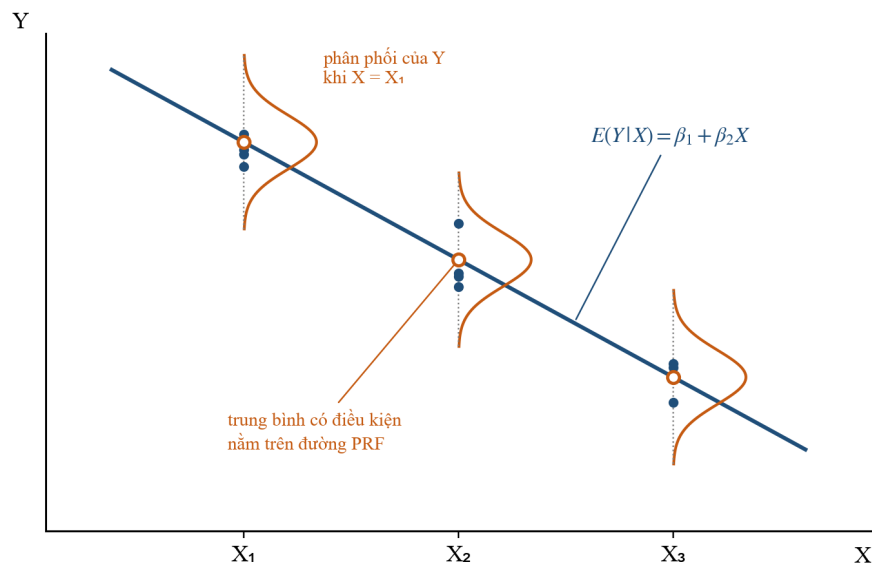
The population regression function (PRF) links those conditional expectations. Assuming a linear relation:

$$E(Y|X_i) = \beta_1 + \beta_2 X_i \quad (2.1)$$

β_1 là hệ số chặn, β_2 là hệ số góc; cả hai là hằng số của tổng thể, chưa biết. Một quan sát cụ thể lệch khỏi trung bình một khoảng U_i gọi là sai số ngẫu nhiên (nhiều):

β_1 is the intercept and β_2 the slope; both are unknown population constants. An individual observation deviates from the mean by U_i , the disturbance (error term):

$$Y_i = \beta_1 + \beta_2 X_i + U_i, \quad U_i = Y_i - E(Y|X_i) \quad (2.2)$$



Hình 2.1. Hàm hồi quy tổng thể: ở mỗi mức X , Y có một phân phối; các trung bình có điều kiện nằm trên đường PRF

Figure 2.1. The population regression function: at each level of X , Y has a distribution; the conditional means lie on the PRF

Sai số U_i gom mọi thứ mô hình không đưa vào: các biến bị bỏ qua (giá đối thủ, nhiệt độ, chi quảng cáo), yếu tố không đo được (thị hiếu), sai số đo lường của Y , dạng hàm chưa đúng hẵn và tính ngẫu nhiên trong hành vi con người. Mô hình cố ý giữ ít biến (nguyên tắc đơn giản); phần còn lại để cho U_i .

The disturbance U_i collects everything the model leaves out: omitted variables (rival prices, temperature, advertising), unmeasurable factors (tastes), measurement error in Y , an imperfect

functional form and the randomness of human behaviour. The model deliberately keeps few variables (parsimony); the rest goes into U_i .

Thực tế chỉ có một mẫu, nên chỉ ước lượng được PRF. Hàm hồi quy mẫu (SRF) và phần dư:

In practice there is only one sample, so the PRF can only be estimated. The sample regression function (SRF) and the residual:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad (2.3)$$

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i = \hat{Y}_i + e_i, \quad e_i = Y_i - \hat{Y}_i \quad (2.4)$$

$\hat{\beta}_1, \hat{\beta}_2$ là ước lượng (estimator) của β_1, β_2 : công thức tính từ mẫu, nên là biến ngẫu nhiên; con số cụ thể tính ra từ một mẫu là giá trị ước lượng (estimate). $\hat{Y}_i$ là giá trị ước lượng (giá trị khớp) của Y_i , đồng thời là ước lượng của $E(Y|X_i)$. Phần dư e_i là ước lượng của sai số U_i .

$\hat{\beta}_1$ and $\hat{\beta}_2$ are estimators of β_1 and β_2 : formulas applied to the sample, hence random variables; the numbers obtained from one sample are estimates. $\hat{Y}_i$ is the fitted value of Y_i and also an estimate of $E(Y|X_i)$. The residual e_i estimates the disturbance U_i .

Tiêu chí <i>Criterion</i>	Hàm hồi quy tổng thể (PRF) <i>Population regression function</i>	Hàm hồi quy mẫu (SRF) <i>Sample regression function</i>
Dạng <i>Form</i>	$E(Y X_i) = \beta_1 + \beta_2 X_i$	$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
Tham số <i>Parameters</i>	β_1, β_2 : hằng số, chưa biết <i>β_1, β_2: constants, unknown</i>	$\hat{\beta}_1, \hat{\beta}_2$: biến ngẫu nhiên, tính được <i>$\hat{\beta}_1, \hat{\beta}_2$: random variables, computable</i>
Số lượng <i>How many</i>	Một, duy nhất <i>One, unique</i>	Mỗi mẫu một đường <i>One line per sample</i>
Độ lệch của Y_i <i>Deviation of Y_i</i>	Sai số U_i , không quan sát được <i>Disturbance U_i, unobservable</i>	Phần dư e_i , tính được <i>Residual e_i, computable</i>
Tổng độ lệch <i>Sum of deviations</i>	$\sum U_i$ trong mẫu thường $\neq 0$ <i>Sample $\sum U_i$ usually $\neq 0$</i>	$\sum e_i = 0$ (khi có hệ số chặn) <i>$\sum e_i = 0$ (with an intercept)</i>
Ký hiệu <i>Notation</i>	Không có dấu mũ <i>No hats</i>	Có dấu mũ trên β và Y <i>Hats on β and Y</i>

LỖI HAY GẶP / COMMON MISTAKE — Viết “ $Y_i = \beta_1 + \beta_2 X_i$ ” (thiếu e_i) hoặc “ $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$ ” (thừa e_i). Hai cách viết đúng: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, hoặc $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$. Giám khảo trừ điểm ký hiệu ngay ở câu a.

Writing “ $Y_i = \beta_1 + \beta_2 X_i$ ” (missing e_i) or “ $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$ ” (extra e_i). The two correct forms are $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, or $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$. Examiners deduct for notation as early as part a.

2.2. Ý nghĩa của từ “tuyến tính” — What “linear” means

Trong kinh tế lượng, “mô hình hồi quy tuyến tính” là tuyến tính theo tham số: mỗi tham số β xuất hiện với lũy thừa bậc nhất, không nhân hay chia cho tham số khác, không nằm

trong hàm mũ, logarit. Biến X , Y thì có thể bình phương, nghịch đảo, lấy logarit. Đặt biến mới ($Z = X^2$, $Z = 1/X$, $W = \ln Y$) là đưa về dạng (2.2) và dùng được OLS.

In econometrics a “linear regression model” is linear in the parameters: each β appears to the first power, is not multiplied or divided by another parameter and does not sit inside an exponential or logarithm. The variables X and Y may be squared, inverted or logged. Defining new variables ($Z = X^2$, $Z = 1/X$, $W = \ln Y$) brings the model to form (2.2), and OLS applies.

Mô hình <i>Model</i>	Tuyến tính theo biến? <i>Linear in variables?</i>	Tuyến tính theo tham số? <i>Linear in parameters?</i>	Dùng OLS được? <i>OLS usable?</i>
$Y_i = \beta_1 + \beta_2 X_i + U_i$	Có <i>Yes</i>	Có <i>Yes</i>	Có <i>Yes</i>
$Y_i = \beta_1 + \beta_2 X_i^2 + U_i$	Không <i>No</i>	Có <i>Yes</i>	Có, đặt $Z = X^2$ <i>Yes, set $Z = X^2$</i>
$Y_i = \beta_1 + \beta_2 (1/X_i) + U_i$	Không <i>No</i>	Có <i>Yes</i>	Có, đặt $Z = 1/X$ <i>Yes, set $Z = 1/X$</i>
$\ln Y_i = \beta_1 + \beta_2 \ln X_i + U_i$	Không (theo X , Y) <i>No (in X, Y)</i>	Có <i>Yes</i>	Có, biến $\ln Y$, $\ln X$ <i>Yes, with $\ln Y$, $\ln X$</i>
$Y_i = \beta_1 \cdot X_i^{\beta_2} \cdot e^{U_i}$	Không <i>No</i>	Sau khi lấy ln: có <i>After taking logs: yes</i>	Có, sau khi log hóa <i>Yes, after logging</i>
$Y_i = \beta_1 + \beta_2^2 X_i + U_i$	Có <i>Yes</i>	Không <i>No</i>	Không trực tiếp <i>Not directly</i>
$Y_i = \beta_1 + (\beta_2/\beta_1) X_i + U_i$	Có <i>Yes</i>	Không <i>No</i>	Không trực tiếp <i>Not directly</i>

Các dạng log–log, log–lin, lin–log, nghịch đảo và cách diễn giải hệ số của chúng được trình bày ở Chương 3. Mục này chỉ cần nhận ra mô hình nào ước lượng được bằng OLS.

The log–log, log–lin, lin–log and reciprocal forms and how to read their coefficients come in Chapter 3. Here it is enough to recognise which models OLS can estimate.

LỖI HAY GẶP / COMMON MISTAKE — Cho rằng $Y = \beta_1 + \beta_2 X^2$ “không tuyến tính nên không dùng OLS”. Mô hình đó phi tuyến theo biến nhưng tuyến tính theo tham số, vẫn là mô hình hồi quy tuyến tính.

Claiming $Y = \beta_1 + \beta_2 X^2$ is “non-linear so OLS cannot be used”. The model is non-linear in the variable but linear in the parameters, so it is still a linear regression model.

2.3. Phương pháp bình phương nhỏ nhất (OLS) — *The ordinary least squares (OLS) method*

Qua đám điểm (X_i, Y_i) có thể kẻ vô số đường thẳng. OLS chọn đường có tổng bình phương phần dư nhỏ nhất:

Infinitely many lines can be drawn through the scatter of (X_i, Y_i) . OLS picks the line with the smallest sum of squared residuals:

$$\min \text{RSS} = \sum e_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad (2.5)$$

Không dùng $\sum e_i$ vì phần dư dương bù phần dư âm: mọi đường đi qua $(\bar{X}, \bar{Y})$ đều có $\sum e_i = 0$. Bình phương làm mọi độ lệch thành số dương và phạt nặng độ lệch lớn. Lấy đạo hàm riêng của RSS theo $\hat{\beta}_1, \hat{\beta}_2$ và cho bằng 0 được hệ phương trình chuẩn:

$\sum e_i$ is not used because positive residuals cancel negative ones: every line through $(\bar{X}, \bar{Y})$ has $\sum e_i = 0$. Squaring makes every deviation positive and penalises large ones heavily. Setting the partial derivatives of RSS with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ to zero gives the normal equations:

$$\partial \text{RSS} / \partial \hat{\beta}_1 = -2 \sum e_i = 0 \Rightarrow \sum Y_i = n \hat{\beta}_1 + \hat{\beta}_2 \sum X_i \quad (2.6)$$

$$\partial \text{RSS} / \partial \hat{\beta}_2 = -2 \sum e_i X_i = 0 \Rightarrow \sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \quad (2.7)$$

Giải hệ (2.6)–(2.7). Đặt độ lệch so với trung bình $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$ (chữ thường là độ lệch):

Solve (2.6)–(2.7). With deviations from the mean $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$ (lower case means deviation):

$$\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2 \quad (2.8)$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \quad (2.9)$$

Đáp án đề thi thường dùng dạng tính từ tổng thô và trung bình (không phải lập cột độ lệch). Hai dạng cho cùng kết quả vì $\sum x_i y_i = \sum X_i Y_i - n \bar{X} \bar{Y}$ và $\sum x_i^2 = \sum X_i^2 - n \bar{X}^2$:

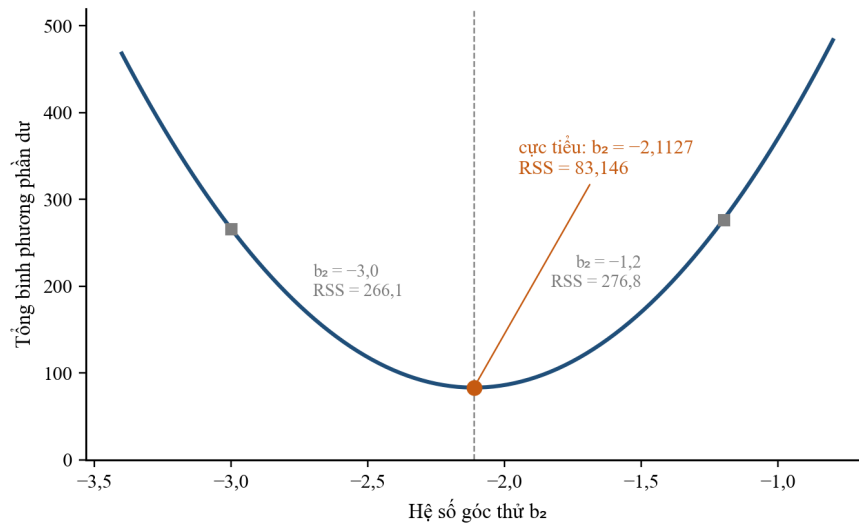
Model answers often use the form based on raw sums and means (no deviation columns). Both forms give the same result because $\sum x_i y_i = \sum X_i Y_i - n \bar{X} \bar{Y}$ and $\sum x_i^2 = \sum X_i^2 - n \bar{X}^2$:

$$\hat{\beta}_2 = (\sum X_i Y_i - n \bar{X} \bar{Y}) / (\sum X_i^2 - n \bar{X}^2) \quad (2.10)$$

$$\hat{\beta}_2 = (\text{mean}(XY) - \bar{X} \cdot \bar{Y}) / (\text{mean}(X^2) - \bar{X}^2) = \text{Cov}(X, Y) / S_x^2 \quad (2.11)$$

Trong (2.11), $\text{mean}(XY) = \sum X_i Y_i / n$, $\text{mean}(X^2) = \sum X_i^2 / n$; tử số là hiệp phương sai mẫu, mẫu số là phương sai mẫu không hiệu chỉnh của X (chia n). Hai số cùng chia n nên thương không đổi; nếu dùng S_x^2 hiệu chỉnh (chia $n - 1$) thì hiệp phương sai cũng phải chia $n - 1$.

In (2.11), $\text{mean}(XY) = \sum X_i Y_i / n$ and $\text{mean}(X^2) = \sum X_i^2 / n$; the numerator is the sample covariance and the denominator the uncorrected sample variance of X (divided by n). Both are divided by n, so the ratio is unchanged; if the corrected variance (divided by $n - 1$) is used, the covariance must also be divided by $n - 1$.



Hình 2.2. Bộ A: tổng bình phương phần dư theo hệ số góc thứ b_2 ($b_1 = \bar{Y} - b_2\bar{X}$); cực tiểu tại β_2 của OLS

Figure 2.2. Set A: the residual sum of squares against a trial slope b_2 ($b_1 = \bar{Y} - b_2\bar{X}$); the minimum is at the OLS β_2

Với $b_1 = \bar{Y} - b_2\bar{X}$, RSS là parabol theo b_2 : $RSS(b_2) = \sum y_i^2 - 2b_2\sum x_i y_i + b_2^2\sum x_i^2$. Bộ A: thử $b_2 = -3,0$ được $RSS = 266,1$; thử $b_2 = -1,2$ được $276,8$; đỉnh parabol ở $b_2 = \sum x_i y_i / \sum x_i^2 = -2,1127$ với $RSS = 83,146$, không đường thẳng nào có RSS nhỏ hơn.

With $b_1 = \bar{Y} - b_2\bar{X}$, RSS is a parabola in b_2 : $RSS(b_2) = \sum y_i^2 - 2b_2\sum x_i y_i + b_2^2\sum x_i^2$. For set A, $b_2 = -3.0$ gives $RSS = 266.1$; $b_2 = -1.2$ gives 276.8 ; the vertex is at $b_2 = \sum x_i y_i / \sum x_i^2 = -2.1127$ with $RSS = 83.146$, and no straight line has a smaller RSS.

2.4. Bài tính tay đầy đủ trên bộ A — A full hand calculation on set A

VÍ DỤ / EXAMPLE — Ước lượng hàm hồi quy mẫu của lượng bán Y theo giá X cho Công ty Mây Xanh (10 tháng). Lập bảng tính, tìm β_1 , β_2 , các giá trị $\hat{Y}_i$, phần dư e_i và RSS.

Estimate the sample regression function of sales Y on price X for Mây Xanh (ten months). Build the worksheet and find β_1 , β_2 , the fitted values $\hat{Y}_i$, the residuals e_i and RSS.

i	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	$\hat{Y}_i$	e_i	e_i^2
1	20	62	-7,6 -7.6	16,5 16.5	-125,40 -125.40	57,76 57.76	272,25 272.25	61,557 61.557	0,443 0.443	0,196 0.196
2	22	58	-5,6 -5.6	12,5 12.5	-70,00 -70.00	31,36 31.36	156,25 156.25	57,331 57.331	0,669 0.669	0,447 0.447
3	24	53	-3,6 -3.6	7,5 7.5	-27,00 -27.00	12,96 12.96	56,25 56.25	53,106 53.106	-0,106 -0.106	0,011 0.011
4	25	50	-2,6 -2.6	4,5 4.5	-11,70 -11.70	6,76 6.76	20,25 20.25	50,993 50.993	-0,993 -0.993	0,986 0.986
5	26	46	-1,6 -1.6	0,5 0.5	-0,80 -0.80	2,56 2.56	0,25 0.25	48,880 48.880	-2,880 -2.880	8,297 8.297

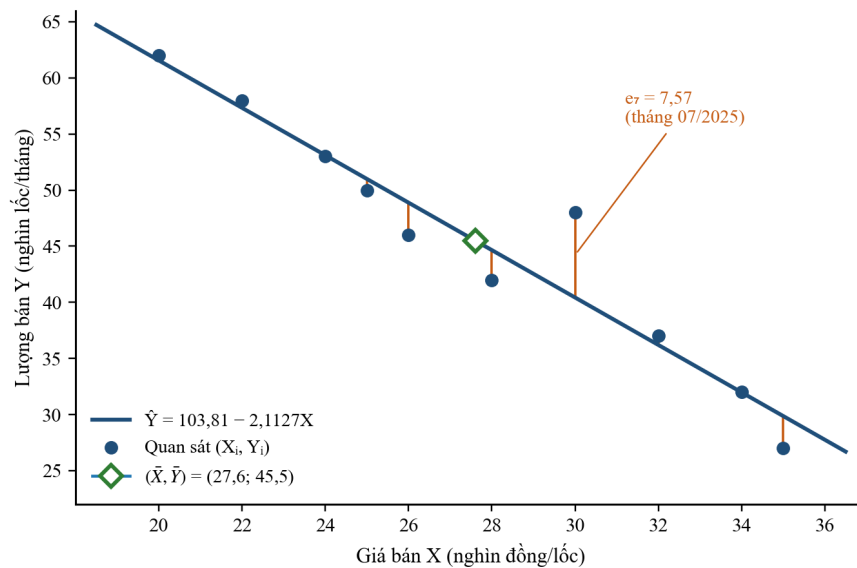
i	X _i	Y _i	x _i	y _i	x _i y _i	x _i ²	y _i ²	Ŷ _i	e _i	e _i ²
6	28	42	0,4 0.4	-3,5 -3.5	-1,40 -1.40	0,16 0.16	12,25 12.25	44,655 44.655	-2,655 -2.655	7,049 7.049
7	30	48	2,4 2.4	2,5 2.5	6,00 6.00	5,76 5.76	6,25 6.25	40,429 40.429	7,571 7.571	57,313 57.313
8	32	37	4,4 4.4	-8,5 -8.5	-37,40 -37.40	19,36 19.36	72,25 72.25	36,204 36.204	0,796 0.796	0,634 0.634
9	34	32	6,4 6.4	-13,5 -13.5	-86,40 -86.40	40,96 40.96	182,25 182.25	31,978 31.978	0,022 0.022	0,000 0.000
10	35	27	7,4 7.4	-18,5 -18.5	-136,90 -136.90	54,76 54.76	342,25 342.25	29,866 29.866	-2,866 -2.866	8,213 8.213
Σ	276	455	0	0	-491,0 -491.0	232,4 232.4	1.120,5 1,120.5	455	0,000 0.000	83,146 83.146

Tổng của cột $\hat{Y}_i$, e_i , e_i^2 tính từ số chưa làm tròn; cộng các số đã làm tròn 3 chữ số có thể lệch ở chữ số cuối. Cột x_i , y_i có tổng bằng 0 đúng: đây là phép kiểm tra đầu tiên của bảng tính.

The totals of the $\hat{Y}_i$, e_i and e_i^2 columns come from unrounded values; adding the figures rounded to three decimals may differ in the last digit. The x_i and y_i columns sum to exactly zero: the first check of the worksheet.

1. Trung bình: $\bar{X} = \Sigma X_i/n = 276/10 = 27,6$ nghìn đồng/lốc; $\bar{Y} = \Sigma Y_i/n = 455/10 = 45,5$ nghìn lốc/tháng.
Means: $\bar{X} = \Sigma X_i/n = 276/10 = 27.6$ (VND thousand a pack); $\bar{Y} = \Sigma Y_i/n = 455/10 = 45.5$ (thousand packs a month).
2. Lập cột độ lệch và cộng: $\Sigma x_i y_i = -491$; $\Sigma x_i^2 = 232,4$; $\Sigma y_i^2 = 1.120,5$.
Build the deviation columns and add: $\Sigma x_i y_i = -491$; $\Sigma x_i^2 = 232.4$; $\Sigma y_i^2 = 1,120.5$.
3. Theo (2.8): $\beta_2 = -491/232,4 = -2,1127$. Theo (2.9): $\beta_1 = 45,5 - (-2,1127) \times 27,6 = 103,8115$.
By (2.8): $\beta_2 = -491/232.4 = -2.1127$. By (2.9): $\beta_1 = 45.5 - (-2.1127) \times 27.6 = 103.8115$.
4. Kiểm lại bằng (2.10): $\Sigma X_i Y_i = 12.067$, $\Sigma X_i^2 = 7.850$; $\beta_2 = (12.067 - 10 \times 27,6 \times 45,5)/(7.850 - 10 \times 27,6^2) = -491,0/232,4 = -2,1127$. Theo (2.11): $\text{mean}(XY) = 1.206,7$, $\text{mean}(X^2) = 785$; $(1.206,7 - 1.255,8)/(785 - 761,76) = -49,10/23,24 = -2,1127$.
Check with (2.10): $\Sigma X_i Y_i = 12,067$, $\Sigma X_i^2 = 7,850$; $\beta_2 = (12,067 - 10 \times 27.6 \times 45.5)/(7,850 - 10 \times 27.6^2) = -491.0/232.4 = -2.1127$. By (2.11): $\text{mean}(XY) = 1,206.7$, $\text{mean}(X^2) = 785$; $(1,206.7 - 1,255.8)/(785 - 761.76) = -49.10/23.24 = -2.1127$.
5. Hàm hồi quy mẫu: $\hat{Y}_i = 103,8115 - 2,1127X_i$. Thế từng X_i được cột $\hat{Y}_i$; tháng 1: $\hat{Y}_1 = 103,8115 - 2,1127 \times 20 = 61,557$, $e_1 = 62 - 61,557 = 0,443$.
Sample regression function: $\hat{Y}_i = 103.8115 - 2.1127X_i$. Substituting each X_i gives the $\hat{Y}_i$ column; month 1: $\hat{Y}_1 = 103.8115 - 2.1127 \times 20 = 61.557$, $e_1 = 62 - 61.557 = 0.443$.

6. $RSS = \sum e_i^2 = 83,146$. Kiểm tra: $\sum e_i = 0$, $\sum \hat{Y}_i = \sum Y_i = 455$. Kết quả khớp phần mềm (statsmodels): $\beta_1 = 103,8115$, $\beta_2 = -2,1127$, $RSS = 83,146$.
 $RSS = \sum e_i^2 = 83.146$. Checks: $\sum e_i = 0$, $\sum \hat{Y}_i = \sum Y_i = 455$. The results match the software (statsmodels): $\beta_1 = 103.8115$, $\beta_2 = -2.1127$, $RSS = 83.146$.



Hình 2.3. Bộ A: quan sát, đường hồi quy mẫu và phần dư (đoạn dọc); đường đi qua điểm trung bình

Figure 2.3. Set A: observations, the sample regression line and residuals (vertical segments); the line passes through the point of means

Phần dư lớn nhất rơi vào tháng 07/2025 ($e_7 = 7,571$ nghìn lốc): ở giá 30 nghìn đồng/lốc lượng bán thực tế 48 nghìn lốc, cao hơn mức trung bình ước lượng 40,429. Tháng 7 là cao điểm nắng nóng; nhiệt độ không có trong mô hình nên ảnh hưởng của nó nằm trong phần dư.

The largest residual is in July 2025 ($e_7 = 7.571$ thousand packs): at a price of VND 30 thousand a pack actual sales were 48 thousand packs, above the estimated mean of 40.429. July is the peak of the hot season; temperature is not in the model, so its effect sits in the residual.

```
Python (statsmodels): import statsmodels.api as sm
                        kq = sm.OLS(Y, sm.add_constant(X)).fit(); print(kq.params,
kq.ssr)
EViews:               ls Y c X
Stata:                regress Y X
R:                    kq <- lm(Y ~ X); coef(kq); sum(resid(kq)^2)
Excel:                =SLOPE(Y;X) -> -2,1127 =INTERCEPT(Y;X) -> 103,8115
```

LỖI HAY GẶP / COMMON MISTAKE — Tính $\beta_2 = \sum x_i y_i / \sum y_i^2$ hoặc $\sum x_i^2 / \sum x_i y_i$ do nhớ lẫn. Mẹo nhớ: mẫu số là tổng bình phương độ lệch của biến giải thích X (biến nằm bên phải phương trình). Lỗi thứ hai: lấy $\sum X_i Y_i / \sum X_i^2$ (tổng thô) mà quên trừ $n\bar{X}\bar{Y}$, $n\bar{X}^2$; bộ A cho 1,5372, sai cả dấu.

Computing $\beta_2 = \sum x_i y_i / \sum y_i^2$ or $\sum x_i^2 / \sum x_i y_i$ through confusion. Memory aid: the denominator is the sum of squared deviations of the explanatory variable X (the right-hand-side variable). Second slip: taking $\sum X_i Y_i / \sum X_i^2$ (raw sums) without subtracting $n\bar{X}\bar{Y}$ and $n\bar{X}^2$; for set A this gives 1.5372, wrong even in sign.