

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.vn>

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.

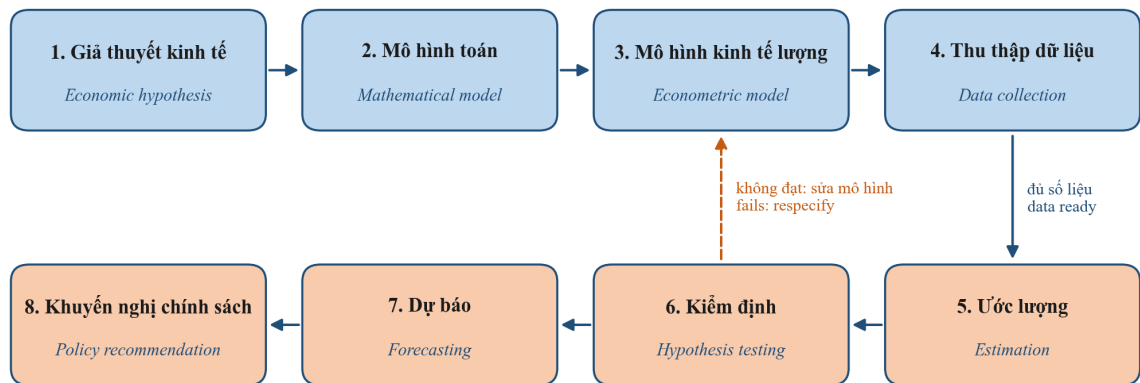
VÍ DỤ / EXAMPLE — Công ty TNHH Thực phẩm Mây Xanh bán sữa chua uống theo lốc 4 chai. Gọi Y là lượng bán (nghìn lốc/tháng), X là giá bán (nghìn đồng/lốc). Mô hình toán $Y = \beta_1 + \beta_2 X$ nói mọi tháng có cùng giá thì cùng lượng bán. Mô hình kinh tế lượng $Y_i = \beta_1 + \beta_2 X_i + u_i$ cho 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 <i>Step</i>	Việc phải làm <i>Task</i>	Ví dụ Mây Xanh <i>Mây Xanh example</i>
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.

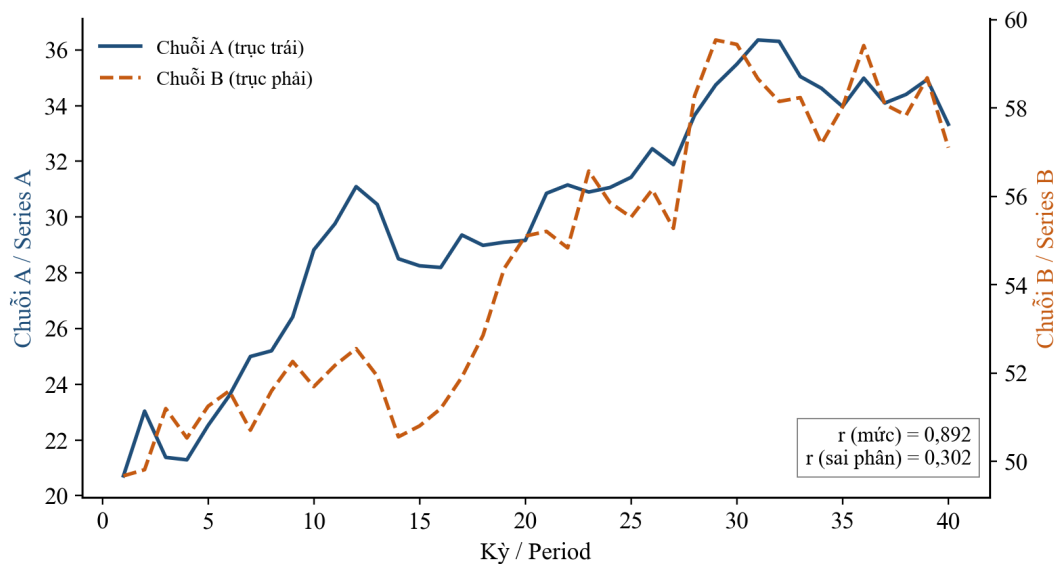
Third variable. Ice-cream sales and swimmer numbers both rise in summer; both are driven by temperature. Regressing ice-cream sales on swimmers gives a positive coefficient with no causal meaning.

- **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.

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

$$S_{XY} = \sum x_i y_i / (n - 1); \quad r = \sum x_i y_i / \sqrt{(\sum x_i^2 \cdot \sum y_i^2)} = S_{XY} / (S_X S_Y) \quad (1.11)$$

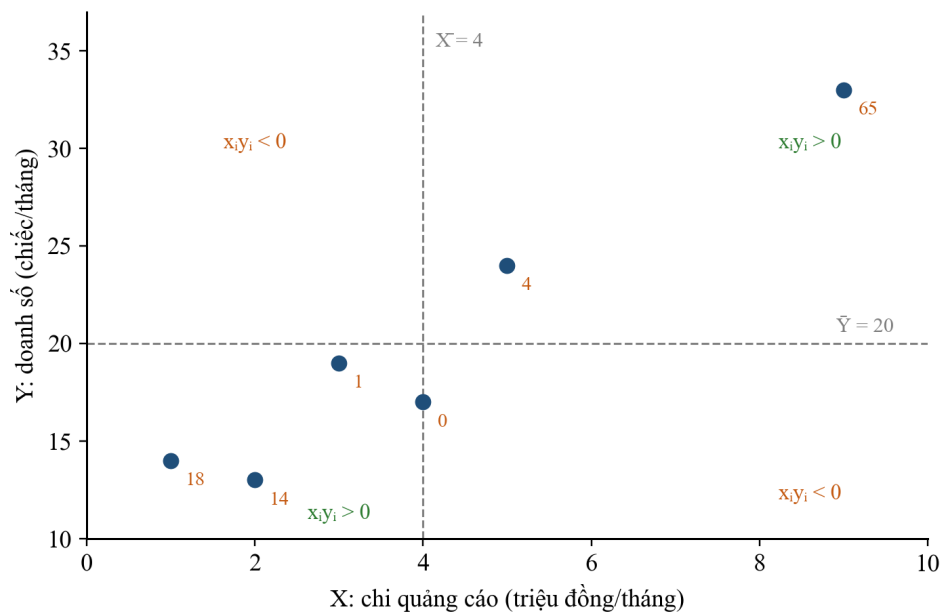
Bảng 1.3. Bài tính tay: chi quảng cáo và doanh số của 6 cửa hàng — *Table 1.3. Hand calculation: advertising and sales of 6 shops*

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**.

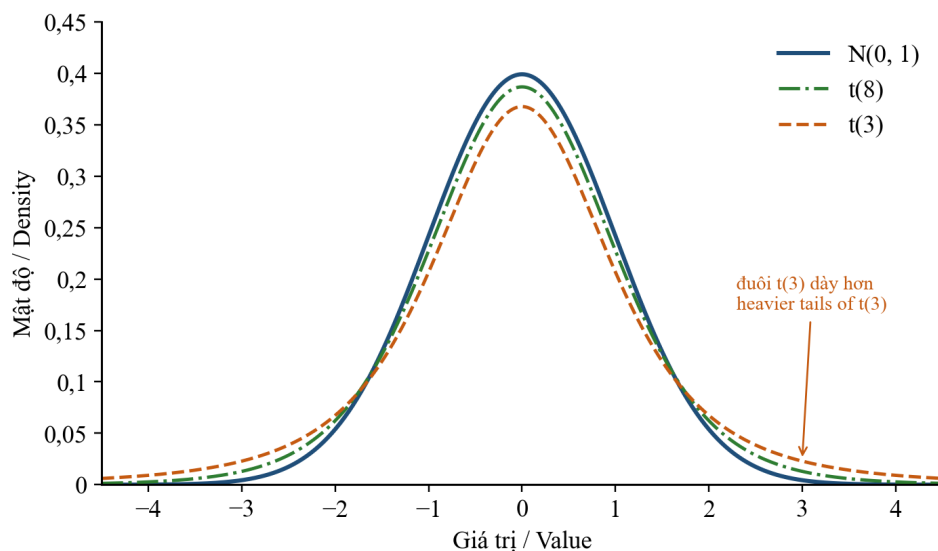
*The next three distributions are all built from standard normals. Let $Z_1, \dots, Z_k$ be independent $N(0, 1)$; $V \sim \chi^2(k)$ with Z independent of V ; $V_1 \sim \chi^2(k_1)$ and $V_2 \sim \chi^2(k_2)$ independent. The parameter k (k_1, k_2) is called the **degrees of freedom**.*

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

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

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

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).

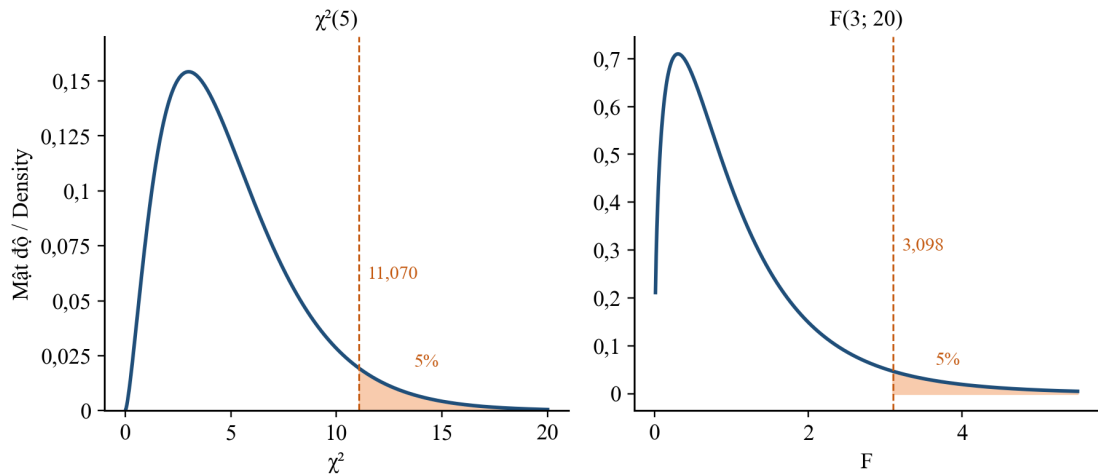
Cách tra bảng — Reading the tables

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: $\bar{X} = \sum X_i/n$ 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: $\bar{X} = \sum X_i/n$ 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 <i>Test type</i>	Cặp giả thuyết <i>Hypotheses</i>	Bác bỏ H_0 khi <i>Reject H_0 when</i>	Ví dụ câu hỏi <i>Example question</i>
Hai phía <i>Two-sided</i>	$H_0: \mu = \mu_0; H_1: \mu \neq \mu_0$	$ 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 <i>Right-tailed</i>	$H_0: \mu \leq \mu_0; H_1: \mu > \mu_0$	$t > t(\alpha; df)$	Có giao chậm hơn cam kết 40 phút không? <i>Is delivery slower than the 40-minute promise?</i>
Một phía trái <i>Left-tailed</i>	$H_0: \mu \geq \mu_0; H_1: \mu < \mu_0$	$t < -t(\alpha; df)$	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.

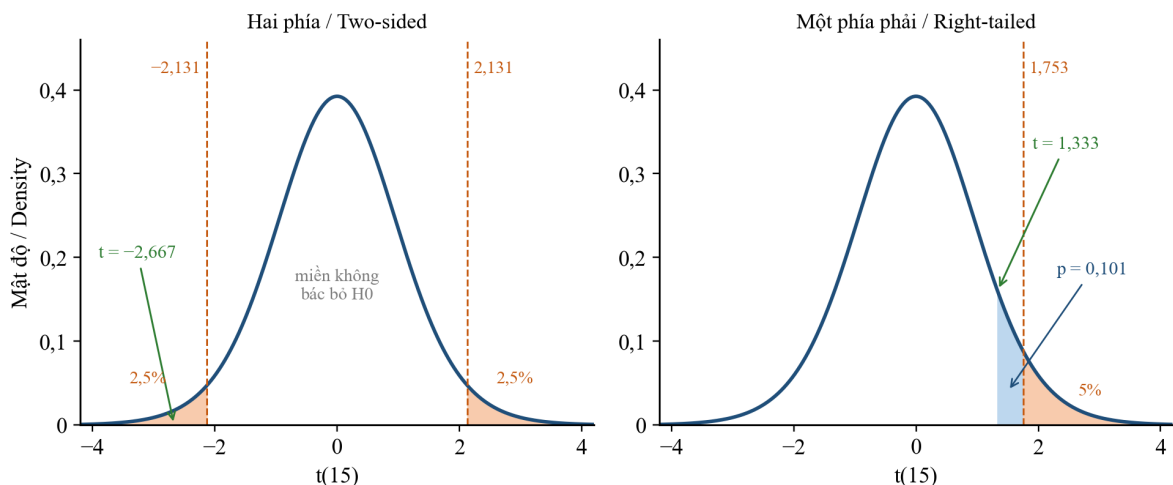
z-test. Continuing the coffee example: $H_0: \mu = 500$; $H_1: \mu \neq 500$. $z = (498 - 500)/1 = -2.00$; $|z| = 2.00 > 1.96$, so reject H_0 ; $p\text{-value} = 2 \cdot P(Z > 2.00) = 0.0455 < 0.05$. Conclusion: at 5%, the mean weight differs from the label. This matches the confidence interval in section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

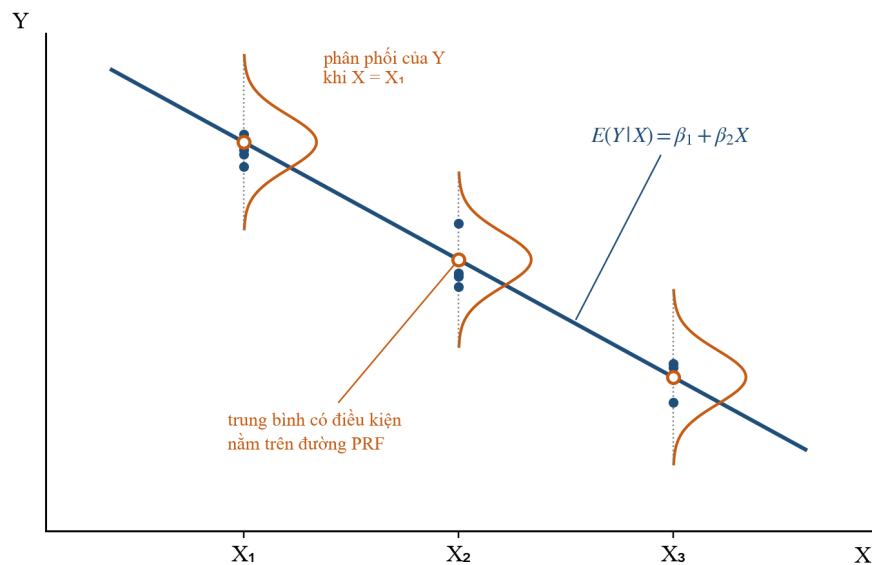
The population regression function (PRF) links those conditional expectations. Assuming a linear relation:

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

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 hẳn 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 $\sum x_i^2 = \sum X_i^2 - n \bar{X}^2$:

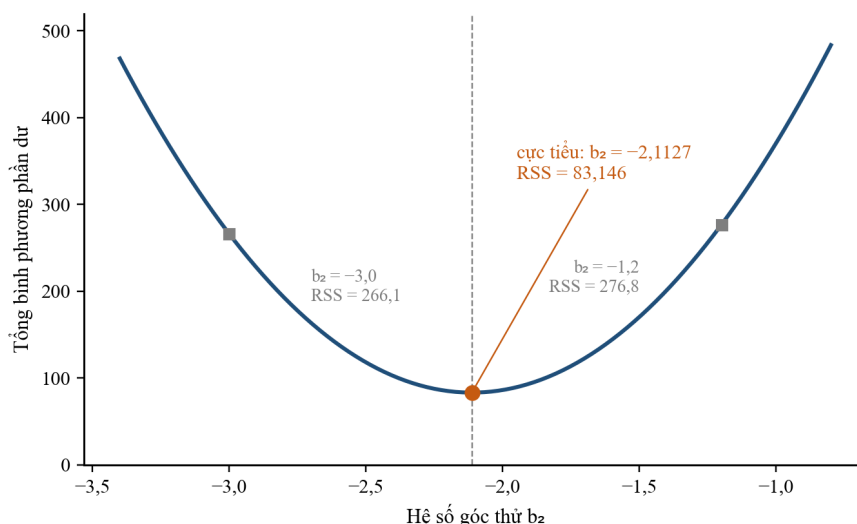
Model answers often use the form based on raw sums and means (no deviation columns). Both forms give the same result because $\sum x_i y_i = \sum X_i Y_i - n \bar{X} \bar{Y}$ and $\sum x_i^2 = \sum X_i^2 - n \bar{X}^2$:

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

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

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

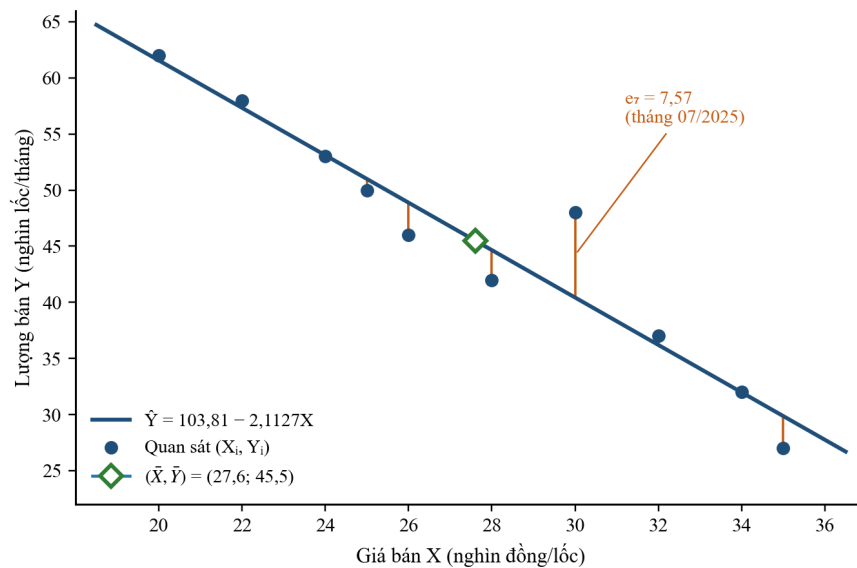
i	X _i	Y _i	x _i	y _i	x _i y _i	x _i ²	y _i ²	Ŷ _i	e _i	e _i ²
6	28	42	0,4 0.4	-3,5 -3.5	-1,40 -1.40	0,16 0.16	12,25 12.25	44,655 44.655	-2,655 -2.655	7,049 7.049
7	30	48	2,4 2.4	2,5 2.5	6,00 6.00	5,76 5.76	6,25 6.25	40,429 40.429	7,571 7.571	57,313 57.313
8	32	37	4,4 4.4	-8,5 -8.5	-37,40 -37.40	19,36 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.

2.5. Ý nghĩa của các hệ số hồi quy — *Interpreting the regression coefficients*

Hệ số góc β_2 là thay đổi trung bình của Y khi X tăng 1 đơn vị (đơn vị của X), tính bằng đơn vị của Y. Câu trả lời đủ điểm có ba phần: chiều tác động (tăng/giảm), độ lớn kèm đơn vị, và chữ “trung bình” (vì SRF ước lượng $E(Y|X)$, không phải từng giá trị Y).

The slope β_2 is the average change in Y when X rises by one unit (in X's units), measured in Y's units. A full-mark answer has three parts: direction (rise/fall), size with units, and the word “average” (the SRF estimates $E(Y|X)$, not each value of Y).

VÍ DỤ / EXAMPLE — Bộ A: $\beta_2 = -2,1127$. Khi giá bán tăng 1 nghìn đồng/lốc, lượng bán trung bình giảm 2,1127 nghìn lốc/tháng (khoảng 2.113 lốc). Dấu âm phù hợp luật cầu. $\beta_1 = 103,8115$: nếu giá bằng 0 thì lượng bán trung bình ước lượng là 103,81 nghìn lốc/tháng; giá 0 nằm xa vùng dữ liệu (20–35 nghìn đồng) và không có thật nên hệ số chặn không có ý nghĩa kinh tế, chỉ giữ vai trò định vị đường hồi quy.

Set A: $\beta_2 = -2.1127$. When the price rises by VND 1 thousand a pack, average sales fall by 2.1127 thousand packs a month (about 2,113 packs). The negative sign fits the law of demand. $\beta_1 = 103.8115$: at a zero price estimated average sales would be 103.81 thousand packs a month; a zero price lies far outside the data (VND 20–35 thousand) and does not occur, so the intercept has no economic meaning and only positions the line.

Tình huống <i>Situation</i>	Hệ số chặn β_1 <i>Intercept β_1</i>	Có ý nghĩa kinh tế? <i>Economic meaning?</i>
Lượng bán theo giá (bộ A) <i>Sales on price (set A)</i>	Lượng bán khi giá bằng 0 <i>Sales at zero price</i>	Không: $X = 0$ ngoài vùng dữ liệu <i>No: $X = 0$ is outside the data</i>
Tổng chi phí theo sản lượng <i>Total cost on output</i>	Chi phí khi sản lượng bằng 0 <i>Cost at zero output</i>	Có: chi phí cố định <i>Yes: fixed cost</i>
Chi tiêu theo thu nhập <i>Spending on income</i>	Chi tiêu khi thu nhập bằng 0 <i>Spending at zero income</i>	Có: chi tiêu tự định (nếu X gần 0 có trong mẫu) <i>Yes: autonomous spending (if X near 0 is in the sample)</i>
Tiền lương theo số năm đi học <i>Wage on years of schooling</i>	Lương khi chưa đi học năm nào <i>Wage with no schooling</i>	Thường không: mẫu ít người 0 năm học <i>Usually not: few zero-schooling cases</i>

Muốn so sánh tác động không phụ thuộc đơn vị, dùng hệ số co giãn tại điểm trung bình: phần trăm thay đổi của Y khi X tăng 1%, tính tại $(\bar{X}, \bar{Y})$:

To compare effects independently of units, use the elasticity at the means: the percentage change in Y when X rises by 1%, evaluated at $(\bar{X}, \bar{Y})$:

$$\epsilon_{YX} = \beta_2 \cdot \bar{X} / \bar{Y} \quad (2.12)$$

Bộ A: $\epsilon = -2,1127 \times 27,6/45,5 = -1,2816$. Tại mức giá và lượng bán trung bình, giá tăng 1% thì lượng bán trung bình giảm khoảng 1,28%: cầu co giãn ($|\epsilon| > 1$), giảm giá làm tăng doanh thu ở quanh điểm này. Co giãn của mô hình tuyến tính thay đổi dọc theo đường hồi quy; mô hình log–log (Chương 3) cho co giãn không đổi.

Set A: $\varepsilon = -2.1127 \times 27.6/45.5 = -1.2816$. At the average price and sales, a 1% price rise lowers average sales by about 1.28%: demand is elastic ($|\varepsilon| > 1$), so a price cut raises revenue around this point. The elasticity of a linear model changes along the line; the log-log model (Chapter 3) gives a constant elasticity.

LỖI HAY GẶP / COMMON MISTAKE — Diễn giải $\beta_2 = -2,1127$ là “giá tăng 1% thì lượng bán giảm 2,1127%”. Đó là cách đọc của mô hình log-log; ở mô hình tuyến tính hệ số góc mang đơn vị gốc (nghìn lốc trên nghìn đồng). Lỗi đi kèm: bỏ chữ “trung bình”, hoặc ghi “giảm 2,1127 lốc” trong khi Y đo bằng nghìn lốc.

Reading $\beta_2 = -2.1127$ as “a 1% price rise cuts sales by 2.1127%”. That is the log-log reading; in a linear model the slope carries the original units (thousand packs per VND thousand). Related slips: dropping the word “average”, or writing “falls by 2.1127 packs” when Y is measured in thousand packs.

2.6. Phân tích biến động: $TSS = ESS + RSS$, hệ số xác định R^2 — *Decomposing variation: $TSS = ESS + RSS$ and R^2*

Độ lệch của mỗi quan sát khỏi trung bình tách thành phần do đường hồi quy giải thích và phần dư: $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + e_i$. Bình phương rồi cộng lại; tích chéo $\sum(\hat{Y}_i - \bar{Y})e_i$ bằng 0 theo tính chất của OLS, nên:

Each observation's deviation from the mean splits into a part explained by the regression line and the residual: $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + e_i$. Square and add; the cross term $\sum(\hat{Y}_i - \bar{Y})e_i$ is zero by the OLS properties, so:

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \quad \text{hay} \quad TSS = ESS + RSS \quad (2.13)$$

Ký hiệu <i>Symbol</i>	Tên <i>Name</i>	Công thức tính nhanh <i>Quick formula</i>	Bộ A <i>Set A</i>
TSS	Tổng bình phương toàn phần <i>Total sum of squares</i>	$\sum y_i^2$	1.120,5 <i>1,120.5</i>
ESS	Tổng bình phương được giải thích <i>Explained sum of squares</i>	$\beta_2 \sum x_i y_i = \beta_2^2 \sum x_i^2$	1.037,354 <i>1,037.354</i>
RSS	Tổng bình phương phần dư <i>Residual sum of squares</i>	$\sum e_i^2 = TSS - ESS$	83,146 <i>83.146</i>

Hệ số xác định R^2 là tỷ lệ biến động của Y quanh trung bình được mô hình giải thích:

The coefficient of determination R^2 is the share of the variation of Y about its mean explained by the model:

$$R^2 = ESS/TSS = 1 - RSS/TSS, \quad 0 \leq R^2 \leq 1 \quad (2.14)$$

$$R^2 = \beta_2^2 \sum x_i^2 / \sum y_i^2 = (\sum x_i y_i)^2 / (\sum x_i^2 \cdot \sum y_i^2) \quad (2.15)$$

Hệ số tương quan mẫu r đo mức độ quan hệ tuyến tính giữa X và Y, $-1 \leq r \leq 1$. Ở hồi quy hai biến có hệ số chặn, $R^2 = r^2$ và r cùng dấu với β_2 :

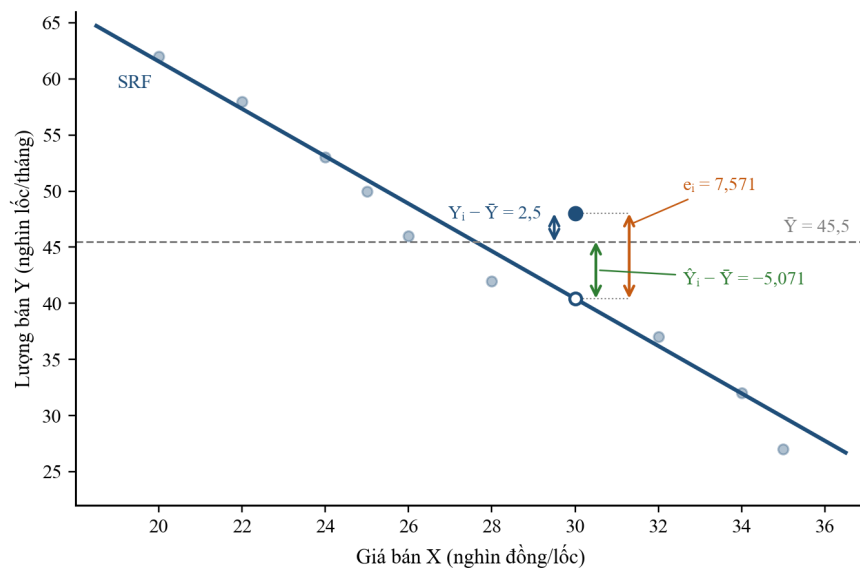
The sample correlation coefficient r measures the strength of the linear relation between X and Y, $-1 \leq r \leq 1$. In a two-variable regression with an intercept, $R^2 = r^2$ and r has the sign of β_2 :

$$r = \Sigma x_i y_i / \sqrt{(\Sigma x_i^2 \cdot \Sigma y_i^2)}, \quad R^2 = r^2, \quad \hat{\beta}_2 = r \cdot S_Y / S_X \quad (2.16)$$

VÍ DỤ / EXAMPLE — Tính TSS, ESS, RSS, R^2 và r cho bộ A.

Compute TSS, ESS, RSS, R^2 and r for set A.

1. $TSS = \Sigma y_i^2 = 1.120,5$.
 $TSS = \Sigma y_i^2 = 1,120.5$.
2. $ESS = \hat{\beta}_2 \cdot \Sigma x_i y_i = (-2,1127) \times (-491) = 1.037,354$; $RSS = TSS - ESS = 83,146$, khớp tổng cột e_i^2 trong bảng của mục 2.4.
 $ESS = \hat{\beta}_2 \cdot \Sigma x_i y_i = (-2.1127) \times (-491) = 1,037.354$; $RSS = TSS - ESS = 83.146$, matching the e_i^2 column total in the section 2.4 worksheet.
3. $R^2 = 1.037,354/1.120,5 = \mathbf{0,9258}$. Giá bán giải thích 92,58% biến động của lượng bán quanh mức trung bình 45,5 nghìn lốc; 7,42% còn lại do các yếu tố ngoài mô hình.
 $R^2 = 1,037.354/1,120.5 = \mathbf{0.9258}$. Price explains 92.58% of the variation of sales about the mean of 45.5 thousand packs; the other 7.42% comes from factors outside the model.
4. $r = -491/\sqrt{(232,4 \times 1.120,5)} = \mathbf{-0,9622}$: tương quan nghịch, chặt. Kiểm tra $r^2 = 0,9258 = R^2$.
 $r = -491/\sqrt{(232.4 \times 1,120.5)} = \mathbf{-0.9622}$: a strong negative correlation. Check: $r^2 = 0.9258 = R^2$.



Hình 2.4. Phân rã độ lệch của tháng 07/2025: $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + e_i$

Figure 2.4. Decomposing the deviation of July 2025: $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + e_i$

Hình 2.4 lấy tháng 07/2025: $Y_7 - \bar{Y} = 48 - 45,5 = 2,5$; phần đường hồi quy giải thích $\hat{Y}_7 - \bar{Y} = -5,071$ (giá 30 cao hơn giá trung bình nên lượng bán dự kiến thấp hơn trung bình); phần dư $e_7 = 7,571$. Cộng lại: $-5,071 + 7,571 = 2,5$.

Figure 2.4 uses July 2025: $Y_7 - \bar{Y} = 48 - 45.5 = 2.5$; the part explained by the line is $\hat{Y}_7 - \bar{Y} = -5.071$ (a price of 30, above the average, means expected sales below average); the residual is $e_7 = 7.571$. Adding up: $-5.071 + 7.571 = 2.5$.

Tiêu chí <i>Criterion</i>	Hệ số xác định R^2 <i>Coefficient of determination R^2</i>	Hệ số tương quan r <i>Correlation coefficient r</i>
Miền giá trị <i>Range</i>	0 đến 1 <i>0 to 1</i>	-1 đến 1 <i>-1 to 1</i>
Dấu <i>Sign</i>	Luôn không âm <i>Never negative</i>	Cùng dấu với β_2 <i>Same sign as β_2</i>
Ý nghĩa <i>Meaning</i>	Tỷ lệ biến động của Y được giải thích <i>Share of variation in Y explained</i>	Mức độ và chiều của quan hệ tuyến tính <i>Strength and direction of the linear relation</i>
Đối xứng X, Y? <i>Symmetric in X, Y?</i>	Có ở hai biến (hồi quy X theo Y cùng R^2) <i>Yes with two variables (regressing X on Y gives the same R^2)</i>	Có: $r(X, Y) = r(Y, X)$ <i>Yes: $r(X, Y) = r(Y, X)$</i>
Bộ A <i>Set A</i>	0,9258 <i>0.9258</i>	-0,9622 <i>-0.9622</i>
Mở rộng <i>Extension</i>	Dùng cho hồi quy bội (Chương 4) <i>Used in multiple regression (Chapter 4)</i>	Chỉ đo từng cặp biến <i>Pairs of variables only</i>

LỖI HAY GẶP / COMMON MISTAKE — Lấy $r = \sqrt{R^2}$ rồi ghi số dương trong khi β_2 âm. Bộ A có $R^2 = 0,9258$ nên $r = -\sqrt{0,9258} = -0,9622$. Lỗi thứ hai: nói “ $R^2 = 0,9258$ nên giá giải thích 92,58% lượng bán”; đúng là giải thích 92,58% biến động của lượng bán quanh trung bình.

Taking $r = \sqrt{R^2}$ and writing a positive number when β_2 is negative. Set A has $R^2 = 0.9258$, so $r = -\sqrt{0.9258} = -0.9622$. Second slip: saying “ $R^2 = 0.9258$, so price explains 92.58% of sales”; correctly it explains 92.58% of the variation in sales about the mean.

2.7. Ước lượng phương sai của sai số ngẫu nhiên — *Estimating the error variance*

$\sigma^2 = \text{Var}(U_i)$ chưa biết vì U_i không quan sát được. Ước lượng không chệch của σ^2 dùng RSS chia cho bậc tự do $n - 2$: mỗi tham số ước lượng (β_1, β_2) lấy đi một bậc tự do, vì phần dư phải thỏa hai ràng buộc $\sum e_i = 0$ và $\sum e_i X_i = 0$.

$\sigma^2 = \text{Var}(U_i)$ is unknown because U_i is unobservable. The unbiased estimator of σ^2 divides RSS by the degrees of freedom $n - 2$: each estimated parameter (β_1, β_2) uses up one degree of freedom, since the residuals must satisfy the two restrictions $\sum e_i = 0$ and $\sum e_i X_i = 0$.

$$\hat{\sigma}^2 = \text{RSS}/(n - 2) = \sum e_i^2/(n - 2), \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2} \quad (2.17)$$

Bộ A: $\hat{\sigma}^2 = 83,146/(10 - 2) = \mathbf{10,3933}$; $\hat{\sigma} = 3,2239$ nghìn lốc. $\hat{\sigma}$ (sai số chuẩn của hồi quy, “S.E. of regression” trong bảng EViews) là độ lệch điển hình của lượng bán thực tế quanh đường hồi quy: khoảng 3.224 lốc/tháng. Nếu mô hình có k tham số, mẫu số là $n - k$ (Chương 4).

Set A: $\sigma^2 = 83.146/(10 - 2) = 10.3933$; $\sigma = 3.2239$ thousand packs. σ (the standard error of the regression, “S.E. of regression” in EViews output) is the typical deviation of actual sales around the line: about 3,224 packs a month. With k parameters the denominator is $n - k$ (Chapter 4).

LỖI HAY GẶP / COMMON MISTAKE — Chia RSS cho n (được 8,3146) hoặc $n - 1$ (được 9,2385) thay vì $n - 2$. Sai σ^2 kéo theo sai se, t và khoảng tin cậy ở mọi câu sau của bài thi.

Dividing RSS by n (giving 8.3146) or $n - 1$ (giving 9.2385) instead of $n - 2$. A wrong σ^2 carries through to the se, t and confidence intervals in every later part of the exam.

2.8. Giả thiết của mô hình hồi quy tuyến tính cổ điển và định lý Gauss–

Markov — *Assumptions of the classical linear regression model and the Gauss–Markov theorem*

OLS cho ra con số với mọi bộ dữ liệu. Để β có tính chất thống kê tốt và để suy diễn (Chương 3) đúng, mô hình hồi quy tuyến tính cổ điển (CLRM) đặt các giả thiết sau:

OLS produces numbers for any data set. For β to have good statistical properties and for inference (Chapter 3) to be valid, the classical linear regression model (CLRM) makes the following assumptions:

Giả thiết <i>Assumption</i>	Ý nghĩa <i>Meaning</i>	Khi vi phạm <i>If violated</i>	Chương <i>Chapter</i>
1. Tuyến tính theo tham số <i>1. Linear in parameters</i>	$Y_i = \beta_1 + \beta_2 X_i + U_i$ đúng dạng <i>The model has the right form</i>	Sai dạng hàm, β chệch <i>Wrong functional form, biased β</i>	3, 9
2. X không ngẫu nhiên hoặc không tương quan với U <i>2. X fixed or uncorrelated with U</i>	X không chứa phần của sai số <i>X carries no part of the error</i>	β chệch và không vững <i>β biased and inconsistent</i>	9
3. $E(U_i X_i) = 0$	Các yếu tố bỏ qua trung bình triệt tiêu ở mọi X <i>Omitted factors average out at every X</i>	Chệch do bỏ sót biến <i>Omitted variable bias</i>	9
4. $\text{Var}(U_i X_i) = \sigma^2$	Phương sai sai số như nhau <i>Equal error variance (homoscedasticity)</i>	Không hiệu quả, se sai <i>Inefficient, wrong se</i>	7
5. $\text{Cov}(U_i, U_j) = 0, i \neq j$	Không tự tương quan <i>No autocorrelation</i>	Không hiệu quả, se sai <i>Inefficient, wrong se</i>	8
6. $n > \text{số tham số}, \Sigma x_i^2 > 0$ <i>6. $n > \text{parameters}, \Sigma x_i^2 > 0$</i>	X có biến thiên, đủ quan sát <i>X varies, enough observations</i>	Không tính được β_2 <i>β_2 cannot be computed</i>	2

Giả thiết <i>Assumption</i>	Ý nghĩa <i>Meaning</i>	Khi vi phạm <i>If violated</i>	Chương <i>Chapter</i>
7. Không đa cộng tuyến hoàn hảo (hồi quy bội) <i>7. No perfect multicollinearity (multiple regression)</i>	Không biến X nào là tổ hợp tuyến tính của biến khác <i>No X is a linear combination of others</i>	Không tách được tác động riêng <i>Separate effects not identified</i>	6
8. $U_i \sim N(0, \sigma^2)$	Sai số phân phối chuẩn <i>Normal disturbances</i>	Kiểm định t, F mẫu nhỏ không chính xác <i>Small-sample t, F tests inexact</i>	3, 9

Định lý Gauss–Markov: với các giả thiết 1–6 (không cần giả thiết 8), ước lượng OLS là ước lượng tuyến tính không chệch tốt nhất (BLUE): trong mọi ước lượng tuyến tính theo Y_i và không chệch, β OLS có phương sai nhỏ nhất.

Gauss–Markov theorem: under assumptions 1–6 (assumption 8 is not needed), the OLS estimator is the best linear unbiased estimator (BLUE): among all estimators that are linear in Y_i and unbiased, the OLS β has the smallest variance.

Chữ <i>Letter</i>	Nghĩa <i>Meaning</i>	Biểu hiện ở β_2 <i>What it means for β_2</i>
B (Best)	Tốt nhất: phương sai nhỏ nhất trong lớp <i>Best: smallest variance in the class</i>	$\text{Var}(\beta_2) = \sigma^2 / \sum x_i^2 \leq \text{Var}(\tilde{\beta}_2)$ của mọi ước lượng khác cùng lớp <i>$\text{Var}(\beta_2) = \sigma^2 / \sum x_i^2 \leq \text{Var}(\tilde{\beta}_2)$ of any rival in the class</i>
L (Linear)	Tuyến tính theo Y_i <i>Linear in Y_i</i>	$\beta_2 = \sum k_i Y_i$ với $k_i = x_i / \sum x_i^2$ <i>$\beta_2 = \sum k_i Y_i$ with $k_i = x_i / \sum x_i^2$</i>
U (Unbiased)	Không chệch <i>Unbiased</i>	$E(\beta_2) = \beta_2$: trung bình qua nhiều mẫu đúng bằng tham số <i>$E(\beta_2) = \beta_2$: the average over many samples equals the parameter</i>
E (Estimator)	Ước lượng <i>Estimator</i>	Công thức áp lên mẫu, không phải một con số cố định <i>A formula applied to a sample, not a fixed number</i>

Vững (consistent) là tính chất khi n lớn: β hội tụ về β khi n tăng. OLS vững dưới giả thiết 2–3. Không chệch nói về trung bình qua các mẫu cùng cỡ; vững nói về một mẫu ngày càng lớn.

Consistency is a large-sample property: β converges to β as n grows. OLS is consistent under assumptions 2–3. Unbiasedness concerns the average over samples of the same size; consistency concerns one ever larger sample.

LỖI HAY GẶP / COMMON MISTAKE — Ghi “định lý Gauss–Markov cần sai số phân phối chuẩn”. BLUE không cần giả thiết chuẩn; giả thiết chuẩn cần cho phân phối t , F của thống kê kiểm định ở mẫu nhỏ. Lỗi ngược lại: cho rằng có phương sai thay đổi thì β bị chệch; thực tế β vẫn không chệch, chỉ mất tính “tốt nhất” và se công thức bị sai.

Writing “the Gauss–Markov theorem needs normal disturbances”. BLUE needs no normality; normality is needed for the t and F distributions of test statistics in small samples. The opposite slip: thinking heteroscedasticity makes β biased; in fact β stays unbiased, losing only “best” and the usual se formula.

2.9. Phương sai và sai số chuẩn của β_1 , β_2 — Variances and standard errors of β_1 and β_2

β thay đổi từ mẫu này sang mẫu khác. Dưới giả thiết của CLRM:

β varies from sample to sample. Under the CLRM assumptions:

$$\text{Var}(\beta_2) = \sigma^2 / \sum x_i^2 \quad (2.18)$$

$$\text{Var}(\beta_1) = \sigma^2 \cdot \sum x_i^2 / (n \cdot \sum x_i^2) \quad (2.19)$$

$$\text{Cov}(\beta_1, \beta_2) = -\bar{X} \cdot \sigma^2 / \sum x_i^2 \quad (2.20)$$

Thay σ^2 bằng $\hat{\sigma}^2$ và lấy căn được sai số chuẩn (standard error), cột “Std. Error” của bảng EViews:

Replacing σ^2 with $\hat{\sigma}^2$ and taking square roots gives the standard errors, the “Std. Error” column of EViews output:

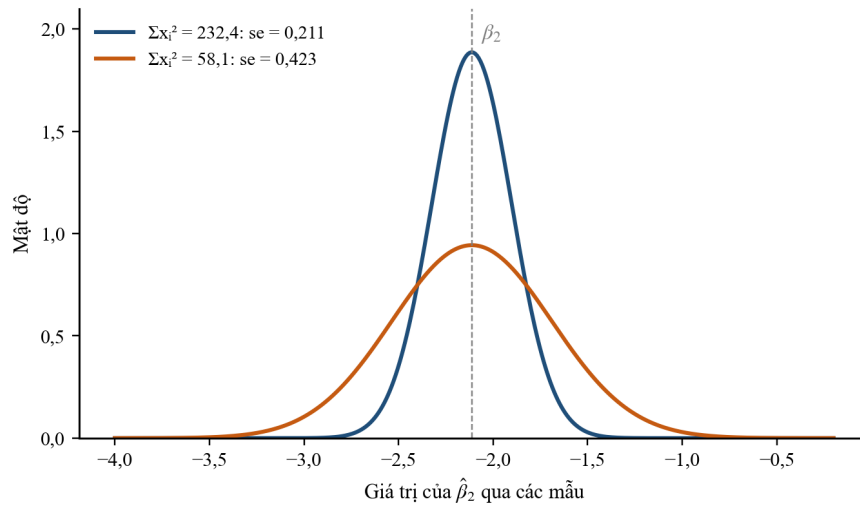
$$\text{se}(\hat{\beta}_2) = \sqrt{(\hat{\sigma}^2 / \sum x_i^2)}, \quad \text{se}(\hat{\beta}_1) = \sqrt{(\hat{\sigma}^2 \cdot \sum x_i^2 / (n \cdot \sum x_i^2))} \quad (2.21)$$

VÍ DỤ / EXAMPLE — Tính sai số chuẩn của β_1 , β_2 và hiệp phương sai ước lượng cho bộ A.
Compute the standard errors of β_1 and β_2 and the estimated covariance for set A.

- $\text{se}(\beta_2) = \sqrt{(10,3933/232,4)} = \sqrt{0,044722} = \mathbf{0,2115}$.
 $\text{se}(\beta_2) = \sqrt{(10.3933/232.4)} = \sqrt{0.044722} = \mathbf{0.2115}$.
- $\sum x_i^2 = 7.850$; $\text{se}(\beta_1) = \sqrt{(10,3933 \times 7.850 / (10 \times 232,4))} = \sqrt{35,1064} = \mathbf{5,9251}$.
 $\sum x_i^2 = 7,850$; $\text{se}(\beta_1) = \sqrt{(10.3933 \times 7,850 / (10 \times 232.4))} = \sqrt{35.1064} = \mathbf{5.9251}$.
- $\text{Cov}(\beta_1, \beta_2) = -27,6 \times 10,3933/232,4 = -1,2343$: âm vì $\bar{X} > 0$; mẫu nào có β_2 lớn hơn thì β_1 nhỏ hơn (đường xoay quanh điểm trung bình).
 $\text{Cov}(\beta_1, \beta_2) = -27.6 \times 10.3933/232.4 = -1.2343$: negative because $\bar{X} > 0$; a sample with a larger β_2 has a smaller β_1 (the line pivots about the point of means).
- Đối chiếu phần mềm (statsmodels): $\text{se} = (5,9251; 0,2115)$, khớp hai bước trên.
Software check (statsmodels): $\text{se} = (5.9251, 0.2115)$, matching the two steps above.

Từ (2.18): $\text{se}(\beta_2)$ nhỏ khi (1) σ^2 nhỏ (dữ liệu bám sát đường hồi quy), (2) $\sum x_i^2$ lớn: X trải rộng hoặc n lớn. Muốn ước lượng chính xác tác động của giá, công ty nên thử nhiều mức giá cách xa nhau thay vì giữ giá gần như cố định.

From (2.18), $\text{se}(\beta_2)$ is small when (1) σ^2 is small (data hug the line) and (2) $\sum x_i^2$ is large: X is widely spread or n is large. To estimate the price effect precisely, the firm should try well-separated prices rather than keep the price almost fixed.



Hình 2.5. Phân phối của β_2 với cùng σ^2 và n : X trải rộng như bộ A (xanh) và X chỉ phân tán một nửa (cam, Σx_i^2 còn 1/4, se gấp đôi)

Figure 2.5. The distribution of β_2 with the same σ^2 and n : X spread as in set A (blue) and X spread half as much (orange, Σx_i^2 a quarter, se doubled)

Hình 2.5 giữ $\sigma^2 = 10,3933$, $n = 10$. Nếu các mức giá chỉ phân tán một nửa quanh 27,6 thì $\Sigma x_i^2 = 58,1$, $se(\beta_2)$ tăng từ 0,2115 lên 0,4229: hai ước lượng cùng không chệch nhưng ước lượng bên phải rải rộng gấp đôi.

Figure 2.5 keeps $\sigma^2 = 10.3933$, $n = 10$. If the prices were spread only half as widely around 27.6, then $\Sigma x_i^2 = 58.1$ and $se(\beta_2)$ rises from 0.2115 to 0.4229: both estimators are unbiased, but the second is twice as dispersed.

Đại lượng Quantity	Ý nghĩa Meaning	Bộ A Set A
σ^2	Phương sai sai số tổng thể, chưa biết Population error variance, unknown	—
σ^2	Ước lượng của $\sigma^2 = RSS/(n - 2)$ Estimate of σ^2 , $RSS/(n - 2)$	10,3933 10.3933
σ	Sai số chuẩn của hồi quy Standard error of the regression	3,2239 3.2239
$Var(\beta_2)$	Phương sai của ước lượng β_2 qua các mẫu Sampling variance of β_2	0,044722 0.044722
$se(\beta_2)$	Căn của $Var(\beta_2)$ ước lượng Square root of the estimated $Var(\beta_2)$	0,2115 0.2115
$se(\beta_1)$	Sai số chuẩn của hệ số chặn Standard error of the intercept	5,9251 5.9251

LỖI HAY GẶP / COMMON MISTAKE — Nhầm σ (sai số chuẩn của hồi quy) với $se(\beta_2)$, hoặc quên lấy căn: ghi $se(\beta_2) = 0,0447$ thay vì 0,2115. Với $se(\beta_1)$, hay dùng Σx_i^2 ở tử số thay cho ΣX_i^2 (tổng thô), cho 1,0195, sai.

Confusing σ (the standard error of the regression) with $se(\beta_2)$, or forgetting the square root: writing $se(\beta_2) = 0.0447$ instead of 0.2115. For $se(\beta_1)$, using $\sum x_i^2$ in the numerator instead of $\sum X_i^2$ (raw sum) gives 1.0195, which is wrong.

2.10. Tính chất của đường hồi quy mẫu — *Properties of the sample regression line*

Từ hai phương trình chuẩn (2.6)–(2.7) suy ra các tính chất đại số sau, đúng với mọi hồi quy OLS có hệ số chặn, không cần giả thiết thống kê nào:

The two normal equations (2.6)–(2.7) imply the following algebraic properties, valid for every OLS regression with an intercept and requiring no statistical assumption:

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \quad (\text{đường hồi quy đi qua điểm } (\bar{X}, \bar{Y})) \quad (2.22)$$

$$\sum e_i = 0, \quad \text{mean}(\hat{Y}) = \bar{Y} \quad (2.23)$$

$$\sum e_i X_i = 0, \quad \sum e_i \hat{Y}_i = 0 \quad (2.24)$$

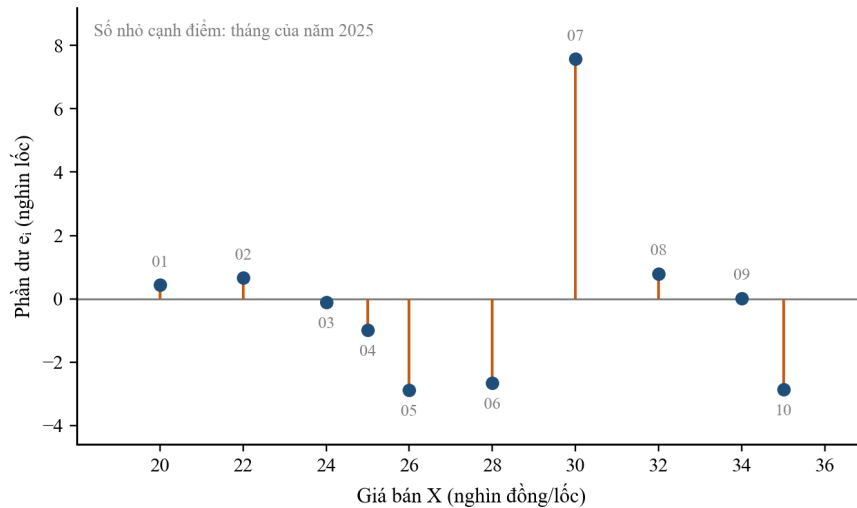
Bảng dưới kiểm các tính chất trên bộ A; tổng tính từ số chưa làm tròn.

The table checks the properties on set A; totals use unrounded values.

i	X_i^2	$X_i Y_i$	e_i	$e_i X_i$	$e_i \hat{Y}_i$	$e_i Y_i$
1	400	1.240	0,443 <i>0.443</i>	8,864 <i>8.864</i>	27,282 <i>27.282</i>	27,478 <i>27.478</i>
2	484	1.276	0,669 <i>0.669</i>	14,711 <i>14.711</i>	38,336 <i>38.336</i>	38,783 <i>38.783</i>
3	576	1.272	−0,106 <i>−0.106</i>	−2,540 <i>−2.540</i>	−5,621 <i>−5.621</i>	−5,610 <i>−5.610</i>
4	625	1.250	−0,993 <i>−0.993</i>	−24,828 <i>−24.828</i>	−50,642 <i>−50.642</i>	−49,656 <i>−49.656</i>
5	676	1.196	−2,880 <i>−2.880</i>	−74,890 <i>−74.890</i>	−140,794 <i>−140.794</i>	−132,497 <i>−132.497</i>
6	784	1.176	−2,655 <i>−2.655</i>	−74,337 <i>−74.337</i>	−118,555 <i>−118.555</i>	−111,506 <i>−111.506</i>
7	900	1.440	7,571 <i>7.571</i>	227,117 <i>227.117</i>	306,074 <i>306.074</i>	363,387 <i>363.387</i>
8	1.024	1.184	0,796 <i>0.796</i>	25,473 <i>25.473</i>	28,820 <i>28.820</i>	29,454 <i>29.454</i>
9	1.156	1.088	0,022 <i>0.022</i>	0,731 <i>0.731</i>	0,688 <i>0.688</i>	0,688 <i>0.688</i>
10	1.225	945	−2,866 <i>−2.866</i>	−100,301 <i>−100.301</i>	−85,588 <i>−85.588</i>	−77,375 <i>−77.375</i>
Σ	7.850	12.067	0,000 <i>0.000</i>	0,000 <i>0.000</i>	0,000 <i>0.000</i>	83,146 <i>83.146</i>

Ba cột e_i , $e_i X_i$, $e_i \hat{Y}_i$ có tổng bằng 0. Cột cuối có tổng $\sum e_i Y_i = 83,146 = \text{RSS}$, không phải 0: $\sum e_i Y_i = \sum e_i \hat{Y}_i + \sum e_i^2$. Điểm trung bình: $\hat{Y}(\bar{X}) = 103,8115 - 2,1127 \times 27,6 = 45,5 = \bar{Y}$. Dự báo tại giá 27 nghìn đồng/lốc (gần $\bar{X}$): $\hat{Y} = 46,768$ nghìn lốc/tháng.

The e_i , $e_i X_i$ and $e_i \hat{Y}_i$ columns sum to zero. The last column sums to $\sum e_i Y_i = 83.146 = \text{RSS}$, not zero: $\sum e_i Y_i = \sum e_i \hat{Y}_i + \sum e_i^2$. Point of means: $\hat{Y}(\bar{X}) = 103.8115 - 2.1127 \times 27.6 = 45.5 = \bar{Y}$. At a price of VND 27 thousand a pack (near $\bar{X}$): $\hat{Y} = 46.768$ thousand packs a month.



Hình 2.6. Bộ A: phần dư theo giá bán; các phần dư dương và âm bù trừ nhau ($\sum e_i = 0$, $\sum e_i X_i = 0$)

Figure 2.6. Set A: residuals against price; positive and negative residuals offset each other ($\sum e_i = 0$, $\sum e_i X_i = 0$)

Đồ thị phần dư là công cụ chẩn đoán chính ở các chương sau: phần dư rải đều quanh 0 không theo hình dạng nào là dấu hiệu tốt; hình phễu gợi ý phương sai thay đổi (Chương 7); chuỗi phần dư cùng dấu liên tiếp theo thời gian gợi ý tự tương quan (Chương 8); hình cong gợi ý sai dạng hàm (Chương 9). Với 10 quan sát của bộ A chưa thấy hình dạng rõ; tháng 07 là điểm nổi bật cần giải thích bằng yếu tố mùa vụ.

The residual plot is the main diagnostic tool in later chapters: residuals scattered evenly around zero with no pattern are a good sign; a funnel suggests heteroscedasticity (Chapter 7); runs of same-signed residuals over time suggest autocorrelation (Chapter 8); a curve suggests a wrong functional form (Chapter 9). The ten observations of set A show no clear pattern; July stands out and calls for a seasonal explanation.

Dễ nhầm <i>Easily confused</i>	Phân biệt <i>Difference</i>
Sai số U_i – phần dư e_i <i>Disturbance U_i – residual e_i</i>	U_i lệch khỏi PRF, không quan sát được; e_i lệch khỏi SRF, tính được <i>U_i deviates from the PRF, unobservable; e_i deviates from the SRF, computable</i>
Ước lượng – giá trị ước lượng <i>Estimator – estimate</i>	Công thức (biến ngẫu nhiên) – con số tính từ một mẫu <i>A formula (random variable) – a number from one sample</i>

Dễ nhầm <i>Easily confused</i>	Phân biệt <i>Difference</i>
Tuyến tính theo biến – theo tham số <i>Linear in variables – in parameters</i>	Kinh tế lượng chỉ đòi hỏi tuyến tính theo tham số <i>Econometrics only requires linearity in parameters</i>
$\sigma - se(\beta_2)$	Độ lệch điển hình của Y quanh đường – độ chính xác của β_2 <i>Typical deviation of Y about the line – precision of β_2</i>
$R^2 - r$	Tỷ lệ giải thích, không âm – tương quan, có dấu <i>Share explained, non-negative – correlation, signed</i>
Không chệch – vững <i>Unbiased – consistent</i>	Đúng trung bình ở mọi cỡ mẫu – hội tụ khi n lớn <i>Right on average at any sample size – converges as n grows</i>
$x_i - \bar{X}$	Độ lệch $X_i - \bar{X}$ – giá trị gốc; $\sum x_i^2 \neq \sum X_i^2$ <i>Deviation $X_i - \bar{X}$ – raw value; $\sum x_i^2 \neq \sum X_i^2$</i>

LỖI HAY GẶP / COMMON MISTAKE — Khi bảng tính có tổng cột e_i khác 0 rõ rệt (ví dụ 3,5) mà vẫn viết tiếp các câu sau. Tổng phần dư khác 0 (ngoài sai số làm tròn) nghĩa là β_1 hoặc β_2 đã tính sai; dừng lại tính lại trước khi làm R^2 , σ^2 và kiểm định.

Carrying on when the e_i column clearly does not sum to zero (say 3.5). A non-zero residual sum (beyond rounding) means β_1 or β_2 is wrong; stop and recompute before working out R^2 , σ^2 and tests.

PHẢI THUỘC / MEMORISE — Công thức phải thuộc. $\beta_2 = \sum x_i y_i / \sum x_i^2 = (\sum x_i y_i - n\bar{X}\bar{Y}) / (\sum x_i^2 - n\bar{X}^2)$; $\beta_1 = \bar{Y} - \beta_2 \bar{X}$. TSS = $\sum y_i^2$, ESS = $\beta_2 \sum x_i y_i$, RSS = TSS – ESS; $R^2 = \text{ESS} / \text{TSS}$; $r = \sum x_i y_i / \sqrt{(\sum x_i^2 \sum y_i^2)}$, $R^2 = r^2$, r cùng dấu β_2 . $\sigma^2 = \text{RSS} / (n - 2)$; $se(\beta_2) = \sqrt{(\sigma^2 / \sum x_i^2)}$; $se(\beta_1) = \sqrt{(\sigma^2 \sum x_i^2 / (n \sum x_i^2))}$. Co giãn tại trung bình $\beta_2 \bar{X} / \bar{Y}$. **Trình tự bài tính tay:** (1) $\bar{X}$, $\bar{Y}$; (2) cột x_i , y_i , $x_i y_i$, x_i^2 , y_i^2 và tổng (kiểm $\sum x_i = \sum y_i = 0$); (3) β_2 , β_1 , viết SRF có dấu mũ; (4) diễn giải hệ số có đơn vị và chữ “trung bình”; (5) ESS, RSS, R^2 ; (6) σ^2 , se. **Tính chất OLS:** qua $(\bar{X}, \bar{Y})$; $\sum e_i = 0$; $\sum e_i X_i = 0$; $\sum e_i \hat{Y}_i = 0$. **Gauss–Markov:** giả thiết 1–6 đủ cho BLUE, không cần phân phối chuẩn.

Formulas to memorise. $\beta_2 = \sum x_i y_i / \sum x_i^2 = (\sum x_i y_i - n\bar{X}\bar{Y}) / (\sum x_i^2 - n\bar{X}^2)$; $\beta_1 = \bar{Y} - \beta_2 \bar{X}$. TSS = $\sum y_i^2$, ESS = $\beta_2 \sum x_i y_i$, RSS = TSS – ESS; $R^2 = \text{ESS} / \text{TSS}$; $r = \sum x_i y_i / \sqrt{(\sum x_i^2 \sum y_i^2)}$, $R^2 = r^2$, r has the sign of β_2 . $\sigma^2 = \text{RSS} / (n - 2)$; $se(\beta_2) = \sqrt{(\sigma^2 / \sum x_i^2)}$; $se(\beta_1) = \sqrt{(\sigma^2 \sum x_i^2 / (n \sum x_i^2))}$. Elasticity at the means $\beta_2 \bar{X} / \bar{Y}$. **Hand-calculation sequence:** (1) $\bar{X}$, $\bar{Y}$; (2) columns x_i , y_i , $x_i y_i$, x_i^2 , y_i^2 and totals (check $\sum x_i = \sum y_i = 0$); (3) β_2 , β_1 , write the SRF with hats; (4) interpret with units and the word “average”; (5) ESS, RSS, R^2 ; (6) σ^2 , se. **OLS properties:** through $(\bar{X}, \bar{Y})$; $\sum e_i = 0$; $\sum e_i X_i = 0$; $\sum e_i \hat{Y}_i = 0$. **Gauss–Markov:** assumptions 1–6 suffice for BLUE; normality is not needed.

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>
Biến phụ thuộc	Dependent variable	Biến Y cần giải thích <i>The variable Y to be explained</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Biến giải thích	Explanatory variable	Biến X dùng để giải thích Y <i>The variable X used to explain Y</i>
Hàm hồi quy tổng thể (PRF)	Population regression function	$E(Y X) = \beta_1 + \beta_2 X$
Hàm hồi quy mẫu (SRF)	Sample regression function	$\hat{Y} = \beta_1 + \beta_2 X$ ước lượng từ mẫu <i>$\hat{Y} = \beta_1 + \beta_2 X$ estimated from a sample</i>
Kỳ vọng có điều kiện	Conditional expectation	Trung bình của Y khi X cố định <i>Mean of Y for given X</i>
Sai số ngẫu nhiên	Disturbance (error term)	$U_i = Y_i - E(Y X_i)$
Phần dư	Residual	$e_i = Y_i - \hat{Y}_i$
Giá trị ước lượng (khớp)	Fitted value	$\hat{Y}_i$ trên đường hồi quy mẫu <i>$\hat{Y}_i$ on the sample regression line</i>
Bình phương nhỏ nhất thông thường	Ordinary least squares (OLS)	Cực tiểu $\sum e_i^2$ <i>Minimise $\sum e_i^2$</i>
Hệ phương trình chuẩn	Normal equations	Điều kiện bậc nhất của OLS <i>First-order conditions of OLS</i>
Hệ số chặn	Intercept	β_1 , giá trị trung bình của Y khi X = 0 <i>β_1, mean of Y at X = 0</i>
Hệ số góc	Slope coefficient	β_2 , thay đổi trung bình của Y khi X tăng 1 đơn vị <i>β_2, average change in Y per unit of X</i>
Ước lượng	Estimator	Công thức tính tham số từ mẫu <i>Formula giving a parameter from a sample</i>
Không chệch	Unbiased	$E(\hat{\beta}) = \beta$
Hiệu quả	Efficient	Phương sai nhỏ nhất trong lớp <i>Smallest variance in the class</i>
Vững	Consistent	$\hat{\beta}$ hội tụ về β khi n tăng <i>$\hat{\beta}$ converges to β as n grows</i>
Ước lượng tuyến tính không chệch tốt nhất	BLUE	Kết luận của định lý Gauss–Markov <i>Conclusion of the Gauss–Markov theorem</i>
Tổng bình phương toàn phần	Total sum of squares (TSS)	$\sum (Y_i - \bar{Y})^2$

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Tổng bình phương được giải thích	Explained sum of squares (ESS)	$\Sigma(\hat{Y}_i - \bar{Y})^2$
Tổng bình phương phần dư	Residual sum of squares (RSS)	Σe_i^2
Hệ số xác định	Coefficient of determination	$R^2 = ESS/TSS$
Hệ số tương quan	Correlation coefficient	r , từ -1 đến 1 <i>r, from -1 to 1</i>
Bậc tự do	Degrees of freedom	$n - 2$ ở hồi quy hai biến <i>n - 2 in two-variable regression</i>
Sai số chuẩn	Standard error	Căn bậc hai phương sai ước lượng của β <i>Square root of the estimated variance of β</i>
Phương sai sai số không đổi	Homoscedasticity	$\text{Var}(U_i) = \sigma^2$ ở mọi quan sát <i>Var(U_i) = σ^2 for all i</i>

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

Bài 1. Chi tiêu theo thu nhập của 8 hộ gia đình — *Exercise 1. Spending against income for eight households*

Dữ kiện đề bài: khảo sát 8 hộ gia đình ở Cần Thơ, X là thu nhập, Y là chi tiêu tiêu dùng (cùng đơn vị triệu đồng/tháng). X: 8; 10; 12; 14; 16; 18; 20; 22. Y: 7; 8; 9; 11; 12; 12; 15; 18. (a) Ước lượng hàm hồi quy mẫu của Y theo X bằng OLS, nêu ý nghĩa hệ số. (b) Tính TSS, ESS, RSS, R^2 và hệ số tương quan r . (c) Ước lượng σ^2 , tính $se(\beta_1)$, $se(\beta_2)$. (d) Kiểm tra đường hồi quy đi qua điểm trung bình.

Problem data: a survey of eight households in Can Tho; X is income and Y is consumer spending (both in VND million a month). X: 8, 10, 12, 14, 16, 18, 20, 22. Y: 7, 8, 9, 11, 12, 12, 15, 18. (a) Estimate the sample regression function of Y on X by OLS and interpret the coefficients. (b) Compute TSS, ESS, RSS, R^2 and the correlation coefficient r . (c) Estimate σ^2 and compute $se(\beta_1)$, $se(\beta_2)$. (d) Check that the regression line passes through the point of means.

- $\bar{X} = 120/8 = 15$; $\bar{Y} = 92/8 = 11,5$. Độ lệch $x_i = X_i - \bar{X}$: -7; -5; -3; -1; 1; 3; 5; 7; $y_i = Y_i - \bar{Y}$: -4,5; -3,5; -2,5; -0,5; 0,5; 0,5; 3,5; 6,5. Cộng các cột: $\Sigma x_i y_i = 122$; $\Sigma x_i^2 = 168$; $\Sigma y_i^2 = 94$.
 $\bar{X} = 120/8 = 15$; $\bar{Y} = 92/8 = 11.5$. Deviations $x_i = X_i - \bar{X}$: -7, -5, -3, -1, 1, 3, 5, 7; $y_i = Y_i - \bar{Y}$: -4.5, -3.5, -2.5, -0.5, 0.5, 0.5, 3.5, 6.5. Column totals: $\Sigma x_i y_i = 122$; $\Sigma x_i^2 = 168$; $\Sigma y_i^2 = 94$.
- (a) $\beta_2 = 122/168 = 0,7262$; $\beta_1 = 11,5 - 0,7262 \times 15 = 0,6071$. Hàm hồi quy mẫu: $\hat{Y} = 0,6071 + 0,7262X$. Thu nhập tăng 1 triệu đồng/tháng thì chi tiêu trung bình tăng

0,7262 triệu đồng/tháng (khuynh hướng tiêu dùng biên ước lượng 0,73); $\beta_1 = 0,6071$ triệu đồng là mức chi tiêu ước lượng khi thu nhập bằng 0, nằm ngoài vùng dữ liệu (8–22) nên chỉ đọc như chi tiêu tự định, không dùng để dự báo.

(a) $\beta_2 = 122/168 = 0.7262$; $\beta_1 = 11.5 - 0.7262 \times 15 = 0.6071$. SRF: $\hat{Y} = 0.6071 + 0.7262X$. A rise of VND 1 million a month in income raises average spending by VND 0.7262 million a month (estimated marginal propensity to consume 0.73); $\beta_1 = \text{VND } 0.6071 \text{ million is estimated spending at zero income, outside the data range (8–22), so it is read as autonomous spending only, not used for forecasting.}$

3. (b) $TSS = \sum y_i^2 = 94$; $ESS = \beta_2 \cdot \sum x_i y_i = 0.7262 \times 122 = 88.5952$; $RSS = TSS - ESS = 5.4048$. $R^2 = ESS/TSS = 0.9425$: 94,25% biến động của chi tiêu quanh mức trung bình được giải thích bởi thu nhập. $r = \sum x_i y_i / \sqrt{(\sum x_i^2 \cdot \sum y_i^2)} = 122 / \sqrt{(168 \times 94)} = 0.9708$; $r^2 = 0.9425 = R^2$, r cùng dấu với β_2 .

(b) $TSS = \sum y_i^2 = 94$; $ESS = \beta_2 \cdot \sum x_i y_i = 0.7262 \times 122 = 88.5952$; $RSS = TSS - ESS = 5.4048$. $R^2 = ESS/TSS = 0.9425$: income explains 94.25% of the variation in spending about its mean. $r = \sum x_i y_i / \sqrt{(\sum x_i^2 \cdot \sum y_i^2)} = 122 / \sqrt{(168 \times 94)} = 0.9708$; $r^2 = 0.9425 = R^2$, and r has the sign of β_2 .

4. (c) $\sigma^2 = RSS/(n - 2) = 5.4048/6 = 0.9008$; $\sigma = 0.9491$. $se(\beta_2) = \sqrt{(\sigma^2 / \sum x_i^2)} = \sqrt{(0.9008/168)} = 0.0732$. $\sum X_i^2 = 1.968$, $se(\beta_1) = \sqrt{(\sigma^2 \cdot \sum X_i^2 / (n \cdot \sum X_i^2))} = \sqrt{(0.9008 \times 1.968 / (8 \times 168))} = 1.1485$.

(c) $\sigma^2 = RSS/(n - 2) = 5.4048/6 = 0.9008$; $\sigma = 0.9491$. $se(\beta_2) = \sqrt{(\sigma^2 / \sum x_i^2)} = \sqrt{(0.9008/168)} = 0.0732$. $\sum X_i^2 = 1.968$, $se(\beta_1) = \sqrt{(\sigma^2 \cdot \sum X_i^2 / (n \cdot \sum X_i^2))} = \sqrt{(0.9008 \times 1.968 / (8 \times 168))} = 1.1485$.

5. (d) Thế $X = \bar{X} = 15$: $\hat{Y} = 0.6071 + 0.7262 \times 15 = 11.5000 = \bar{Y}$. Kết quả đúng với mọi hồi quy OLS có hệ số chặn vì β_1 được tính từ chính đẳng thức $\bar{Y} = \beta_1 + \beta_2 \bar{X}$.

(d) Put $X = \bar{X} = 15$: $\hat{Y} = 0.6071 + 0.7262 \times 15 = 11.5000 = \bar{Y}$. This holds in every OLS regression with an intercept, because β_1 is computed from the identity $\bar{Y} = \beta_1 + \beta_2 \bar{X}$ itself.

Bài 2. Hàm chi phí khi đề chỉ cho các tổng — Exercise 2. A cost function when only the sums are given

Dữ kiện đề bài: xưởng in bao bì Thanh Trúc theo dõi 12 tháng, X là sản lượng (nghìn hộp/tháng), Y là tổng chi phí (triệu đồng/tháng). Đề cho: $n = 12$; $\sum X_i = 240$; $\sum Y_i = 2.636$; $\sum X_i^2 = 5.528$; $\sum Y_i^2 = 596.214$; $\sum X_i Y_i = 56.244$. (a) Ước lượng hàm chi phí $Y = \beta_1 + \beta_2 X + U$, nêu ý nghĩa kinh tế của từng hệ số. (b) Tính R^2 . (c) Tính σ^2 và $se(\beta_2)$. (d) Ước lượng chi phí trung bình khi sản lượng 20 nghìn hộp.

Problem data: the Thanh Truc packaging print shop tracks 12 months, X being output (thousand boxes a month) and Y total cost (VND million a month). Given: $n = 12$; $\sum X_i = 240$; $\sum Y_i = 2,636$; $\sum X_i^2 = 5,528$; $\sum Y_i^2 = 596,214$; $\sum X_i Y_i = 56,244$. (a) Estimate the cost function $Y = \beta_1 + \beta_2 X + U$ and give the economic meaning of each coefficient. (b) Compute R^2 . (c) Compute σ^2 and $se(\beta_2)$. (d) Estimate the average cost at an output of 20 thousand boxes.

1. $\bar{X} = 240/12 = 20$; $\bar{Y} = 2.636/12 = 219.6667$. Đổi tổng sang tổng độ lệch: $\sum x_i^2 = \sum X_i^2 - n\bar{X}^2 = 5.528 - 12 \times 20^2 = 728$; $\sum x_i y_i = \sum X_i Y_i - n\bar{X}\bar{Y} = 56.244 - 12 \times 20 \times 219.6667 = 3.524$; $\sum y_i^2 = \sum Y_i^2 - n\bar{Y}^2 = 17.172,6667$.

$\bar{X} = 240/12 = 20$; $\bar{Y} = 2,636/12 = 219.6667$. Convert raw sums into deviation sums: $\sum x_i^2 =$

$$\sum X_i^2 - n\bar{X}^2 = 5,528 - 12 \times 20^2 = 728; \sum x_i y_i = \sum X_i Y_i - n\bar{X}\bar{Y} = 56,244 - 12 \times 20 \times 219.6667 = 3,524; \sum y_i^2 = \sum Y_i^2 - n\bar{Y}^2 = 17,172.6667.$$

2. (a) $\beta_2 = 3.524/728 = 4,8407$; $\beta_1 = 219,6667 - 4,8407 \times 20 = 122,8535$. Cách tính theo trung bình (dạng hay gặp trong đáp án): $(\text{mean } XY - \bar{X} \cdot \bar{Y})/(\text{mean } X^2 - \bar{X}^2) = (4.687 - 20 \times 219,6667)/(460,6667 - 400) = 293,6667/60,6667 = 4,8407$, cùng kết quả. $\hat{Y} = 122,8535 + 4,8407X$: thêm 1 nghìn hộp thì tổng chi phí trung bình tăng 4,8407 triệu đồng (chi phí biên khoảng 4,84 nghìn đồng/hộp); $\beta_1 = 122,85$ triệu đồng/tháng là chi phí cố định ước lượng, hệ số chặn có ý nghĩa kinh tế.
 (a) $\beta_2 = 3,524/728 = 4.8407$; $\beta_1 = 219.6667 - 4.8407 \times 20 = 122.8535$. Using means (the form common in model answers): $(\text{mean } XY - \bar{X} \cdot \bar{Y})/(\text{mean } X^2 - \bar{X}^2) = (4.687 - 20 \times 219.6667)/(460.6667 - 400) = 293.6667/60.6667 = 4.8407$, the same result. $\hat{Y} = 122.8535 + 4.8407X$: one thousand more boxes raises average total cost by VND 4.8407 million (a marginal cost of about VND 4.84 thousand a box); $\beta_1 = \text{VND } 122.85 \text{ million a month is the estimated fixed cost, so the intercept has an economic meaning.}$
3. (b) $ESS = \beta_2 \cdot \sum x_i y_i = 17.058,4835$; $R^2 = ESS/\sum y_i^2 = 17.058,4835/17.172,6667 = 0,9934$. Cách khác: $R^2 = (\sum x_i y_i)^2/(\sum x_i^2 \cdot \sum y_i^2) = 3.524^2/(728 \times 17.172,6667) = 0,9934$.
 (b) $ESS = \beta_2 \cdot \sum x_i y_i = 17,058.4835$; $R^2 = ESS/\sum y_i^2 = 17,058.4835/17,172.6667 = 0,9934$.
 Alternatively $R^2 = (\sum x_i y_i)^2/(\sum x_i^2 \cdot \sum y_i^2) = 3,524^2/(728 \times 17,172.6667) = 0.9934$.
4. (c) $RSS = \sum y_i^2 - ESS = 114,1832$; $\sigma^2 = RSS/(12 - 2) = 11,4183$; $se(\beta_2) = \sqrt{(11,4183/728)} = 0,1252$; $se(\beta_1) = \sqrt{(\sigma^2 \cdot \sum X_i^2)/(n \cdot \sum x_i^2)} = 2,6880$.
 (c) $RSS = \sum y_i^2 - ESS = 114.1832$; $\sigma^2 = RSS/(12 - 2) = 11.4183$; $se(\beta_2) = \sqrt{(11.4183/728)} = 0.1252$; $se(\beta_1) = \sqrt{(\sigma^2 \cdot \sum X_i^2)/(n \cdot \sum x_i^2)} = 2.6880$.
5. (d) $X_0 = 20 = \bar{X}$ nên $\hat{Y}_0 = \bar{Y} = 219,67$ triệu đồng/tháng (không cần thế số: đường hồi quy mẫu đi qua điểm trung bình). Khoảng tin cậy cho giá trị này thuộc Chương 3.
 (d) $X_0 = 20 = \bar{X}$, so $\hat{Y}_0 = \bar{Y} = \text{VND } 219.67 \text{ million a month (no substitution needed: the sample regression line passes through the point of means). The confidence interval for this value belongs to Chapter 3.}$

Bài 3. Tìm lại phần dư bị mất — Exercise 3. Recovering lost residuals

Dữ kiện đề bài: cửa hàng điện thoại Ánh Dương hồi quy doanh số tháng Y (chục triệu đồng) của 6 nhân viên theo số năm kinh nghiệm X bằng OLS, được $\hat{Y} = 20 + 3X$. X : 2; 4; 5; 7; 8; 10. Bảng phần dư bị nhòe hai ô cuối; bốn ô đầu: 1,0; -2,0; 0,5; 1,5. (a) Tìm e_5 , e_6 . (b) Tìm Y_5 , Y_6 . (c) Tính RSS , σ^2 , $se(\beta_2)$, $se(\beta_1)$. (d) Tính R^2 .

Problem data: the Anh Duong phone shop regresses monthly sales Y (VND 10 million) of six salespeople on years of experience X by OLS and obtains $\hat{Y} = 20 + 3X$. X : 2, 4, 5, 7, 8, 10. The last two cells of the residual column are smudged; the first four are 1.0, -2.0, 0.5, 1.5. (a) Find e_5 and e_6 . (b) Find Y_5 and Y_6 . (c) Compute RSS , σ^2 , $se(\beta_2)$, $se(\beta_1)$. (d) Compute R^2 .

1. (a) Hai tính chất của phần dư OLS khi có hệ số chặn: $\sum e_i = 0$ và $\sum e_i X_i = 0$. Hệ hai phương trình: $e_5 + e_6 = -(1,0 + -2,0 + 0,5 + 1,5) = -1,0$; $8e_5 + 10e_6 = -(2 \times (1,0) + 4 \times (-2,0) + 5 \times (0,5) + 7 \times (1,5)) = -7,0$. Giải: $e_5 = -1,5$; $e_6 = 0,5$.
 (a) Two properties of OLS residuals with an intercept: $\sum e_i = 0$ and $\sum e_i X_i = 0$. The system: $e_5 + e_6 = -(1.0 + -2.0 + 0.5 + 1.5) = -1.0$; $8e_5 + 10e_6 = -(2 \times (1.0) + 4 \times (-2.0) + 5 \times (0.5) + 7 \times (1.5)) = -7.0$. Solving: $e_5 = -1.5$, $e_6 = 0.5$.

2. (b) $Y_i = \hat{Y}_i + e_i$; $Y_5 = 20 + 3 \times 8 + (-1,5) = 42,5$; $Y_6 = 20 + 3 \times 10 + 0,5 = 50,5$. Hồi quy lại 6 cặp số cho đúng $\hat{Y} = 20 + 3X$.
 (b) $Y_i = \hat{Y}_i + e_i$; $Y_5 = 20 + 3 \times 8 + (-1.5) = 42.5$; $Y_6 = 20 + 3 \times 10 + 0.5 = 50.5$. Re-running the regression on the six pairs gives exactly $\hat{Y} = 20 + 3X$.
3. (c) $RSS = \sum e_i^2 = 1,00 + 4,00 + 0,25 + 2,25 + 2,25 + 0,25 = 10$; $\sigma^2 = 10/(6 - 2) = 2,5$. $\bar{X} = 6$, $\sum X_i^2 = 42$, $\sum X_i Y_i = 258$. $se(\beta_2) = \sqrt{(2,5/42)} = 0,2440$; $se(\beta_1) = \sqrt{(2,5 \times 258/(6 \times 42))} = 1,5999$.
 (c) $RSS = \sum e_i^2 = 1.00 + 4.00 + 0.25 + 2.25 + 2.25 + 0.25 = 10$; $\sigma^2 = 10/(6 - 2) = 2.5$. $\bar{X} = 6$, $\sum X_i^2 = 42$, $\sum X_i Y_i = 258$. $se(\beta_2) = \sqrt{(2.5/42)} = 0.2440$; $se(\beta_1) = \sqrt{(2.5 \times 258/(6 \times 42))} = 1.5999$.
4. (d) $ESS = \beta_2^2 \cdot \sum X_i^2 = 3^2 \times 42 = 378$; $TSS = ESS + RSS = 388$; $R^2 = 378/388 = \mathbf{0,9742}$. Không cần biết $Y_1 \dots Y_6$ vẫn tính được R^2 .
 (d) $ESS = \beta_2^2 \cdot \sum X_i^2 = 3^2 \times 42 = 378$; $TSS = ESS + RSS = 388$; $R^2 = 378/388 = \mathbf{0.9742}$. R^2 is found without knowing $Y_1 \dots Y_6$.

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

Câu 1. Phân biệt hàm hồi quy tổng thể và hàm hồi quy mẫu; sai số ngẫu nhiên và phần dư. Minh họa bằng số liệu của Mây Xanh.

Question 1. Distinguish the population regression function from the sample regression function, and the disturbance from the residual. Illustrate with the Mây Xanh data.

- PRF: $E(Y|X_i) = \beta_1 + \beta_2 X_i$, mô tả trung bình của Y ở từng mức X trong tổng thể; β_1, β_2 là hằng số chưa biết.
PRF: $E(Y|X_i) = \beta_1 + \beta_2 X_i$ describes the mean of Y at each level of X in the population; β_1, β_2 are unknown constants.
- SRF: $\hat{Y}_i = \beta_1 + \beta_2 X_i$, ước lượng từ một mẫu; với Mây Xanh $\hat{Y} = 103,8115 - 2,1127X$. Mẫu khác cho số khác: β là biến ngẫu nhiên.
SRF: $\hat{Y}_i = \beta_1 + \beta_2 X_i$, estimated from one sample; for Mây Xanh $\hat{Y} = 103.8115 - 2.1127X$. Another sample gives other numbers: β is a random variable.
- Sai số $U_i = Y_i - E(Y|X_i)$: khoảng cách tới đường tổng thể, không quan sát được, gom các yếu tố bỏ qua và nhiễu.
The disturbance $U_i = Y_i - E(Y|X_i)$: the distance to the population line, unobservable, collecting omitted factors and noise.
- Phần dư $e_i = Y_i - \hat{Y}_i$: tính được sau khi ước lượng; tháng 07/2025 $e = 7,571$ nghìn lốc. e_i là ước lượng của U_i , $\sum e_i = 0$ luôn đúng nhưng $\sum U_i$ trong mẫu không nhất thiết bằng 0.
The residual $e_i = Y_i - \hat{Y}_i$: computable after estimation; in July 2025 $e = 7.571$ thousand packs. e_i estimates U_i ; $\sum e_i = 0$ always holds, whereas the sample $\sum U_i$ need not be zero.
- Viết đúng ký hiệu: PRF không có dấu mũ, SRF có dấu mũ; phương trình có Y (không mũ) thì phải có số hạng U hoặc e .
Use the notation correctly: no hats in the PRF, hats in the SRF; an equation for Y (no hat) must carry a U or e term.

Câu 2. Nêu các giả thiết của mô hình hồi quy tuyến tính cổ điển và phát biểu định lý Gauss–Markov. Giả thiết phân phối chuẩn của sai số cần cho việc gì?

Question 2. *State the assumptions of the classical linear regression model and the Gauss–Markov theorem. What is the normality assumption on the disturbance needed for?*

- Mô hình tuyến tính theo tham số; X không ngẫu nhiên (hoặc không tương quan với U) và có biến thiên ($\sum x_i^2 > 0$).
The model is linear in parameters; X is non-random (or uncorrelated with U) and varies ($\sum x_i^2 > 0$).
- $E(U_i|X_i) = 0$; $\text{Var}(U_i|X_i) = \sigma^2$ như nhau ở mọi quan sát; $\text{Cov}(U_i, U_j) = 0$ khi $i \neq j$; $n >$ số tham số; mô hình chỉ định đúng.
 $E(U_i|X_i) = 0$; $\text{Var}(U_i|X_i) = \sigma^2$, the same for every observation; $\text{Cov}(U_i, U_j) = 0$ for $i \neq j$; n exceeds the number of parameters; the model is correctly specified.
- Gauss–Markov: dưới các giả thiết trên, β OLS là ước lượng tuyến tính không chệch có phương sai nhỏ nhất (BLUE) trong lớp ước lượng tuyến tính không chệch.
Gauss–Markov: under these assumptions the OLS β is the best linear unbiased estimator (BLUE), having the smallest variance among linear unbiased estimators.
- Định lý không cần U phân phối chuẩn. Giả thiết chuẩn dùng để biết phân phối của β (chuẩn), của $(\hat{\beta} - \beta)/\text{se}(\hat{\beta})$ (Student t) và RSS/σ^2 (χ^2) khi mẫu nhỏ, làm cơ sở cho khoảng tin cậy và kiểm định ở Chương 3.
The theorem does not require normal disturbances. Normality gives the distributions of β (normal), of $(\hat{\beta} - \beta)/\text{se}(\hat{\beta})$ (Student t) and of RSS/σ^2 (χ^2) in small samples, the basis for the intervals and tests of Chapter 3.
- Vi phạm phương sai bằng nhau (Chương 7) hay không tự tương quan (Chương 8): β vẫn không chệch nhưng không còn BLUE, se công thức thường bị sai.
If equal variance (Chapter 7) or no autocorrelation (Chapter 8) fails, β stays unbiased but is no longer BLUE and the usual se formula is wrong.

Câu 3. Vì sao OLS cực tiểu tổng bình phương phần dư mà không cực tiểu tổng phần dư? Trình bày các tính chất đại số của đường hồi quy mẫu và cách dùng chúng khi làm bài.

Question 3. *Why does OLS minimise the sum of squared residuals rather than the sum of residuals? Set out the algebraic properties of the sample regression line and how to use them in an exam.*

- Tổng phần dư có thể bằng 0 với vô số đường thẳng (mọi đường qua điểm trung bình), phần dư dương bù phần dư âm: tiêu chí không chọn được đường duy nhất.
The sum of residuals can be zero for infinitely many lines (every line through the point of means), positive residuals offsetting negative ones: the criterion picks no unique line.
- Bình phương làm mọi sai lệch đều dương, phạt nặng sai lệch lớn, cho hàm khả vi; điều kiện bậc nhất là hai phương trình chuẩn tuyến tính, có nghiệm duy nhất khi $\sum x_i^2 > 0$.
Squaring makes every deviation positive, penalises large ones heavily and gives a differentiable function; the first-order conditions are two linear normal equations with a unique solution when $\sum x_i^2 > 0$.

- Tính chất: đường qua $(\bar{X}, \bar{Y})$; trung bình $\hat{Y}$ bằng $\bar{Y}$; $\sum e_i = 0$; $\sum e_i X_i = 0$; $\sum e_i \hat{Y}_i = 0$.
Properties: the line passes through $(\bar{X}, \bar{Y})$; the mean of $\hat{Y}$ equals $\bar{Y}$; $\sum e_i = 0$; $\sum e_i X_i = 0$; $\sum e_i \hat{Y}_i = 0$.
- Dùng khi làm bài: kiểm tra cột e_i (tổng phải xấp xỉ 0); tìm ô phần dư bị khuyết; dự báo nhanh tại $\bar{X}$; tìm β_1 khi đã có β_2 và hai trung bình.
Exam uses: check the e_i column (its total must be about 0); recover missing residuals; forecast quickly at $\bar{X}$; find β_1 from β_2 and the two means.
- Các tính chất này chỉ đúng khi mô hình có hệ số chặn; hồi quy qua gốc tọa độ (Chương 3) không bảo đảm $\sum e_i = 0$.
These properties need an intercept; regression through the origin (Chapter 3) does not guarantee $\sum e_i = 0$.

2.13. Câu hỏi trắc nghiệm — Multiple-choice questions

9. Bộ số liệu Mây Xanh có $\sum x_i y_i = -491$ và $\sum x_i^2 = 232,4$. Hệ số góc β_2 bằng:
The Mây Xanh data give $\sum x_i y_i = -491$ and $\sum x_i^2 = 232,4$. The slope β_2 equals:
- A. -2,1127 / -2.1127
 - B. 2,1127 / 2.1127
 - C. -0,4733 / -0.4733
 - D. -0,4382 / -0.4382
10. Mô hình nào KHÔNG tuyến tính theo tham số?
Which model is NOT linear in the parameters?
- A. $Y_i = \beta_1 + \beta_2 X_i^2 + U_i$
 - B. $Y_i = \beta_1 + \beta_2 (1/X_i) + U_i$
 - C. $\ln Y_i = \beta_1 + \beta_2 \ln X_i + U_i$
 - D. $Y_i = \beta_1 + \beta_2^3 X_i + U_i$
11. Hồi quy hai biến với $n = 20$ quan sát có $RSS = 48,6$. Ước lượng của phương sai sai số σ^2 là:
A two-variable regression on $n = 20$ observations has $RSS = 48.6$. The estimate of the error variance σ^2 is:
- A. 2,43 / 2.43
 - B. 1,64 / 1.64
 - C. 2,70 / 2.70
 - D. 2,56 / 2.56
12. Cho $\sum x_i y_i = 60$, $\sum x_i^2 = 40$, $\sum y_i^2 = 100$. Hệ số xác định R^2 bằng:
Given $\sum x_i y_i = 60$, $\sum x_i^2 = 40$, $\sum y_i^2 = 100$. The coefficient of determination R^2 is:
- A. 0,60 / 0.60
 - B. 0,90 / 0.90
 - C. 0,95 / 0.95
 - D. 1,50 / 1.50

13. Với hồi quy OLS có hệ số chặn, đẳng thức nào KHÔNG luôn đúng?

In an OLS regression with an intercept, which equality does NOT always hold?

- A. $\sum e_i = 0$
- B. $\sum e_i X_i = 0$
- C. $\bar{Y} = \beta_1 + \beta_2 \bar{X}$
- D. $\sum e_i Y_i = 0$

14. Định lý Gauss–Markov KHÔNG cần giả thiết nào sau đây?

Which assumption is NOT needed for the Gauss–Markov theorem?

- A. Sai số ngẫu nhiên có phân phối chuẩn / *Normally distributed disturbances*
- B. Phương sai sai số không đổi / *Constant error variance*
- C. Các sai số không tương quan với nhau / *Uncorrelated disturbances*
- D. Kỳ vọng có điều kiện của sai số bằng 0 / *Zero conditional mean of the disturbance*

15. Hồi quy hai biến có $\sigma^2 = 12,5$ và $\sum x_i^2 = 200$. Sai số chuẩn $se(\beta_2)$ bằng:

A two-variable regression has $\sigma^2 = 12.5$ and $\sum x_i^2 = 200$. The standard error $se(\beta_2)$ is:

- A. 0,0625 / *0.0625*
- B. 0,0177 / *0.0177*
- C. 0,2500 / *0.2500*
- D. 4,0000 / *4.0000*

16. Mây Xanh: $\hat{Y} = 103,8115 - 2,1127X$, X là giá (nghìn đồng/lốc), Y là lượng bán (nghìn lốc/tháng). Cách diễn giải đúng là:

Mây Xanh: $\hat{Y} = 103.8115 - 2.1127X$, X being price (VND thousand a pack) and Y sales (thousand packs a month). The correct reading is:

- A. Giá tăng 1% thì lượng bán giảm 2,1127% / *A 1% price rise cuts sales by 2.1127%*
- B. Giá tăng 1 nghìn đồng/lốc thì lượng bán trung bình giảm 2,1127 nghìn lốc/tháng / *A VND 1 thousand rise in price cuts average sales by 2.1127 thousand packs a month*
- C. Giá tăng 1 nghìn đồng/lốc thì lượng bán giảm đúng 2,1127 lốc mỗi tháng / *A VND 1 thousand rise in price cuts sales by exactly 2.1127 packs every month*
- D. Khi giá bằng 0 công ty bán được 103,8115 nghìn lốc, đó là sức chứa của thị trường / *At a zero price the firm sells 103.8115 thousand packs, the capacity of the market*

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

Mục tiêu chương — Chapter objectives

- Nêu được phân phối của β_1 , β_2 và σ^2 khi sai số ngẫu nhiên có phân phối chuẩn; lập khoảng tin cậy cho β_1 , β_2 , σ^2 .
State the sampling distributions of β_1 , β_2 and σ^2 under normal disturbances; build confidence intervals for β_1 , β_2 and σ^2 .
- Kiểm định t hai phía, một phía và $H_0: \beta_2 = \beta^*$; kết luận bằng miền bác bỏ, khoảng tin cậy và p-value.
Carry out two-sided, one-sided and $H_0: \beta_2 = \beta^$ t-tests; conclude by rejection region, confidence interval and p-value.*
- Lập bảng ANOVA, kiểm định F; dự báo giá trị trung bình và giá trị cá biệt kèm khoảng dự báo.
Build the ANOVA table and run the F-test; forecast the mean and an individual value with prediction intervals.
- Xử lý đổi đơn vị đo, chọn và diễn giải các dạng hàm log–log, log–lin, lin–log, nghịch đảo, hồi quy qua gốc tọa độ.
Handle changes of units; choose and interpret log–log, log–lin, lin–log, reciprocal forms and regression through the origin.
- Đọc từng ô của bảng kết quả EViews và giải trọn một câu thi hồi quy hai biến dạng đề UFM.
Read every cell of an EViews output table and solve a full two-variable regression exam question in the UFM format.

3.1. Phân phối xác suất của các ước lượng OLS — Sampling distributions of the OLS estimators

Chương 2 đã ước lượng hàm hồi quy mẫu của bộ A (Công ty TNHH Thực phẩm Mây Xanh, sữa chua uống, 10 tháng năm 2025): $\hat{Y} = 103,8115 - 2,1127X$, với X là giá bán (nghìn đồng/lốc 4 chai) và Y là lượng bán (nghìn lốc/tháng). $\beta_2 = -2,1127$ chỉ là một giá trị từ một mẫu; mẫu 10 tháng khác cho β_2 khác. Để nói β_2 tổng thể nằm trong khoảng nào, hay β_2 có khác 0 không, cần biết β_2 dao động ra sao quanh β_2 .

Chapter 2 estimated the sample regression function of data set A (Mây Xanh Foods Co., Ltd., drinking yoghurt, 10 months of 2025): $\hat{Y} = 103.8115 - 2.1127X$, where X is the price (thousand VND per pack of 4 bottles) and Y is the quantity sold (thousand packs a month). $\beta_2 = -2.1127$ is a single value from one sample; another 10-month sample gives another β_2 . To say where the population β_2 lies, or whether it differs from 0, we need to know how β_2 varies around β_2 .

PHẢI THUỘC / MEMORISE — Giả thiết bổ sung: $u_i \sim N(0, \sigma^2)$, độc lập. Khi đó β_1 , β_2 là tổ hợp tuyến tính của các Y_i có phân phối chuẩn nên cũng có **phân phối chuẩn**; $\sigma^2 = \text{RSS}/(n - 2)$ liên hệ với phân phối χ^2 bậc $n - 2$. Thay σ chưa biết bằng σ thì thống kê chuẩn hóa có

phân phối **t bậc n – 2**.

*Added assumption: $u_i \sim N(0, \sigma^2)$, independent. Then β_1 and β_2 , being linear combinations of normal Y_i , are **normally distributed**; $\sigma^2 = \text{RSS}/(n - 2)$ is linked to the χ^2 distribution with $n - 2$ degrees of freedom. Replacing the unknown σ with $\hat{\sigma}$ gives a standardised statistic with a **t distribution with $n - 2$ degrees of freedom**.*

$$\hat{\beta}_2 \sim N(\beta_2 ; \sigma^2/\Sigma x_i^2) \quad (3.1)$$

$$\hat{\beta}_1 \sim N(\beta_1 ; \sigma^2 \cdot \Sigma x_i^2 / (n \cdot \Sigma x_i^2)) \quad (3.2)$$

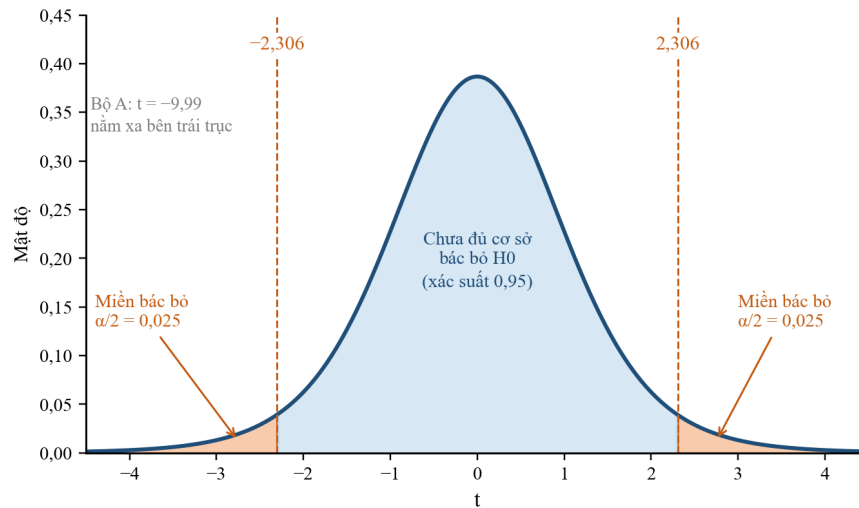
$$(n - 2) \cdot \hat{\sigma}^2 / \sigma^2 = \text{RSS} / \sigma^2 \sim \chi^2(n - 2) \quad (3.3)$$

$$t = (\hat{\beta}_j - \beta_j) / \text{se}(\hat{\beta}_j) \sim t(n - 2), \quad j = 1, 2 \quad (3.4)$$

Bậc tự do $n - 2$ vì hai tham số β_1, β_2 đã được ước lượng từ dữ liệu, tạo ra hai ràng buộc $\Sigma e_i = 0$ và $\Sigma e_i X_i = 0$ lên các phần dư. Với n nhỏ (bài thi thường $n = 5-15$), phân phối t có đuôi dày hơn phân phối chuẩn nên giá trị tới hạn t lớn hơn $z = 1,96$. Khi $n - 2 > 120$, t gần như trùng với z .

There are $n - 2$ degrees of freedom because β_1 and β_2 have been estimated, imposing the two constraints $\Sigma e_i = 0$ and $\Sigma e_i X_i = 0$ on the residuals. With small n (exam questions usually have $n = 5-15$) the t distribution has fatter tails than the normal, so the t critical value exceeds $z = 1.96$. When $n - 2 > 120$, t is almost identical to z .

Đại lượng <i>Quantity</i>	Công thức <i>Formula</i>	Bộ A <i>Data set A</i>
Σx^2	$\Sigma X_i^2 - n\bar{X}^2$	232,4 <i>232.4</i>
σ^2	$\text{RSS}/(n - 2)$	$83,146/8 = 10,3933$ <i>83.146/8 = 10.3933</i>
$\text{se}(\hat{\beta}_2)$	$\sqrt{(\sigma^2/\Sigma x^2)}$	$\sqrt{(10,3933/232,4)} = 0,2115$ <i>$\sqrt{(10.3933/232.4)} = 0.2115$</i>
$\text{se}(\hat{\beta}_1)$	$\sqrt{(\sigma^2 \cdot \Sigma X^2 / (n \cdot \Sigma x^2))}$	$\sqrt{(10,3933 \cdot 7.850 / (10 \cdot 232,4))} = 5,9251$ <i>$\sqrt{(10.3933 \cdot 7.850 / (10 \cdot 232.4))} = 5.9251$</i>
Bậc tự do <i>Degrees of freedom</i>	$n - 2$	8
t tới hạn hai phía 5% <i>Two-sided 5% critical t</i>	$t_{0,025}^{(8)}$ <i>$t_{0.025}^{(8)}$</i>	2,306 <i>2.306</i>
t tới hạn một phía 5% <i>One-sided 5% critical t</i>	$t_{0,05}^{(8)}$ <i>$t_{0.05}^{(8)}$</i>	1,860 <i>1.860</i>



Hình 3.1. Phân phối t với 8 bậc tự do và miền bác bỏ hai phía ở mức ý nghĩa 5%

Figure 3.1. The t distribution with 8 degrees of freedom and the two-sided 5% rejection region

LỖI HAY GẶP / COMMON MISTAKE — Dùng sai bậc tự do. Hồi quy hai biến dùng $n - 2$, không dùng n hay $n - 1$. Bộ A có $n = 10$: $t_{0,025}(8) = 2,306$; nếu tra nhầm 10 bậc tự do được 2,228, nếu dùng z được 1,96. Với n nhỏ, sai lệch đủ làm đổi kết luận ở những kiểm định sát ngưỡng.

Using the wrong degrees of freedom. Two-variable regression uses $n - 2$, not n or $n - 1$. Data set A has $n = 10$: $t_{0,025}(8) = 2.306$; looking up 10 df by mistake gives 2.228, and using z gives 1.96. With small n the gap can reverse borderline verdicts.

3.2. Khoảng tin cậy của hệ số hồi quy và của phương sai sai số — Confidence intervals for the coefficients and the disturbance variance

Từ (3.4), $P(-t(\alpha/2) \leq (\beta_j - \hat{\beta}_j)/\text{se}(\hat{\beta}_j) \leq t(\alpha/2)) = 1 - \alpha$. Biến đổi bất đẳng thức để β_j đứng giữa được khoảng tin cậy $1 - \alpha$. Khoảng là ngẫu nhiên (đổi theo mẫu), còn β_j là hằng số: “95%” nghĩa là nếu lấy nhiều mẫu thì khoảng 95% số khoảng lập ra chứa β_j .

From (3.4), $P(-t(\alpha/2) \leq (\beta_j - \hat{\beta}_j)/\text{se}(\hat{\beta}_j) \leq t(\alpha/2)) = 1 - \alpha$. Rearranging so that β_j sits in the middle gives the $1 - \alpha$ confidence interval. The interval is random (it changes with the sample) while β_j is fixed: “95%” means that in repeated sampling about 95% of the intervals built this way contain β_j .

$$\hat{\beta}_j - t_{\alpha/2}(n-2) \cdot \text{se}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2}(n-2) \cdot \text{se}(\hat{\beta}_j) \quad (3.5)$$

$$(n-2)\hat{\sigma}^2/\chi^2_{\alpha/2}(n-2) \leq \sigma^2 \leq (n-2)\hat{\sigma}^2/\chi^2_{1-\alpha/2}(n-2) \quad (3.6)$$

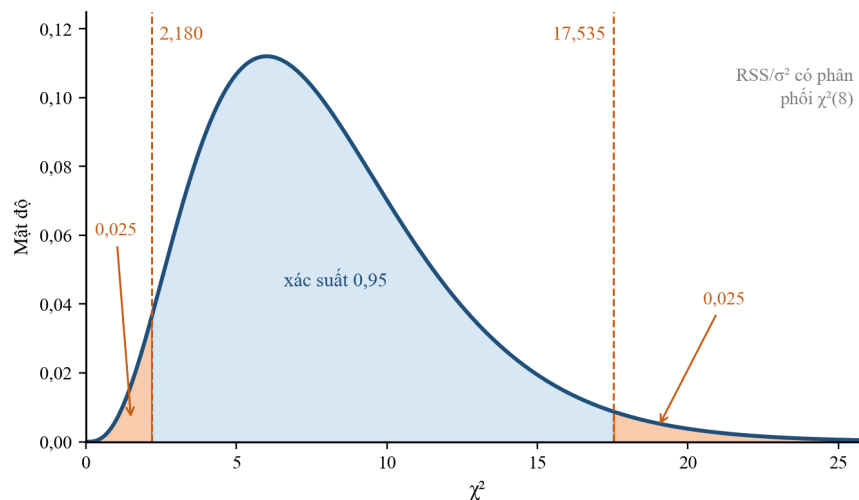
Trong (3.6), $\chi^2(\alpha/2)$ là giá trị có xác suất vượt quá bằng $\alpha/2$ (đuôi phải, số lớn) nên đứng ở mẫu cận dưới; $\chi^2(1 - \alpha/2)$ là số nhỏ, đứng ở mẫu cận trên. Phân phối χ^2 lệch phải nên khoảng tin cậy của σ^2 không đối xứng quanh $\hat{\sigma}^2$.

In (3.6), $\chi^2(\alpha/2)$ is the value exceeded with probability $\alpha/2$ (right tail, the large number), so it sits in the denominator of the lower limit; $\chi^2(1 - \alpha/2)$ is the small number, in the denominator of the upper limit. The χ^2 distribution is right-skewed, so the interval for σ^2 is not symmetric around $\hat{\sigma}^2$.

VÍ DỤ / EXAMPLE — Bộ A, độ tin cậy 95%: $t_{0,025}^{(8)} = 2,306$; $\chi^2_{0,025}^{(8)} = 17,535$; $\chi^2_{0,975}^{(8)} = 2,180$.

Data set A, 95% confidence: $t_{0,025}^{(8)} = 2.306$; $\chi^2_{0,025}^{(8)} = 17.535$; $\chi^2_{0,975}^{(8)} = 2.180$.

1. $\beta_2: -2,1127 \pm 2,306 \times 0,2115 = -2,1127 \pm 0,4877 \Rightarrow -2,6004 \leq \beta_2 \leq -1,6251$. Với độ tin cậy 95%, khi giá tăng 1 nghìn đồng/lốc, lượng bán trung bình giảm từ 1,625 đến 2,600 nghìn lốc/tháng.
 $\beta_2: -2.1127 \pm 2.306 \times 0.2115 = -2.1127 \pm 0.4877 \Rightarrow -2.6004 \leq \beta_2 \leq -1.6251$. With 95% confidence, a price rise of 1 thousand VND per pack lowers mean sales by between 1.625 and 2.600 thousand packs a month.
2. $\beta_1: 103,8115 \pm 2,306 \times 5,9251 = 103,8115 \pm 13,6632 \Rightarrow 90,148 \leq \beta_1 \leq 117,475$. Giá bằng 0 nằm xa khoảng dữ liệu (20–35 nghìn đồng) nên β_1 không có ý nghĩa kinh tế trực tiếp; khoảng này chủ yếu dùng để kiểm định.
 $\beta_1: 103.8115 \pm 2.306 \times 5.9251 = 103.8115 \pm 13.6632 \Rightarrow 90.148 \leq \beta_1 \leq 117.475$. A zero price lies far outside the data range (20–35 thousand VND), so β_1 has no direct economic meaning; the interval is used mainly for testing.
3. $\sigma^2: \text{RSS} = (n - 2)\sigma^2 = 83,146$; $83,146/17,535 \leq \sigma^2 \leq 83,146/2,180 \Rightarrow 4,742 \leq \sigma^2 \leq 38,141$. Ước lượng điểm $\sigma^2 = 10,393$ nằm lệch về phía cận dưới.
 $\sigma^2: \text{RSS} = (n - 2)\sigma^2 = 83.146$; $83.146/17.535 \leq \sigma^2 \leq 83.146/2.180 \Rightarrow 4.742 \leq \sigma^2 \leq 38.141$. The point estimate $\sigma^2 = 10.393$ lies towards the lower limit.



Hình 3.2. Phân phối $\chi^2(8)$ và hai giá trị tới hạn dùng cho khoảng tin cậy 95% của σ^2

Figure 3.2. The $\chi^2(8)$ distribution and the two critical values for the 95% confidence interval of σ^2

Độ rộng khoảng tin cậy của β_2 bằng $2 \cdot t \cdot \text{se}(\beta_2) = 2 \cdot t \cdot \hat{\sigma} / \sqrt{\Sigma x^2}$. Khoảng hẹp lại khi: $\hat{\sigma}$ nhỏ (mô hình khớp tốt), Σx^2 lớn (X trải rộng), n lớn (t nhỏ và Σx^2 tăng), hoặc chấp nhận độ tin cậy thấp hơn.

The width of the interval for β_2 is $2 \cdot t \cdot \text{se}(\beta_2) = 2 \cdot t \cdot \hat{\sigma} / \sqrt{\Sigma x^2}$. It narrows when $\hat{\sigma}$ is small (a good fit), Σx^2 is large (X is spread out), n is large (smaller t , larger Σx^2), or a lower confidence level is accepted.

LỖI HAY GẶP / COMMON MISTAKE — Đảo chỗ hai giá trị χ^2 trong (3.6), ra cận dưới lớn hơn cận trên; hoặc dùng σ (độ lệch chuẩn) thay cho $(n-2)\sigma^2$ ở tử số. Kiểm tra nhanh: cận dưới phải nhỏ hơn σ^2 , cận trên phải lớn hơn σ^2 .

Swapping the two χ^2 values in (3.6), so the lower limit exceeds the upper; or using σ (the standard deviation) instead of $(n-2)\sigma^2$ in the numerator. Quick check: the lower limit must be below σ^2 and the upper limit above it.

3.3. Kiểm định giả thuyết về hệ số hồi quy — Hypothesis tests on a regression coefficient

Kiểm định hai phía $H_0: \beta_2 = 0$ — Two-sided test of $H_0: \beta_2 = 0$

Câu hỏi “X có ảnh hưởng đến Y không?” được viết thành $H_0: \beta_2 = 0$ (X không ảnh hưởng) và $H_1: \beta_2 \neq 0$. Thống kê kiểm định lấy từ (3.4) với $\beta_2 = 0$ theo H_0 .

The question “does X affect Y?” is written as $H_0: \beta_2 = 0$ (no effect) against $H_1: \beta_2 \neq 0$. The test statistic comes from (3.4) with $\beta_2 = 0$ under H_0 .

$$t = (\hat{\beta}_2 - \beta^*) / \text{se}(\hat{\beta}_2); \quad \beta^* = 0 \Rightarrow t = \hat{\beta}_2 / \text{se}(\hat{\beta}_2) \quad (3.7)$$

$$\text{Bác bỏ } H_0 \text{ khi } |t| > t_{\alpha/2}(n-2); \quad p\text{-value} = 2 \cdot P(T > |t|) \quad (3.8)$$

VÍ DỤ / EXAMPLE — Bộ A, $\alpha = 5\%$. $H_0: \beta_2 = 0$; $H_1: \beta_2 \neq 0$. $t = -2,1127/0,2115 = -9,990$. $|t| = 9,990 > t_{0,025}(8) = 2,306$ nên bác bỏ H_0 . Kết luận: với mức ý nghĩa 5%, giá bán có ảnh hưởng đến lượng bán sữa chua uống của Mây Xanh. $p\text{-value} < 0,0001$ (EViews in 0.0000).

Data set A, $\alpha = 5\%$. $H_0: \beta_2 = 0$; $H_1: \beta_2 \neq 0$. $t = -2.1127/0.2115 = -9.990$. $|t| = 9.990 > t_{0.025}(8) = 2.306$, so reject H_0 . Conclusion: at the 5% level, price affects Mây Xanh’s sales of drinking yoghurt. $p\text{-value} < 0.0001$ (EViews prints 0.0000).

Kiểm định một phía — One-sided tests

Khi lý thuyết hoặc câu hỏi đã chỉ rõ hướng tác động, dùng kiểm định một phía. Câu hỏi “giảm giá có làm tăng lượng bán không?” tương đương “ β_2 có âm không?": $H_0: \beta_2 \geq 0$, $H_1: \beta_2 < 0$. Toàn bộ α đặt ở đuôi trái, giá trị tới hạn là t_α (không chia đôi). Giả thuyết cần chứng minh luôn đặt ở H_1 , dấu bằng luôn nằm ở H_0 .

When theory or the question fixes the direction of the effect, use a one-sided test. “Does a price cut raise sales?” is the same as “is β_2 negative?": $H_0: \beta_2 \geq 0$, $H_1: \beta_2 < 0$. All of α goes in the left tail and the critical value is t_α (not halved). The claim to be shown always goes in H_1 ; the equality sign always stays in H_0 .

Câu hỏi <i>Question</i>	H0	H1	Bác bỏ H0 khi <i>Reject H0 if</i>
X có ảnh hưởng đến Y? <i>Does X affect Y?</i>	$\beta_2 = \beta^*$	$\beta_2 \neq \beta^*$	$ t > t(\alpha/2)$
X tác động cùng chiều? <i>Positive effect?</i>	$\beta_2 \leq \beta^*$	$\beta_2 > \beta^*$	$t > t_\alpha$
X tác động ngược chiều? <i>Negative effect?</i>	$\beta_2 \geq \beta^*$	$\beta_2 < \beta^*$	$t < -t_\alpha$

VÍ DỤ / EXAMPLE — Bộ A, $\alpha = 5\%$: $H_0: \beta_2 \geq 0$; $H_1: \beta_2 < 0$. $t = -9,990 < -t_{0,05^{(8)}} = -1,860$ nên bác bỏ H_0 : giảm giá làm tăng lượng bán. p-value một phía = $P(T_8 < -9,99) < 0,0001$, bằng một nửa Prob. hai phía mà EViews in, vì t mang dấu đúng với H_1 .
Data set A, $\alpha = 5\%$: $H_0: \beta_2 \geq 0$; $H_1: \beta_2 < 0$. $t = -9.990 < -t_{0.05^{(8)}} = -1.860$, so reject H_0 : a price cut raises sales. One-sided p-value = $P(T_8 < -9.99) < 0.0001$, half the two-sided Prob. printed by EViews, because t has the sign required by H_1 .

LỖI HAY GẶP / COMMON MISTAKE — Chia đôi Prob. của EViews mà không xem dấu. Prob. là p-value **hai phía**. Với $H_1: \beta_2 < 0$, chỉ được lấy Prob./2 khi $t < 0$; nếu $t > 0$ thì p-value một phía là $1 - \text{Prob.}/2$ và chắc chắn chưa đủ cơ sở bác bỏ H_0 . Lỗi ngược lại: kiểm định một phía mà vẫn tra $t(\alpha/2)$.
Halving the EViews Prob. without checking the sign. Prob. is a two-sided p-value. With $H_1: \beta_2 < 0$, Prob./2 applies only when $t < 0$; if $t > 0$, the one-sided p-value is $1 - \text{Prob.}/2$ and H_0 is certainly not rejected. The reverse mistake: a one-sided test looked up with $t(\alpha/2)$.

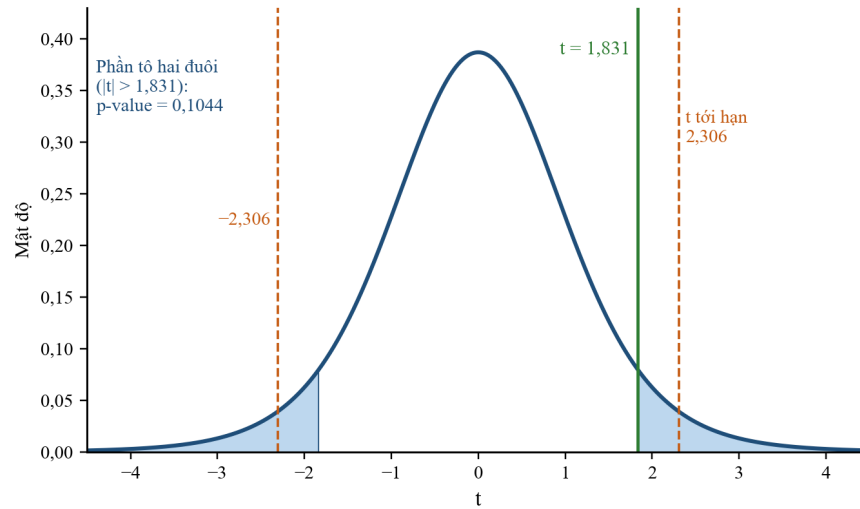
Kiểm định $H_0: \beta_2 = \beta^*$ với β^* khác 0 — Testing $H_0: \beta_2 = \beta^*$ with β^* other than 0

Đề thi hay cho một ý kiến về độ lớn tác động, ví dụ phòng kinh doanh của Mây Xanh cho rằng “tăng giá 1 nghìn đồng/lốc thì lượng bán giảm 2,5 nghìn lốc/tháng”. Ý kiến này là $H_0: \beta_2 = -2,5$; $H_1: \beta_2 \neq -2,5$. Tử số của t là $\beta_2 - \beta^*$, không phải β_2 .

Exam questions often state a claim about the size of an effect; for example Mây Xanh's sales department says “a price rise of 1 thousand VND per pack cuts sales by 2.5 thousand packs a month”. The claim is $H_0: \beta_2 = -2.5$ against $H_1: \beta_2 \neq -2.5$. The numerator of t is $\beta_2 - \beta^$, not β_2 .*

- Giả thuyết: $H_0: \beta_2 = -2,5$; $H_1: \beta_2 \neq -2,5$; $\alpha = 5\%$.
Hypotheses: $H_0: \beta_2 = -2.5$; $H_1: \beta_2 \neq -2.5$; $\alpha = 5\%$.
- Thống kê: $t = (-2,1127 - (-2,5))/0,2115 = 0,3873/0,2115 = 1,831$.
Statistic: $t = (-2.1127 - (-2.5))/0.2115 = 0.3873/0.2115 = 1.831$.
- So sánh: $|t| = 1,831 < 2,306$; p-value = $0,1044 > 0,05$; -2,5 nằm trong khoảng tin cậy $(-2,6004; -1,6251)$.
Compare: $|t| = 1.831 < 2.306$; p-value = $0.1044 > 0.05$; -2.5 lies inside the confidence interval $(-2.6004; -1.6251)$.
- Kết luận: chưa đủ cơ sở bác bỏ H_0 . Số liệu 10 tháng không mâu thuẫn với ý kiến của phòng kinh doanh. Câu này không chứng minh β_2 đúng bằng -2,5: mọi giá trị trong khoảng tin cậy đều không bị bác bỏ.

Conclusion: insufficient evidence to reject H_0 . The 10 months of data do not contradict the sales department's claim. This does not prove β_2 equals -2.5 : every value inside the confidence interval is equally not rejected.



Hình 3.3. p-value của kiểm định $H_0: \beta_2 = -2,5$ trên bộ A: phần tô là hai đuôi vượt quá $|t|$
Figure 3.3. The p-value of the test of $H_0: \beta_2 = -2.5$ on data set A: the shaded area is the two tails beyond $|t|$

Ba cách kết luận một kiểm định — Three ways to conclude a test

Cách Approach	Quy tắc bác bỏ H_0 Reject H_0 when	Ưu điểm Advantage	Bộ A, $H_0: \beta_2 = -2,5$ Data set A, $H_0: \beta_2 = -2.5$
Miền bác bỏ (giá trị tới hạn) Rejection region (critical value)	$ t > t(\alpha/2)$	Đúng dạng đáp án đề thi tính tay Matches hand-worked exam answers	$ 1,831 < 2,306$ $ 1.831 < 2.306$
Khoảng tin cậy Confidence interval	β^* nằm ngoài khoảng β^* lies outside the interval	Cho biết mọi β^* không bị bác bỏ Shows every β^* not rejected	$-2,5 \in (-2,600; -1,625)$ $-2.5 \in (-2.600; -1.625)$
p-value	p-value $< \alpha$	Dùng được ngay với bảng phần mềm, mọi α Read straight from software, any α	$0,1044 > 0,05$ $0.1044 > 0.05$

Ba cách luôn cho cùng kết luận nếu dùng cùng α , cùng loại kiểm định (hai phía) và cùng bậc tự do. p-value là mức ý nghĩa nhỏ nhất mà tại đó H_0 bị bác bỏ; p-value không phải xác suất H_0 đúng.

The three always agree when they use the same α , the same (two-sided) test and the same degrees of freedom. The p-value is the smallest significance level at which H_0 is rejected; it is not the probability that H_0 is true.

LỖI HAY GẶP / COMMON MISTAKE — Viết “chấp nhận H_0 ” hay “ $\beta_2 = -2,5$ là đúng”. Kiểm định chỉ cho biết số liệu có đủ bằng chứng chống lại H_0 hay không. Viết đúng: “chưa đủ cơ sở bác bỏ H_0 ” và diễn giải bằng lời của đề: “số liệu không mâu thuẫn với ý kiến...”.
Writing “accept H_0 ” or “ $\beta_2 = -2.5$ is true”. A test only says whether the data give enough evidence against H_0 . Correct wording: “insufficient evidence to reject H_0 ”, followed by the question’s own terms: “the data do not contradict the claim...”.

3.4. Phân tích phương sai và kiểm định F — Analysis of variance and the F-test

Tổng bình phương toàn phần $TSS = ESS + RSS$ (Chương 2). Mỗi tổng chia cho bậc tự do của nó được trung bình bình phương; tỷ số hai trung bình bình phương là thống kê F kiểm định $H_0: \beta_2 = 0$ (mô hình không giải thích được Y).

The total sum of squares splits as $TSS = ESS + RSS$ (Chapter 2). Dividing each sum by its degrees of freedom gives a mean square; the ratio of the two mean squares is the F statistic for $H_0: \beta_2 = 0$ (the model explains nothing).

$$F = (ESS/1)/(RSS/(n-2)) = R^2(n-2)/(1-R^2) \sim F(1; n-2) \quad (3.9)$$

$$\text{Hồi quy hai biến: } F = t^2, \quad F_{\alpha}(1; n-2) = (t_{\alpha/2}^{(n-2)})^2 \quad (3.10)$$

Nguồn biến động <i>Source</i>	Tổng bình phương <i>Sum of squares</i>	Bậc tự do <i>df</i>	Trung bình bình phương <i>Mean square</i>	F
Hồi quy (ESS) <i>Regression (ESS)</i>	$\beta_2^2 \Sigma x^2 = 1.037,354$ $\beta_2^2 \Sigma x^2 = 1,037.354$	1	1.037,354 $1,037.354$	$1.037,354/10,393 = 99,810$ $1,037.354/10.393 = 99.810$
Phần dư (RSS) <i>Residual (RSS)</i>	83,146 83.146	8	$\sigma^2 = 10,393$ $\sigma^2 = 10.393$	
Toàn phần (TSS) <i>Total (TSS)</i>	$\Sigma y^2 = 1.120,5$ $\Sigma y^2 = 1,120.5$	9		

VÍ DỤ / EXAMPLE — Bộ A: $F = 99,810$; $F_{0,05}(1; 8) = 5,318$. $F > 5,318$ nên bác bỏ H_0 : mô hình phù hợp. Kiểm tra quan hệ (3.10): $t^2 = (-9,9905)^2 = 99,810 = F$; $2,306^2 = 5,318$, khớp $F_{0,05}(1; 8)$. Theo R^2 : $F = 0,9258 \times 8/(1 - 0,9258) = 99,81$.
Data set A: $F = 99.810$; $F_{0.05}(1; 8) = 5.318$. $F > 5.318$, so reject H_0 : the model is significant. Check (3.10): $t^2 = (-9.9905)^2 = 99.810 = F$; $2.306^2 = 5.318$, matching $F_{0.05}(1; 8)$. From R^2 : $F = 0.9258 \times 8/(1 - 0.9258) = 99.81$.

Ở hồi quy hai biến, kiểm định F sự phù hợp của mô hình và kiểm định t hai phía $H_0: \beta_2 = 0$ là một: cùng p-value, cùng kết luận. F không dùng được cho kiểm định một phía và không dùng cho $H_0: \beta_2 = \beta^* \neq 0$ dưới dạng (3.9). Ở hồi quy bội (Chương 4) hai kiểm định tách rời nhau.

In two-variable regression the overall F-test and the two-sided t-test of $H_0: \beta_2 = 0$ are the same test: same p-value, same verdict. F cannot handle a one-sided alternative, and form (3.9) does not test $H_0: \beta_2 = \beta^ \neq 0$. In multiple regression (Chapter 4) the two tests separate.*

3.5. Dự báo giá trị trung bình và giá trị cá biệt — *Forecasting the mean and an individual value*

Cho $X = X_0$, ước lượng điểm của cả $E(Y | X_0)$ lẫn Y_0 đều là $\hat{Y}_0 = \beta_1 + \beta_2 X_0$. Khác nhau ở sai số: dự báo trung bình chỉ chịu sai số ước lượng đường hồi quy; dự báo cá biệt chịu thêm sai số ngẫu nhiên u_0 của quan sát mới.

Given $X = X_0$, the point forecast of both $E(Y | X_0)$ and Y_0 is $\hat{Y}_0 = \beta_1 + \beta_2 X_0$. They differ in error: the mean forecast carries only the error of the estimated line; the individual forecast also carries the disturbance u_0 of the new observation.

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 \quad (3.11)$$

$$\text{Var}(\hat{Y}_0) = \sigma^2 [1/n + (X_0 - \bar{X})^2 / \sum x_i^2] \quad (3.12)$$

$$\text{Var}(Y_0 - \hat{Y}_0) = \sigma^2 [1 + 1/n + (X_0 - \bar{X})^2 / \sum x_i^2] \quad (3.13)$$

$$\hat{Y}_0 \pm t_{\alpha/2}^{(n-2)} \cdot se, \quad se = \sqrt{\text{Var}} \text{ (thay } \sigma^2 \text{ bằng } \hat{\sigma}^2) \quad (3.14)$$

VÍ DỤ / EXAMPLE — Máy Xanh định giá $X_0 = 30$ nghìn đồng/lốc. Dự báo lượng bán với độ tin cậy 95%.

Máy Xanh sets a price of $X_0 = 30$ thousand VND per pack. Forecast sales with 95% confidence.

- $\hat{Y}_0 = 103,8115 - 2,1127 \times 30 = 40,429$ nghìn lốc.
 $\hat{Y}_0 = 103.8115 - 2.1127 \times 30 = 40.429$ thousand packs.
- $\text{Var}(\hat{Y}_0) = 10,3933 \times [1/10 + (30 - 27,6)^2 / 232,4] = 1,2969$; $se = 1,1388$. Khoảng trung bình: $40,429 \pm 2,306 \times 1,1388 = (37,803; 43,056)$.
 $\text{Var}(\hat{Y}_0) = 10.3933 \times [1/10 + (30 - 27.6)^2 / 232.4] = 1.2969$; $se = 1.1388$. Mean interval: $40.429 \pm 2.306 \times 1.1388 = (37.803; 43.056)$.
- $\text{Var}(Y_0 - \hat{Y}_0) = 10,3933 + 1,2969 = 11,6902$; $se = 3,4191$. Khoảng cá biệt: $40,429 \pm 2,306 \times 3,4191 = (32,545; 48,314)$.
 $\text{Var}(Y_0 - \hat{Y}_0) = 10.3933 + 1.2969 = 11.6902$; $se = 3.4191$. Individual interval: $40.429 \pm 2.306 \times 3.4191 = (32.545; 48.314)$.
- Diễn giải: ở giá 30 nghìn đồng, lượng bán trung bình của các tháng nằm trong (37,80; 43,06) nghìn lốc; lượng bán của một tháng cụ thể nằm trong (32,54; 48,31) nghìn lốc, với độ tin cậy 95%.

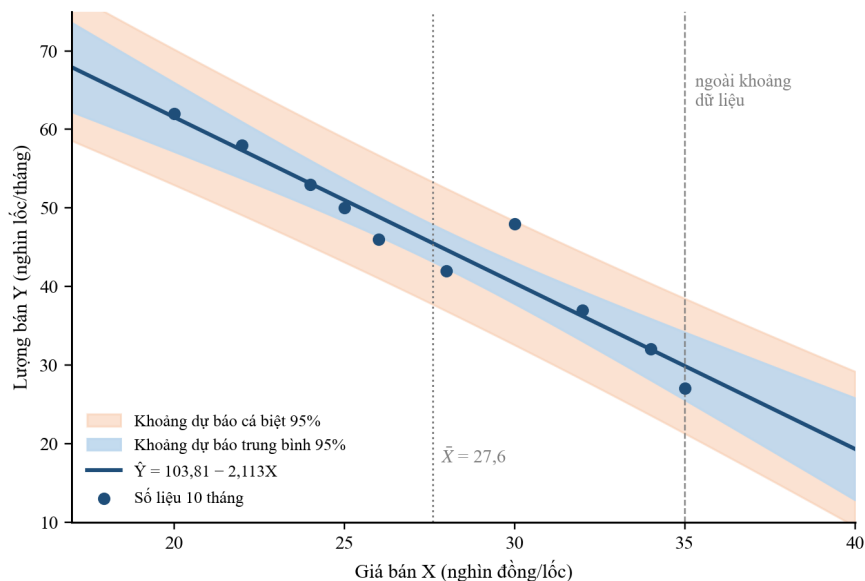
Interpretation: at a price of 30 thousand VND, mean monthly sales lie in (37.80; 43.06) thousand packs, and the sales of one particular month lie in (32.54; 48.31) thousand packs, with 95% confidence.

X_0	$\hat{Y}_0$	se trung bình <i>se of mean</i>	Độ rộng KTC trung bình <i>Mean interval width</i>	se cá biệt <i>se of individual</i>	Độ rộng KTC cá biệt <i>Individual interval width</i>
20,0 20.0	61,56 61.56	1,903 1.903	8,78 8.78	3,744 3.744	17,27 17.27

X_0	$\hat{Y}_0$	se trung bình <i>se of mean</i>	Độ rộng KTC trung bình <i>Mean interval width</i>	se cá biệt <i>se of individual</i>	Độ rộng KTC cá biệt <i>Individual interval width</i>
27,6 <i>27.6</i>	45,50 <i>45.50</i>	1,019 <i>1.019</i>	4,70 <i>4.70</i>	3,381 <i>3.381</i>	15,59 <i>15.59</i>
30,0 <i>30.0</i>	40,43 <i>40.43</i>	1,139 <i>1.139</i>	5,25 <i>5.25</i>	3,419 <i>3.419</i>	15,77 <i>15.77</i>
35,0 <i>35.0</i>	29,87 <i>29.87</i>	1,868 <i>1.868</i>	8,61 <i>8.61</i>	3,726 <i>3.726</i>	17,18 <i>17.18</i>
40,0 <i>40.0</i>	19,30 <i>19.30</i>	2,813 <i>2.813</i>	12,98 <i>12.98</i>	4,279 <i>4.279</i>	19,73 <i>19.73</i>

Bảng trên: khoảng hẹp nhất tại $X_0 = \bar{X} = 27,6$ và loe dần khi X_0 ra xa. $X_0 = 40$ nằm ngoài khoảng giá đã quan sát (20–35): công thức vẫn tính được nhưng giả định quan hệ tuyến tính giữ nguyên ngoài khoảng dữ liệu chưa được kiểm chứng.

Table above: the intervals are narrowest at $X_0 = \bar{X} = 27.6$ and widen as X_0 moves away. $X_0 = 40$ lies outside the observed price range (20–35): the formula still works, but the assumption that the linear relation holds beyond the data is untested.



Hình 3.4. Dải dự báo trung bình và cá biệt 95% của bộ A: hẹp nhất tại $\bar{X}$, loe ra hai phía
Figure 3.4. 95% bands for the mean and individual forecasts of data set A: narrowest at $\bar{X}$, flaring on both sides

LỖI HAY GẶP / COMMON MISTAKE — Đề hỏi “dự báo lượng bán của một tháng” (cá biệt) mà dùng công thức (3.12) của dự báo trung bình, hoặc quên bình phương $(X_0 - \bar{X})$. Đọc kỹ đề: “trung bình”, “bình quân các tháng” dùng (3.12); “một tháng”, “một cửa hàng cụ thể” dùng (3.13).

The question asks for “sales in one month” (individual) but the mean formula (3.12) is used, or $(X_0 -$

$\bar{X}$) is not squared. Read carefully: “mean”, “average over months” uses (3.12); “one month”, “a particular shop” uses (3.13).

3.6. Đổi đơn vị đo của X và Y — Changing the units of X and Y

Đề thi hay hỏi “viết lại hàm hồi quy khi Y tính bằng ...”. Đổi đơn vị là nhân biến với hằng số: $Y^* = w_1 Y$, $X^* = w_2 X$. Không cần ước lượng lại, chỉ cần quy tắc dưới đây.

Exam questions often ask to “rewrite the regression when Y is measured in ...”. A change of units multiplies a variable by a constant: $Y^* = w_1 Y$, $X^* = w_2 X$. No re-estimation is needed, only the rules below.

$$Y^* = w_1 Y, X^* = w_2 X \Rightarrow \hat{\beta}_2^* = (w_1/w_2)\hat{\beta}_2, \hat{\beta}_1^* = w_1\hat{\beta}_1, \hat{\sigma}^* = w_1\hat{\sigma} \quad (3.15)$$

Trường hợp Case	w_1	w_2	β_1	β_2	$se(\beta_2)$	$t(\beta_2)$	R^2	σ
Gốc: Y nghìn lốc, X nghìn đồng <i>Original: Y thousand packs, X thousand VND</i>	1	1	103,81 <i>103.81</i>	-2,1127 <i>-2.1127</i>	0,2115 <i>0.2115</i>	-9,99 0 <i>-9.990</i>	0,925 8 <i>0.9258</i>	3,224 <i>3.224</i>
Y tính bằng lốc <i>Y in packs</i>	1.00 0 <i>1,00</i> 0	1	103.811,5 3 <i>103,811.5</i> 3	-2.112,7 367 <i>-2,112.73</i> 67	211,4747 <i>211.4747</i>	-9,99 0 <i>-9.990</i>	0,925 8 <i>0.9258</i>	3.223,9 <i>3,223.9</i>
X tính bằng đồng <i>X in VND</i>	1	1.00 0 <i>1,00</i> 0	103,81 <i>103.81</i>	-0,00211 27 <i>-0.002112</i> 7	0,000211 5 <i>0.000211</i> 5	-9,99 0 <i>-9.990</i>	0,925 8 <i>0.9258</i>	3,224 <i>3.224</i>
Y bằng lốc, X bằng đồng <i>Y in packs, X in VND</i>	1.00 0 <i>1,00</i> 0	1.00 0 <i>1,00</i> 0	103.811,5 3 <i>103,811.5</i> 3	-2,11273 67 <i>-2.112736</i> 7	0,211474 7 <i>0.211474</i> 7	-9,99 0 <i>-9.990</i>	0,925 8 <i>0.9258</i>	3.223,9 <i>3,223.9</i>

Quy luật đọc từ bảng: (i) thống kê t , p -value và R^2 không đổi trong mọi trường hợp, nên kết luận kiểm định không phụ thuộc đơn vị; (ii) đổi đơn vị Y nhân β_1 , β_2 , se và σ với w_1 ; (iii) đổi đơn vị X chia β_2 và $se(\beta_2)$ cho w_2 , β_1 giữ nguyên. Ví dụ: Y tính bằng lốc thì $Y^* = 103.811,5 - 2.112,7X$: giá tăng 1 nghìn đồng, lượng bán giảm 2.112,7 lốc, cùng nội dung kinh tế với $\beta_2 = -2,1127$ nghìn lốc.

Rules read off the table: (i) the t statistic, p -value and R^2 never change, so test verdicts do not depend on units; (ii) rescaling Y multiplies β_1 , β_2 , the standard errors and σ by w_1 ; (iii) rescaling X divides β_2 and $se(\beta_2)$ by w_2 and leaves β_1 unchanged. Example: with Y in packs, $Y^* = 103,811.5 - 2,112.7X$: a price rise of 1 thousand VND cuts sales by 2,112.7 packs, the same economic content as $\beta_2 = -2.1127$ thousand packs.

LỖI HAY GẶP / COMMON MISTAKE — Đổi đơn vị X mà nhân cả β_1 với w_2 , hoặc chia β_2 theo chiều ngược (X tính bằng đồng thì β_2 phải NHỎ đi 1.000 lần vì 1 đồng tác động ít hơn 1 nghìn đồng). Kiểm bằng ví dụ một điểm: thế $\bar{X}$ vào hàm mới phải ra $\bar{Y}$ đổi đơn vị.
Rescaling X and also multiplying β_1 by w_2 , or dividing β_2 the wrong way (with X in VND, β_2 must become 1,000 times SMALLER, since 1 VND has less effect than 1 thousand VND). Check with one point: substituting $\bar{X}$ into the new function must return $\bar{Y}$ in the new units.

3.7. Các dạng hàm hồi quy — Functional forms

“Tuyến tính” trong mô hình hồi quy tuyến tính là tuyến tính theo tham số. Lấy log hay nghịch đảo của biến vẫn ước lượng bằng OLS như thường, chỉ thay đổi cách đọc hệ số. Bảng dưới đây ước lượng cả năm dạng hàm trên bộ A.

“Linear” in the linear regression model means linear in the parameters. Taking logs or reciprocals of the variables still allows ordinary OLS; only the reading of the coefficients changes. The table below fits all five forms to data set A.

$$\text{Co giãn của mô hình tuyến tính tại trung bình: } E = \beta_2 \cdot \bar{X} / \bar{Y} \quad (3.16)$$

$$\text{Log-log: } \ln Y = \beta_1 + \beta_2 \ln X + u, \quad \beta_2 = (\Delta Y / Y) / (\Delta X / X) \quad (3.17)$$

$$\text{Log-lin: } \ln Y_t = \beta_1 + \beta_2 t + u, \quad r = e^{\beta_2} - 1 \quad (3.18)$$

$$\text{Lin-log: } Y = \beta_1 + \beta_2 \ln X + u, \quad X \text{ tăng } 1\% \Rightarrow \Delta Y \approx \beta_2 / 100 \quad (3.19)$$

$$\text{Nghịch đảo: } Y = \beta_1 + \beta_2 (1/X) + u, \quad X \text{ tăng mãi} \Rightarrow E(Y) \text{ tiến về } \beta_1 \quad (3.20)$$

Dạng hàm Form	β_2 (se)	R^2	Đọc β_2 trên bộ A Reading β_2 on data set A
Tuyến tính $Y - X$ Linear $Y - X$	-2,1127 (0,2115) -2.1127 (0.2115)	0,9258 0.9258	Giá tăng 1 nghìn đồng, lượng bán giảm 2,113 nghìn gói; co giãn tại trung bình -1,282 A price rise of 1 thousand VND cuts sales by 2.113 thousand packs; elasticity at the mean -1.282
Log-log $\ln Y - \ln X$	-1,3201 (0,1763) -1.3201 (0.1763)	0,8752 0.8752	Giá tăng 1%, lượng bán giảm 1,320% (co giãn không đổi) A 1% price rise cuts sales by 1.320% (constant elasticity)
Log-lin $\ln Y - X$	-0,0493 (0,0058) -0.0493 (0.0058)	0,9007 0.9007	Giá tăng 1 nghìn đồng, lượng bán giảm khoảng 4,93% (chính xác 4,81%) A price rise of 1 thousand VND cuts sales by about 4.93% (exactly 4.81%)
Lin-log $Y - \ln X$	-57,158 (6,070) -57.158 (6.070)	0,9173 0.9173	Giá tăng 1%, lượng bán giảm 0,5716 nghìn gói A 1% price rise cuts sales by 0.5716 thousand packs

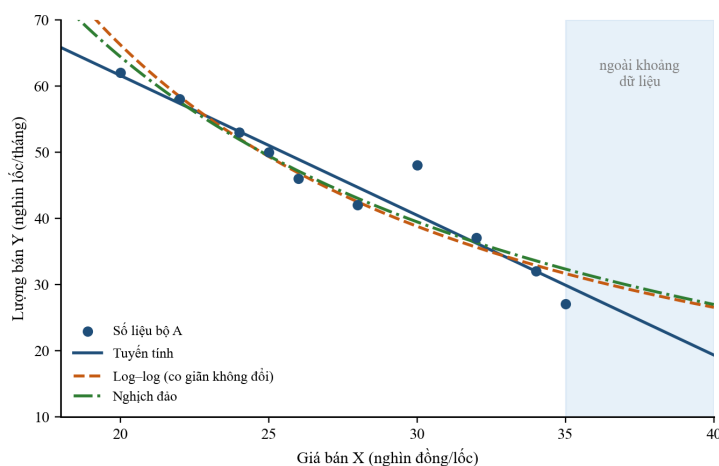
Dạng hàm Form	β_2 (se)	R ²	Đọc β_2 trên bộ A Reading β_2 on data set A
Nghịch đảo Y – 1/X <i>Reciprocal Y – 1/X</i>	1.497,70 (178,26) <i>1,497.70 (178.26)</i>	0,8982 <i>0.8982</i>	Lượng bán giảm dần về $\beta_1 = -10,49$ khi giá tăng; β_1 âm nên dạng này không hợp lý ở giá cao <i>Sales fall towards $\beta_1 = -10.49$ as price rises; a negative β_1 makes the form implausible at high prices</i>

Hệ số co giãn của mô hình log–log là $-1,320$: cầu sữa chua uống của Mây Xanh co giãn theo giá ($|E| > 1$), nên giảm giá làm tổng doanh thu tăng. Kiểm định $H_0: \beta_2 = -1$ (co giãn đơn vị): $t = (-1,3201 + 1)/0,1763 = -1,816$, $p\text{-value} = 0,1069 > 0,05$, chưa đủ cơ sở bác bỏ H_0 ; với 10 quan sát, số liệu chưa phân biệt được co giãn $-1,32$ với co giãn đơn vị. Ở mô hình tuyến tính, co giãn thay đổi dọc đường hồi quy; tại trung bình $E = -2,1127 \times 27,6/45,5 = -1,282$.

The elasticity in the log–log model is -1.320 : demand for Mây Xanh's drinking yoghurt is price-elastic ($|E| > 1$), so a price cut raises total revenue. Test $H_0: \beta_2 = -1$ (unit elasticity): $t = (-1.3201 + 1)/0.1763 = -1.816$, $p\text{-value} = 0.1069 > 0.05$, so H_0 is not rejected; with 10 observations the data cannot tell an elasticity of -1.32 from unit elasticity. In the linear model the elasticity changes along the line; at the mean $E = -2.1127 \times 27.6/45.5 = -1.282$.

VÍ DỤ / EXAMPLE — Tốc độ tăng trưởng bằng mô hình log–lin trên bộ E (Công ty CP Điện máy Sao Bắc, doanh số 40 quý 2016Q1–2025Q4). Đặt $t = 1, \dots, 40$; hồi quy $\ln Y$ theo t được $\ln \hat{Y} = 3,1158 + 0,01425t$ (se = 0,00118). Tốc độ tăng gộp mỗi quý: $\exp(0,01425) - 1 = 1,435\%$; quy ra năm: $(1 + 0,01435)^4 - 1 = 5,87\%$ /năm.

Growth rate from a log–lin model on data set E (Sao Bắc Electronics JSC, sales over 40 quarters 2016Q1–2025Q4). Let $t = 1, \dots, 40$; regressing $\ln Y$ on t gives $\ln \hat{Y} = 3.1158 + 0.01425t$ (se = 0.00118). Compound growth per quarter: $\exp(0.01425) - 1 = 1.435\%$; per year: $(1 + 0.01435)^4 - 1 = 5.87\%$ a year.



Hình 3.5. Bộ A với ba dạng hàm: tuyến tính, log–log và nghịch đảo khác nhau rõ nhất ngoài khoảng dữ liệu

Figure 3.5. Data set A with three functional forms: linear, log–log and reciprocal differ most outside the data range

LỖI HAY GẶP / COMMON MISTAKE — Diễn giải hệ số log sai. Log–log: “X tăng 1% thì Y thay đổi $\beta_2\%$ ”, không phải “tăng 1 đơn vị”. Log–lin: “X tăng 1 đơn vị thì Y thay đổi $100 \cdot \beta_2\%$ ”, không phải $\beta_2\%$. Lin–log: “X tăng 1% thì Y thay đổi $\beta_2/100$ đơn vị”. Một lỗi khác: so R^2 của mô hình Y với R^2 của mô hình $\ln Y$ để chọn dạng hàm.

Misreading log coefficients. Log–log: “a 1% rise in X changes Y by $\beta_2\%$ ”, not “a one-unit rise”. Log–lin: “a one-unit rise in X changes Y by $100 \cdot \beta_2\%$ ”, not $\beta_2\%$. Lin–log: “a 1% rise in X changes Y by $\beta_2/100$ units”. Another error: comparing the R^2 of a Y model with that of an $\ln Y$ model to pick a form.

3.8. Hồi quy qua gốc tọa độ — Regression through the origin

Khi lý thuyết khẳng định $X = 0$ thì $Y = 0$ (chi phí nhiên liệu khi xe không chạy, mô hình CAPM với lợi suất vượt trội), có thể bỏ hệ số chặn: $Y_i = \beta_2 X_i + u_i$. Công thức thay đổi: dùng tổng bình phương thô (không trừ trung bình), bậc tự do $n - 1$, và R^2 thông thường không còn dùng được.

When theory says that $X = 0$ implies $Y = 0$ (fuel cost of an idle truck, the CAPM in excess returns), the intercept can be dropped: $Y_i = \beta_2 X_i + u_i$. The formulas change: raw sums of squares (no mean subtracted), $n - 1$ degrees of freedom, and the usual R^2 no longer applies.

$$\hat{\beta}_2 = \Sigma X_i Y_i / \Sigma X_i^2, \quad \hat{\sigma}^2 = \Sigma e_i^2 / (n - 1), \quad \text{Var}(\hat{\beta}_2) = \sigma^2 / \Sigma X_i^2 \quad (3.21)$$

$$R^2_{\text{thô}} = (\Sigma X_i Y_i)^2 / (\Sigma X_i^2 \cdot \Sigma Y_i^2) \quad (3.22)$$

VÍ DỤ / EXAMPLE — Công ty vận tải Cỏ Lau Vàng theo dõi 8 xe tải trong tháng 9/2026 (dữ kiện đề bài): X là quãng đường (nghìn km/tháng) = 3; 4; 5; 6; 7; 8; 9; 10; Y là chi phí nhiên liệu (triệu đồng/tháng) = 7,8; 10,1; 12,4; 15,3; 17,2; 20,5; 22,4; 25,3.

Cỏ Lau Vàng Transport tracks 8 trucks in September 2026 (given data): X is distance (thousand km a month) = 3; 4; 5; 6; 7; 8; 9; 10; Y is fuel cost (million VND a month) = 7.8; 10.1; 12.4; 15.3; 17.2; 20.5; 22.4; 25.3.

- Ước lượng có hệ số chặn: $\hat{Y} = 0,1095 + 2,5024X$; $\text{se}(\hat{\beta}_1) = 0,3115$, $t = 0,352$, $p\text{-value} = 0,7372$. $|t| < t_{0,025}^{(6)} = 2,447$: chưa đủ cơ sở bác bỏ $H_0: \beta_1 = 0$, phù hợp lý thuyết.
With an intercept: $\hat{Y} = 0.1095 + 2.5024X$; $\text{se}(\hat{\beta}_1) = 0.3115$, $t = 0.352$, $p\text{-value} = 0.7372$. $|t| < t_{0,025}^{(6)} = 2.447$: $H_0: \beta_1 = 0$ is not rejected, as theory predicts.
- Qua gốc tọa độ: $\Sigma XY = 956,6$; $\Sigma X^2 = 380$; $\hat{\beta}_2 = 956,6/380 = 2,5174$ triệu đồng/nghìn km (khoảng 2.517 đồng/km).
Through the origin: $\Sigma XY = 956.6$; $\Sigma X^2 = 380$; $\hat{\beta}_2 = 956.6/380 = 2.5174$ million VND per thousand km (about 2,517 VND per km).
- $\text{RSS} = 0,5254$; $\hat{\sigma}^2 = 0,5254/7 = 0,07505$; $\text{se}(\hat{\beta}_2) = \sqrt{(0,07505/380)} = 0,01405$; khoảng tin cậy 95% dùng $t_{0,025}^{(7)} = 2,365$: (2,4841; 2,5506).
 $\text{RSS} = 0.5254$; $\hat{\sigma}^2 = 0.5254/7 = 0.07505$; $\text{se}(\hat{\beta}_2) = \sqrt{(0.07505/380)} = 0.01405$; the 95% interval uses $t_{0,025}^{(7)} = 2.365$: (2.4841; 2.5506).
- $R^2_{\text{thô}} = 956,6^2 / (380 \times 2.408,64) = 0,9998$. Con số gần 1 này không so được với $R^2 = 0,9980$ của mô hình có hệ số chặn vì mẫu số khác nhau (ΣY^2 thay cho Σy^2).
Raw $R^2 = 956.6^2 / (380 \times 2,408.64) = 0.9998$. This near-1 figure cannot be compared with $R^2 = 0.9980$ of the model with an intercept because the denominators differ (ΣY^2 instead of Σy^2).

Nội dung <i>Item</i>	Có hệ số chặn <i>With intercept</i>	Qua gốc tọa độ <i>Through the origin</i>
β_2	$\Sigma x_i y_i / \Sigma x_i^2$ (độ lệch) $\Sigma x_i y_i / \Sigma x_i^2$ (deviations)	$\Sigma X_i Y_i / \Sigma X_i^2$ (số thô) $\Sigma X_i Y_i / \Sigma X_i^2$ (raw)
σ^2	$RSS / (n - 2)$	$RSS / (n - 1)$
Σe_i	$= 0$	Nói chung $\neq 0$ <i>Generally $\neq 0$</i>
R^2	$1 - RSS/TSS$, nằm trong $[0; 1]$ <i>$1 - RSS/TSS$, within $[0; 1]$</i>	R^2 thông thường có thể âm; dùng R^2 thô (3.22) <i>Usual R^2 can be negative; use raw R^2 (3.22)</i>
Rủi ro <i>Risk</i>	Hệ số chặn thừa chỉ làm giảm hiệu quả <i>A redundant intercept only costs efficiency</i>	Bỏ nhầm hệ số chặn làm β_2 chệch <i>Wrongly dropping it biases β_2</i>

Quy tắc thực hành: giữ hệ số chặn trừ khi có lý thuyết mạnh; nếu bỏ, trước hết ước lượng có hệ số chặn và kiểm định $H_0: \beta_1 = 0$ như bước 1.

Rule of practice: keep the intercept unless theory is strong; if dropping it, first estimate with an intercept and test $H_0: \beta_1 = 0$ as in step 1.

3.9. Đọc bảng kết quả EViews của hồi quy hai biến — *Reading the EViews output of a two-variable regression*

Bảng dưới đây là kết quả hồi quy bộ A theo khuôn EViews (lệnh equation eq1.ls Y c X). Bảng phần mềm giữ dấu chấm thập phân như máy in ra.

The table below is the regression of data set A in EViews layout (command equation eq1.ls Y c X). Software tables keep the decimal point as printed.

equation eq1.ls Y c X eq1.wald c(2) = -2.5				
Thông tin mô hình <i>Model information</i>				
Dependent Variable: Y				
Method: Least Squares				
Sample: 2025M01 2025M10				
Included observations: 10				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	103.8115	5.9251	17.5207	0.0000
X	-2.1127	0.2115	-9.9905	0.0000

Thống kê <i>Statistic</i>	Giá trị <i>Value</i>	Thống kê <i>Statistic</i>	Giá trị <i>Value</i>
R-squared	0.9258	Mean dependent var	45.5000
Adjusted R-squared	0.9165	S.D. dependent var	11.1580
S.E. of regression	3.2239	Akaike info criterion	5.3559
Sum squared resid	83.1463	Schwarz criterion	5.4164
Log likelihood	-24.7795	Hannan-Quinn criter.	5.2895
F-statistic	99.8100	Durbin-Watson stat	1.9777
Prob(F-statistic)	0.0000		

Ô trên bảng <i>Output cell</i>	Ý nghĩa và công thức <i>Meaning and formula</i>	Bộ A <i>Data set A</i>
Coefficient	β_1 (dòng C), β_2 (dòng X) <i>β_1 (row C), β_2 (row X)</i>	103,8115; -2,1127 <i>103.8115; -2.1127</i>
Std. Error	se(β_j), công thức (3.1)–(3.2) với σ <i>se(β_j), formulas (3.1)–(3.2) with σ</i>	5,9251; 0,2115 <i>5.9251; 0.2115</i>
t-Statistic	Coefficient/Std. Error, kiểm định $H_0: \beta_j = 0$ <i>Coefficient/Std. Error, tests $H_0: \beta_j = 0$</i>	17,5207; -9,9905 <i>17.5207; -9.9905</i>
Prob.	p-value hai phía của t, bậc tự do $n - 2$ <i>Two-sided p-value of t, $n - 2$ df</i>	< 0,0001; < 0,0001 <i>< 0.0001; < 0.0001</i>
R-squared	ESS/TSS = $1 - \text{RSS}/\text{TSS}$	0,9258 <i>0.9258</i>
Adjusted R-squared	$1 - (1 - R^2)(n - 1)/(n - 2)$	0,9165 <i>0.9165</i>
S.E. of regression	$\sigma = \sqrt{\text{RSS}/(n - 2)}$	3,2239 <i>3.2239</i>
Sum squared resid	$\text{RSS} = \sum e_i^2$	83,1463 <i>83.1463</i>
F-statistic	$R^2(n - 2)/(1 - R^2) = t^2$	99,8100 <i>99.8100</i>
Prob(F-statistic)	p-value của F, bằng Prob. của X <i>p-value of F, equal to Prob. of X</i>	< 0,0001 <i>< 0.0001</i>
Mean dependent var	$\bar{Y}$	45,5000 <i>45.5000</i>

Ô trên bảng <i>Output cell</i>	Ý nghĩa và công thức <i>Meaning and formula</i>	Bộ A <i>Data set A</i>
S.D. dependent var	$\sqrt{(TSS/(n-1))} \Rightarrow TSS = (n-1) \cdot S.D.^2$	11,1580 <i>11.1580</i>
Akaike, Schwarz, Hannan-Quinn	Tiêu chí thông tin, nhỏ hơn thì tốt hơn (Chương 9) <i>Information criteria, smaller is better (Chapter 9)</i>	
Durbin-Watson stat	Kiểm định tự tương quan (Chương 8) <i>Autocorrelation test (Chapter 8)</i>	

Kiểm định $H_0: \beta_2 = -2,5$ không đọc trực tiếp được từ bảng hồi quy (ô t-Statistic luôn ứng với $\beta^* = 0$). Trong EViews dùng kiểm định Wald trên hệ số C(2); kết quả dưới đây khớp tính tay ở mục 3.3: $F = t^2$, Chi-square = F khi chỉ có một ràng buộc.

The test of $H_0: \beta_2 = -2.5$ cannot be read off the regression table (the t-Statistic cell always refers to $\beta^ = 0$). EViews uses a Wald test on coefficient C(2); the result below matches the hand calculation in section 3.3: $F = t^2$, and Chi-square = F with a single restriction.*

Kiểm định <i>Test</i>	Thống kê <i>Statistic</i>	Giá trị <i>Value</i>	Xác suất <i>Probability</i>	Giá trị <i>Value</i>
Wald Test: C(2) = -2.5	t-statistic	1.8313	Prob. t(8)	0.1044
	F-statistic	3.3535	Prob. F(1,8)	0.1044
	Chi-square	3.3535	Prob. Chi-square(1)	0.0671

Prob. của F và của t bằng nhau (0,1044); Prob. Chi-square (0,0671) dùng phân phối tiệm cận nên nhỏ hơn, với mẫu nhỏ đọc dòng F hoặc t.

The Prob. values of F and t are equal (0.1044); the Chi-square Prob. (0.0671) uses the asymptotic distribution and is smaller; in small samples read the F or t line.

LỖI HAY GẶP / COMMON MISTAKE — Đọc Prob. = 0,0000 thành “xác suất bằng 0” hay “ H_0 sai chắc chắn”. 0,0000 chỉ là p-value nhỏ hơn 0,00005 sau khi làm tròn 4 chữ số. Một lỗi khác: lấy S.D. dependent var làm σ ; σ là S.E. of regression.
Reading Prob. = 0.0000 as “probability zero” or “ H_0 is certainly false”. 0.0000 is only a p-value below 0.00005 after rounding to 4 decimals. Another error: taking S.D. dependent var as σ ; σ is the S.E. of regression.

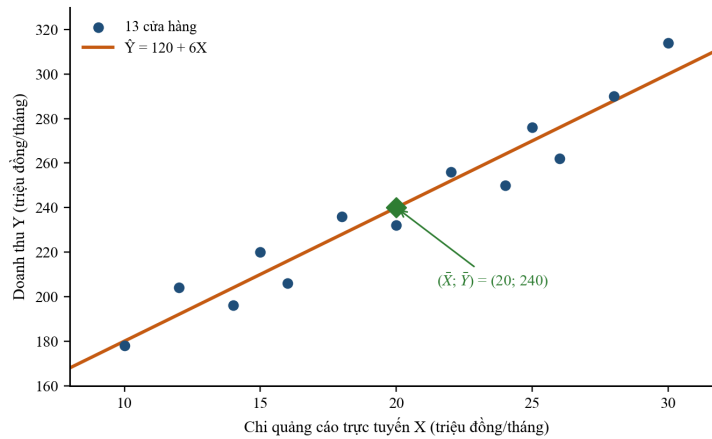
3.10. Bài thi mẫu dạng câu 1 đề UFM — Model exam question in the style of UFM question 1

Đề (dữ kiện đề bài, 4 điểm, khoảng 30 phút). Chuỗi cửa hàng mỹ phẩm Hoa Lài Xanh thu thập số liệu tháng 8/2026 của 13 cửa hàng: X là chi quảng cáo trực tuyến (triệu đồng/tháng), Y là doanh thu (triệu đồng/tháng). a) Ước lượng hàm hồi quy tuyến tính của Y theo X, nêu ý nghĩa các hệ số. b) Tính hệ số xác định, hệ số tương quan và giải thích.

c) Tìm khoảng tin cậy 95% của β_2 và của σ^2 . d) Với mức ý nghĩa 5%, chi quảng cáo có ảnh hưởng đến doanh thu không? e) Trưởng phòng kinh doanh cho rằng mỗi triệu đồng quảng cáo tăng thêm mang lại hơn 5 triệu đồng doanh thu. Kiểm định ý kiến với $\alpha = 5\%$. f) Dự báo doanh thu trung bình và doanh thu của một cửa hàng khi chi quảng cáo 25 triệu đồng, độ tin cậy 95%. g) Viết lại hàm hồi quy khi doanh thu tính bằng tỷ đồng; khi chi quảng cáo tính bằng nghìn đồng.

Question (given data, 4 marks, about 30 minutes). The Hoa Lai Xanh cosmetics chain collects August 2026 data for 13 shops: X is online advertising spend (million VND a month), Y is revenue (million VND a month). a) Estimate the linear regression of Y on X and interpret the coefficients. b) Compute the coefficient of determination and the correlation coefficient, and explain them. c) Find 95% confidence intervals for β_2 and σ^2 . d) At the 5% level, does advertising affect revenue? e) The sales manager claims that each extra million VND of advertising brings more than 5 million VND of revenue. Test the claim at $\alpha = 5\%$. f) Forecast mean revenue and the revenue of one shop at an advertising spend of 25 million VND, with 95% confidence. g) Rewrite the regression with revenue in VND billion; with advertising in thousand VND.

i	X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	10	178	-10	-62	620	100	3.844 3,844
2	12	204	-8	-36	288	64	1.296 1,296
3	14	196	-6	-44	264	36	1.936 1,936
4	15	220	-5	-20	100	25	400
5	16	206	-4	-34	136	16	1.156 1,156
6	18	236	-2	-4	8	4	16
7	20	232	0	-8	0	0	64
8	22	256	2	16	32	4	256
9	24	250	4	10	40	16	100
10	25	276	5	36	180	25	1.296 1,296
11	26	262	6	22	132	36	484
12	28	290	8	50	400	64	2.500 2,500
13	30	314	10	74	740	100	5.476 5,476
Tổng Total	260	3.120 3,120	0	0	2.940 2,940	490	18.824 18,824



Hình 3.6. Bài thi mẫu Hoa Lài Xanh: số liệu 13 cửa hàng và đường hồi quy mẫu đi qua $(\bar{X}; \bar{Y})$
Figure 3.6. Hoa Lài Xanh model exam: data for 13 shops and the sample regression line through $(\bar{X}; \bar{Y})$

Lời giải — Solution

1. a) $\bar{X} = 260/13 = 20$; $\bar{Y} = 3,120/13 = 240$. Từ bảng: $\sum x_i y_i = 2,940$; $\sum x_i^2 = 490$. $\beta_2 = 2,940/490 = 6$; $\beta_1 = 240 - 6 \times 20 = 120$. Hàm hồi quy mẫu: $\hat{Y} = 120 + 6X$.
 a) $\bar{X} = 260/13 = 20$; $\bar{Y} = 3,120/13 = 240$. From the table: $\sum x_i y_i = 2,940$; $\sum x_i^2 = 490$. $\beta_2 = 2,940/490 = 6$; $\beta_1 = 240 - 6 \times 20 = 120$. Sample regression function: $\hat{Y} = 120 + 6X$.
2. Ý nghĩa: $\beta_2 = 6$: chi quảng cáo tăng 1 triệu đồng/tháng thì doanh thu trung bình tăng 6 triệu đồng/tháng (các yếu tố khác không đổi). $\beta_1 = 120$: doanh thu trung bình khi không quảng cáo trực tuyến là 120 triệu đồng/tháng; $X = 0$ nằm ngoài khoảng dữ liệu (10–30) nên chỉ đọc tham khảo.
 Meaning: $\beta_2 = 6$: an extra 1 million VND a month of advertising raises mean revenue by 6 million VND a month (other things equal). $\beta_1 = 120$: mean revenue with no online advertising is 120 million VND a month; $X = 0$ is outside the data range (10–30), so read it with caution.
3. b) $TSS = \sum y_i^2 = 18,824$; $ESS = \beta_2^2 \sum x_i^2 = 6^2 \times 490 = 17,640$; $RSS = 18,824 - 17,640 = 1,184$. $R^2 = 17,640/18,824 = 0,9371$: chi quảng cáo giải thích 93,71% sự biến động của doanh thu giữa các cửa hàng. $r = +\sqrt{R^2} = 0,9680$ (cùng dấu β_2): tương quan thuận, chặt.
 b) $TSS = \sum y_i^2 = 18,824$; $ESS = \beta_2^2 \sum x_i^2 = 6^2 \times 490 = 17,640$; $RSS = 18,824 - 17,640 = 1,184$. $R^2 = 17,640/18,824 = 0,9371$: advertising explains 93.71% of the variation in revenue across shops. $r = +\sqrt{R^2} = 0.9680$ (same sign as β_2): a strong positive correlation.
4. c) $\sigma^2 = RSS/(n - 2) = 1,184/11 = 107,6364$; $se(\beta_2) = \sqrt{(107,6364/490)} = 0,4687$. $t_{0,025}^{(11)} = 2,201$; $\beta_2 \in 6 \pm 2,201 \times 0,4687 = (4,9684; 7,0316)$. σ^2 : $\chi^2_{0,025}^{(11)} = 21,920$, $\chi^2_{0,975}^{(11)} = 3,816$; $1,184/21,920 \leq \sigma^2 \leq 1,184/3,816 \Rightarrow (54,015; 310,273)$.
 c) $\sigma^2 = RSS/(n - 2) = 1,184/11 = 107.6364$; $se(\beta_2) = \sqrt{(107.6364/490)} = 0.4687$. $t_{0.025}^{(11)} = 2.201$; $\beta_2 \in 6 \pm 2.201 \times 0.4687 = (4.9684; 7.0316)$. σ^2 : $\chi^2_{0.025}^{(11)} = 21.920$, $\chi^2_{0.975}^{(11)} = 3.816$; $1,184/21.920 \leq \sigma^2 \leq 1,184/3.816 \Rightarrow (54.015; 310.273)$.
5. d) $H_0: \beta_2 = 0$; $H_1: \beta_2 \neq 0$. $t = 6/0,4687 = 12,802 > 2,201$ nên bác bỏ H_0 : với mức ý nghĩa 5%, chi quảng cáo trực tuyến có ảnh hưởng đến doanh thu.

- d) $H_0: \beta_2 = 0$; $H_1: \beta_2 \neq 0$. $t = 6/0.4687 = 12.802 > 2.201$, so reject H_0 : at the 5% level, online advertising affects revenue.
6. e) “Hơn 5 triệu đồng” là điều cần chứng minh nên đặt ở H_1 : $H_0: \beta_2 \leq 5$; $H_1: \beta_2 > 5$. $t = (6 - 5)/0.4687 = 2.134$. Một phía phải: $t_{0.05}^{(11)} = 1.796$. $t > 1.796$ nên bác bỏ H_0 : số liệu ủng hộ ý kiến của trưởng phòng ở mức 5%.
- e) “More than 5 million” is the claim to be shown, so it goes in H_1 : $H_0: \beta_2 \leq 5$; $H_1: \beta_2 > 5$. $t = (6 - 5)/0.4687 = 2.134$. Right-tailed: $t_{0.05}^{(11)} = 1.796$. $t > 1.796$, so reject H_0 : the data support the manager’s claim at 5%.
7. Chú ý: khoảng tin cậy hai phía ở ý c chứa 5 và $t = 2.134 < 2.201$, nên nếu đề hỏi “mỗi triệu đồng quảng cáo mang lại 5 triệu đồng doanh thu” ($H_1: \beta_2 \neq 5$) thì kết luận là chưa đủ cơ sở bác bỏ H_0 . Cùng số liệu, kiểm định một phía mạnh hơn khi hướng đã biết trước.
- Note: the two-sided interval in part c contains 5 and $t = 2.134 < 2.201$, so if the question read “each million of advertising brings 5 million of revenue” ($H_1: \beta_2 \neq 5$), H_0 would not be rejected. With the same data, a one-sided test has more power when the direction is known in advance.
8. f) $\hat{Y}_0 = 120 + 6 \times 25 = 270$. $\text{Var}(\hat{Y}_0) = 107.6364 \times [1/13 + (25 - 20)^2/490] = 13.7714$; $se = 3.7110$. Trung bình: $270 \pm 2.201 \times 3.7110 = (261.83; 278.17)$ triệu đồng. Cá biệt: $\text{Var} = 107.6364 + 13.7714 = 121.4077$; $se = 11.0185$; $(245.75; 294.25)$ triệu đồng.
- f) $\hat{Y}_0 = 120 + 6 \times 25 = 270$. $\text{Var}(\hat{Y}_0) = 107.6364 \times [1/13 + (25 - 20)^2/490] = 13.7714$; $se = 3.7110$. Mean: $270 \pm 2.201 \times 3.7110 = (261.83; 278.17)$ million VND. Individual: $\text{Var} = 107.6364 + 13.7714 = 121.4077$; $se = 11.0185$; $(245.75; 294.25)$ million VND.
9. g) Y tính bằng tỷ đồng: $Y^* = Y/1.000$, $w_1 = 0.001 \Rightarrow \hat{Y}^* = 0.12 + 0.006X$. X tính bằng nghìn đồng: $X^* = 1.000X$, $w_2 = 1.000 \Rightarrow \hat{Y} = 120 + 0.006X^*$. Cả hai trường hợp t , R^2 giữ nguyên (12,802; 0,9371).
- g) Y in VND billion: $Y^* = Y/1,000$, $w_1 = 0.001 \Rightarrow \hat{Y}^* = 0.12 + 0.006X$. X in thousand VND: $X^* = 1,000X$, $w_2 = 1,000 \Rightarrow \hat{Y} = 120 + 0.006X^*$. In both cases t and R^2 stay the same (12.802; 0.9371).

LỖI HAY GẶP / COMMON MISTAKE — Ở ý e, đặt “ $\beta_2 > 5$ ” vào H_0 rồi kết luận ngược, hoặc dùng $t(\alpha/2) = 2.201$ cho kiểm định một phía. Quy tắc: điều đề muốn kiểm chứng (“hơn”, “ít nhất”, “có làm tăng”) đặt ở H_1 ; dấu bằng luôn ở H_0 .

In part e, putting “ $\beta_2 > 5$ ” in H_0 and reaching the opposite verdict, or using $t(\alpha/2) = 2.201$ for a one-sided test. Rule: what the question wants verified (“more than”, “at least”, “does it raise”) goes in H_1 ; the equality sign always stays in H_0 .

3.11. Tóm tắt, so sánh khái niệm dễ nhầm và thuật ngữ — Summary, commonly confused concepts and glossary

PHẢI THUỘC / MEMORISE — Bậc tự do hồi quy hai biến: $n - 2$. Khoảng tin cậy: $\beta_j \pm t(\alpha/2; n - 2) \cdot se(\beta_j)$; σ^2 : $[(n - 2)\sigma^2/\chi^2(\alpha/2); (n - 2)\sigma^2/\chi^2(1 - \alpha/2)]$. Kiểm định: $t = (\beta_2 - \beta^*)/se(\beta_2)$; hai phía $|t| > t(\alpha/2)$, một phía $t > t_\alpha$ hoặc $t < -t_\alpha$; H_1 chứa điều cần chứng minh. $F = R^2(n - 2)/(1 - R^2) = t^2$. Dự báo: $\hat{Y}_0 \pm t \cdot se$; Var trung bình = $\sigma^2[1/n + (X_0 - \bar{X})^2/\Sigma x^2]$, cá biệt

cộng thêm σ^2 . Đổi đơn vị: $\beta_2^* = (w_1/w_2)\beta_2$, $\beta_1^* = w_1\beta_1$; t , R^2 không đổi. Log–log: co giãn; log–lin: $100 \cdot \beta_2\%$ mỗi đơn vị X , tăng trưởng $\exp(\beta_2) - 1$; lin–log: $\beta_2/100$ đơn vị Y mỗi 1% X . Qua gốc: $\beta_2 = \Sigma XY / \Sigma X^2$, bậc tự do $n - 1$.

Degrees of freedom in two-variable regression: $n - 2$. Confidence interval: $\beta_j \pm t(\alpha/2; n - 2) \cdot se(\beta_j)$; σ^2 : $[(n - 2)\sigma^2/\chi^2(\alpha/2); (n - 2)\sigma^2/\chi^2(1 - \alpha/2)]$. Test: $t = (\beta_2 - \beta_2^)/se(\beta_2)$; two-sided $|t| > t(\alpha/2)$, one-sided $t > t_\alpha$ or $t < -t_\alpha$; H_1 holds the claim to be shown. $F = R^2(n - 2)/(1 - R^2) = t^2$. Forecast: $\hat{Y}_0 \pm t \cdot se$; mean variance = $\sigma^2[1/n + (X_0 - \bar{X})^2/\Sigma x^2]$, individual adds σ^2 . Units: $\beta_2^* = (w_1/w_2)\beta_2$, $\beta_1^* = w_1\beta_1$; t and R^2 unchanged. Log–log: elasticity; log–lin: $100 \cdot \beta_2\%$ per unit of X , growth $\exp(\beta_2) - 1$; lin–log: $\beta_2/100$ units of Y per 1% of X . Through the origin: $\beta_2 = \Sigma XY / \Sigma X^2$, $n - 1$ degrees of freedom.*

Các khái niệm dễ nhầm — Commonly confused concepts

Khái niệm 1 <i>Concept 1</i>	Khái niệm 2 <i>Concept 2</i>	Phân biệt <i>Distinction</i>
Dự báo trung bình $E(Y X_0)$ <i>Mean forecast $E(Y X_0)$</i>	Dự báo cá biệt Y_0 <i>Individual forecast Y_0</i>	Cùng Y_0 ; phương sai cá biệt cộng thêm σ^2 <i>Same Y_0; individual variance adds σ^2</i>
Kiểm định hai phía <i>Two-sided test</i>	Kiểm định một phía <i>One-sided test</i>	$t(\alpha/2)$ và $ t $; t_α và dấu của t <i>$t(\alpha/2)$ with t; t_α with the sign of t</i>
Khoảng tin cậy β_2 <i>Interval for β_2</i>	Khoảng tin cậy σ^2 <i>Interval for σ^2</i>	Phân phối t , đối xứng; phân phối χ^2 , không đối xứng <i>t distribution, symmetric; χ^2, asymmetric</i>
Mức ý nghĩa α <i>Significance level α</i>	p-value	α đặt trước; p-value tính từ mẫu, bác bỏ H_0 khi $p < \alpha$ <i>α is set in advance; p comes from the sample, reject H_0 if $p < \alpha$</i>
“Chưa đủ cơ sở bác bỏ H_0 ” <i>“H_0 not rejected”</i>	“ H_0 đúng” <i>“H_0 is true”</i>	Kiểm định không chứng minh được H_0 <i>A test cannot prove H_0</i>
Kiểm định t ($H_0: \beta_2 = 0$) <i>t-test ($H_0: \beta_2 = 0$)</i>	Kiểm định F sự phù hợp <i>Overall F-test</i>	Hai biến: $F = t^2$, cùng p-value; hồi quy bội thì khác <i>Two variables: $F = t^2$, same p-value; differ in multiple regression</i>
Hệ số co giãn (log–log) <i>Elasticity (log–log)</i>	Tốc độ tăng trưởng (log–lin) <i>Growth rate (log–lin)</i>	% Y theo 1% X ; % Y theo 1 đơn vị X (thời gian) <i>% Y per 1% X; % Y per unit of X (time)</i>
R^2 có hệ số chặn <i>R^2 with intercept</i>	R^2 thô qua gốc <i>Raw R^2 through origin</i>	Mẫu số Σy^2 và ΣY^2 : không so sánh được <i>Denominators Σy^2 and ΣY^2: not comparable</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>
Phân phối mẫu của ước lượng	Sampling distribution	Phân phối của β qua các mẫu <i>Distribution of β across samples</i>
Sai số chuẩn	Standard error	Độ lệch chuẩn ước lượng của β <i>Estimated standard deviation of β</i>
Bậc tự do	Degrees of freedom	$n - 2$ ở hồi quy hai biến <i>$n - 2$ in two-variable regression</i>
Khoảng tin cậy	Confidence interval	Khoảng chứa tham số với độ tin cậy $1 - \alpha$ <i>Range holding the parameter with confidence $1 - \alpha$</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 cho phép <i>Allowed type I error probability</i>
Giá trị tới hạn	Critical value	Ngưỡng tra bảng của thống kê <i>Table threshold of the statistic</i>
Miền bác bỏ	Rejection region	Tập giá trị thống kê dẫn tới bác bỏ H_0 <i>Values leading to rejection of H_0</i>
Giá trị p	p-value	Mức ý nghĩa nhỏ nhất bác bỏ H_0 <i>Smallest level at which H_0 is rejected</i>
Kiểm định hai phía	Two-tailed test	$H_1: \beta \neq \beta^*$
Kiểm định một phía	One-tailed test	$H_1: \beta > \beta^*$ hoặc $\beta < \beta^*$ <i>$H_1: \beta > \beta^*$ or $\beta < \beta^*$</i>
Phân tích phương sai	Analysis of variance (ANOVA)	Tách TSS = ESS + RSS <i>Splitting TSS = ESS + RSS</i>
Kiểm định F	F-test	So ESS/1 với RSS/($n - 2$) <i>Compares ESS/1 with RSS/($n - 2$)</i>
Kiểm định Wald	Wald test	Kiểm định ràng buộc trên hệ số <i>Test of restrictions on coefficients</i>
Dự báo điểm	Point forecast	$\hat{Y}_0 = \beta_1 + \beta_2 X_0$
Khoảng dự báo	Prediction interval	$\hat{Y}_0 \pm t \cdot se$ cho Y_0 hoặc $E(Y X_0)$ <i>$\hat{Y}_0 \pm t \cdot se$ for Y_0 or $E(Y X_0)$</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Dự báo ngoài mẫu	Out-of-sample forecast	X_0 ngoài khoảng dữ liệu <i>X_0 beyond the data range</i>
Hệ số co giãn	Elasticity	% thay đổi Y khi X tăng 1% <i>% change in Y per 1% rise in X</i>
Mô hình log–log	Log–log (double-log) model	ln Y theo ln X <i>ln Y on ln X</i>
Mô hình bán log	Semi-log model	Log–lin hoặc lin–log <i>Log–lin or lin–log</i>
Mô hình nghịch đảo	Reciprocal model	Y theo 1/X <i>Y on 1/X</i>
Hồi quy qua gốc tọa độ	Regression through the origin	Mô hình không có hệ số chặn <i>Model with no intercept</i>
R ² thô	Raw R-squared	$(\Sigma XY)^2 / (\Sigma X^2 \Sigma Y^2)$

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

Bài 1. Tính hồi quy từ các tổng và kiểm định giá trị cho trước — *Exercise 1.*

Regression from sums and a test of a stated value

Công ty phân phối thiết bị văn phòng Mộc Lan Đông khảo sát 12 nhân viên bán hàng: X là số năm kinh nghiệm, Y là doanh số tháng (triệu đồng). Dữ kiện đề bài: $\Sigma X = 60$; $\Sigma Y = 828$; $\Sigma X^2 = 386$; $\Sigma Y^2 = 59.786$; $\Sigma XY = 4.614$. (a) Ước lượng hàm hồi quy Y theo X. (b) Tính R². (c) Với mức ý nghĩa 1%, kinh nghiệm có ảnh hưởng đến doanh số không? (d) Tìm khoảng tin cậy 95% của β_1 . (e) Trưởng phòng nhân sự cho rằng mỗi năm kinh nghiệm làm doanh số tăng 5 triệu đồng. Kiểm định ý kiến này với $\alpha = 5\%$.

Mộc Lan Đông, an office-equipment distributor, surveys 12 sales staff: X is years of experience and Y is monthly sales (million VND). Given data: $\Sigma X = 60$; $\Sigma Y = 828$; $\Sigma X^2 = 386$; $\Sigma Y^2 = 59,786$; $\Sigma XY = 4,614$. (a) Estimate the regression of Y on X. (b) Compute R². (c) At the 1% significance level, does experience affect sales? (d) Find the 95% confidence interval for β_1 . (e) The HR manager claims that each year of experience raises sales by 5 million VND. Test the claim at $\alpha = 5\%$.

- $\bar{X} = 60/12 = 5$; $\bar{Y} = 828/12 = 69$. $\Sigma x^2 = \Sigma X^2 - n\bar{X}^2 = 386 - 12 \cdot 5^2 = 86$; $\Sigma xy = \Sigma XY - n\bar{X}\bar{Y} = 4.614 - 12 \cdot 5 \cdot 69 = 474$; $\Sigma y^2 = \Sigma Y^2 - n\bar{Y}^2 = 2.654$.
 $\bar{X} = 60/12 = 5$; $\bar{Y} = 828/12 = 69$. $\Sigma x^2 = \Sigma X^2 - n\bar{X}^2 = 386 - 12 \cdot 5^2 = 86$; $\Sigma xy = \Sigma XY - n\bar{X}\bar{Y} = 4,614 - 12 \cdot 5 \cdot 69 = 474$; $\Sigma y^2 = \Sigma Y^2 - n\bar{Y}^2 = 2,654$.
- (a) $\beta_2 = 474/86 = 5,5116$; $\beta_1 = 69 - 5,5116 \cdot 5 = 41,4419$. $\hat{Y} = 41,4419 + 5,5116X$: thêm một năm kinh nghiệm, doanh số trung bình tăng 5,51 triệu đồng/tháng.
(a) $\beta_2 = 474/86 = 5.5116$; $\beta_1 = 69 - 5.5116 \cdot 5 = 41.4419$. $\hat{Y} = 41.4419 + 5.5116X$: one more year of experience raises mean sales by 5.51 million VND a month.

3. (b) $ESS = \beta_2 \cdot \Sigma xy = 2.612,512$; $R^2 = ESS/\Sigma y^2 = 2.612,512/2.654 = 0,9844$: kinh nghiệm giải thích 98,44% biến động doanh số. $RSS = 2.654 - 2.612,512 = 41,488$; $\sigma^2 = RSS/(n - 2) = 4,1488$.
 (b) $ESS = \beta_2 \cdot \Sigma xy = 2,612.512$; $R^2 = ESS/\Sigma y^2 = 2,612.512/2,654 = 0.9844$: experience explains 98.44% of the variation in sales. $RSS = 2,654 - 2,612.512 = 41.488$; $\sigma^2 = RSS/(n - 2) = 4.1488$.
4. (c) $se(\beta_2) = \sqrt{(\sigma^2/\Sigma x^2)} = 0,2196$; $H_0: \beta_2 = 0$, $H_1: \beta_2 \neq 0$; $t = 5,5116/0,2196 = 25,094$. Tra bảng $t_{0,005}^{(10)} = 3,169$. $|t| > 3,169$ nên bác bỏ H_0 : ở mức 1%, kinh nghiệm có ảnh hưởng đến doanh số.
 (c) $se(\beta_2) = \sqrt{(\sigma^2/\Sigma x^2)} = 0.2196$; $H_0: \beta_2 = 0$, $H_1: \beta_2 \neq 0$; $t = 5.5116/0.2196 = 25.094$. Table value $t_{0,005}^{(10)} = 3.169$. $|t| > 3.169$, so reject H_0 : at 1%, experience affects sales.
5. (d) $se(\beta_1) = \sqrt{(\sigma^2 \cdot \Sigma X^2 / (n \cdot \Sigma x^2))} = \sqrt{(4,1488 \cdot 386 / (12 \cdot 86))} = 1,2457$. $t_{0,025}^{(10)} = 2,228$; $\beta_1 \in 41,4419 \pm 2,228 \cdot 1,2457 = (38,666; 44,217)$ triệu đồng.
 (d) $se(\beta_1) = \sqrt{(\sigma^2 \cdot \Sigma X^2 / (n \cdot \Sigma x^2))} = \sqrt{(4.1488 \cdot 386 / (12 \cdot 86))} = 1.2457$. $t_{0,025}^{(10)} = 2.228$; $\beta_1 \in 41.4419 \pm 2.228 \cdot 1.2457 = (38.666; 44.217)$ million VND.
6. (e) $H_0: \beta_2 = 5$, $H_1: \beta_2 \neq 5$; $t = (5,5116 - 5)/0,2196 = 2,329 > 2,228$ nên bác bỏ H_0 . Mức tăng 5 triệu đồng/năm không phù hợp với số liệu; ước lượng điểm 5,51 triệu đồng cao hơn ý kiến của trưởng phòng.
 (e) $H_0: \beta_2 = 5$, $H_1: \beta_2 \neq 5$; $t = (5.5116 - 5)/0.2196 = 2.329 > 2.228$, so reject H_0 . A rise of 5 million VND per year is not consistent with the data; the point estimate of 5.51 million is higher than the manager's figure.

Bài 2. Hệ số co giãn chi tiêu ăn uống theo thu nhập (bộ D) — Exercise 2. Income elasticity of food spending (data set D)

Hồi quy log-log chi tiêu ăn uống (Y, triệu đồng/tháng) theo thu nhập (TN, triệu đồng/tháng) của 40 hộ gia đình (bộ D) cho kết quả: $\ln \hat{Y} = -0,6199 + 0,8231 \cdot \ln TN$; $se(\beta_2) = 0,0451$; $R^2 = 0,8975$. (a) Nêu ý nghĩa của β_2 . (b) Thu nhập có ảnh hưởng đến chi tiêu ăn uống không ($\alpha = 5\%$)? (c) Tìm khoảng tin cậy 95% của hệ số co giãn. (d) Có thể coi ăn uống là hàng có co giãn đơn vị ($\beta_2 = 1$) không? (e) Hộ có thu nhập tăng 10% thì chi tiêu ăn uống thay đổi bao nhiêu phần trăm?

A log-log regression of food spending (Y, million VND a month) on income (TN, million VND a month) for 40 households (data set D) gives: $\ln \hat{Y} = -0.6199 + 0.8231 \cdot \ln TN$; $se(\beta_2) = 0.0451$; $R^2 = 0.8975$. (a) Interpret β_2 . (b) Does income affect food spending ($\alpha = 5\%$)? (c) Find the 95% confidence interval for the elasticity. (d) Can food be treated as having unit elasticity ($\beta_2 = 1$)? (e) By what percentage does food spending change when a household's income rises by 10%?

1. (a) $\beta_2 = 0,8231$ là hệ số co giãn của chi tiêu ăn uống theo thu nhập: thu nhập tăng 1% thì chi tiêu ăn uống trung bình tăng khoảng 0,823%.
 (a) $\beta_2 = 0.8231$ is the income elasticity of food spending: a 1% rise in income raises mean food spending by about 0.823%.
2. (b) $H_0: \beta_2 = 0$; $t = 0,8231/0,0451 = 18,24$; bậc tự do $n - 2 = 38$; $t_{0,025}^{(38)} = 2,024$. $|t| > 2,024$ nên bác bỏ H_0 : thu nhập có ảnh hưởng.
 (b) $H_0: \beta_2 = 0$; $t = 0.8231/0.0451 = 18.24$; degrees of freedom $n - 2 = 38$; $t_{0,025}^{(38)} = 2.024$. $|t| > 2.024$, so reject H_0 : income has an effect.

3. (c) $\beta_2 \in 0,8231 \pm 2,024 \cdot 0,0451 = (0,7317; 0,9144)$. Khoảng nằm hẳn dưới 1.
 (c) $\beta_2 \in 0.8231 \pm 2.024 \cdot 0.0451 = (0.7317; 0.9144)$. The interval lies entirely below 1.
4. (d) $H_0: \beta_2 = 1, H_1: \beta_2 \neq 1; t = (0,8231 - 1)/0,0451 = -3,921; |t| > 2,024$ nên bác bỏ H_0 (cùng kết luận với khoảng tin cậy không chứa 1). Co giãn nhỏ hơn 1: ăn uống là hàng thiết yếu.
 (d) $H_0: \beta_2 = 1, H_1: \beta_2 \neq 1; t = (0.8231 - 1)/0.0451 = -3.921; |t| > 2.024$, so reject H_0 (the same verdict as the interval that excludes 1). Elasticity below 1: food is a necessity.
5. (e) Thu nhập tăng 10%: chi tiêu ăn uống tăng xấp xỉ $10 \times 0,8231 = 8,23\%$; tính chính xác $1,1^{\beta_2} - 1 = 8,16\%$. Bộ D có phương sai sai số thay đổi (Chương 7) nên se ở đây chỉ dùng cho bài tập này; báo cáo thật phải dùng se vững.
 (e) A 10% rise in income raises food spending by about $10 \times 0.8231 = 8.23\%$; exactly $1.1^{\beta_2} - 1 = 8.16\%$. Data set D is heteroscedastic (Chapter 7), so the standard error here serves this exercise only; a real report must use robust standard errors.

Bài 3. Tốc độ tăng trưởng doanh thu bằng mô hình log-lin — Exercise 3. Revenue growth rate from a log-lin model

Doanh thu (tỷ đồng) của Công ty CP Nội thất Gỗ Tràm Bạc giai đoạn 2016–2025 (dữ kiện đề bài): 48,2; 52,0; 57,1; 61,5; 63,0; 70,4; 77,8; 83,1; 90,5; 99,6. Đặt $t = 1$ cho năm 2016, ..., $t = 10$ cho năm 2025. Hồi quy $\ln Y$ theo t được $\ln \hat{Y} = 3,7901 + 0,07942t$; $se(\beta_2) = 0,00210$; $R^2 = 0,9944$. (a) Tính tốc độ tăng trưởng doanh thu bình quân năm. (b) Tìm khoảng tin cậy 95% cho tốc độ tăng trưởng. (c) Ban giám đốc công bố doanh thu tăng bình quân 8%/năm. Kiểm định ở mức 5%. (d) Dự báo điểm doanh thu năm 2026.
 Revenue (VND billion) of Gỗ Tràm Bạc Furniture JSC for 2016–2025 (given data): 48.2; 52.0; 57.1; 61.5; 63.0; 70.4; 77.8; 83.1; 90.5; 99.6. Let $t = 1$ for 2016, ..., $t = 10$ for 2025. Regressing $\ln Y$ on t gives $\ln \hat{Y} = 3.7901 + 0.07942t$; $se(\beta_2) = 0.00210$; $R^2 = 0.9944$. (a) Find the average annual revenue growth rate. (b) Find a 95% confidence interval for the growth rate. (c) The board states that revenue grows by 8% a year on average. Test at the 5% level. (d) Give a point forecast of revenue for 2026.

1. (a) Mô hình log-lin: $\beta_2 = 0,07942$ là tốc độ tăng trưởng liên tục. Tốc độ tăng trưởng gộp năm $r = \exp(\beta_2) - 1 = 8,27\%/năm$.
 (a) In the log-lin model $\beta_2 = 0.07942$ is the continuous growth rate. The compound annual rate is $r = \exp(\beta_2) - 1 = 8.27\%$ a year.
2. (b) $t_{0,025}^{(8)} = 2,306$; $\beta_2 \in 0,07942 \pm 2,306 \cdot 0,00210 = (0,07457; 0,08426)$. Đổi sang tốc độ gộp: (7,74%; 8,79%).
 (b) $t_{0,025}^{(8)} = 2.306$; $\beta_2 \in 0.07942 \pm 2.306 \cdot 0.00210 = (0.07457; 0.08426)$. As compound rates: (7.74%; 8.79%).
3. (c) 8%/năm tương ứng $\beta^* = \ln 1,08 = 0,07696$. $H_0: \beta_2 = 0,07696$; $t = (0,07942 - 0,07696)/0,00210 = 1,168$; $|t| < 2,306$ nên chưa đủ cơ sở bác bỏ H_0 : con số 8%/năm không mâu thuẫn với số liệu.
 (c) 8% a year corresponds to $\beta^* = \ln 1.08 = 0.07696$. $H_0: \beta_2 = 0.07696$; $t = (0.07942 - 0.07696)/0.00210 = 1.168$; $|t| < 2.306$, so there is insufficient evidence to reject H_0 : the 8% figure is consistent with the data.

4. Lưu ý: kiểm định phải viết trên β_2 (thang log), không viết $H_0: \beta_2 = 0,08$. Sai lệch giữa 0,08 và $\ln 1,08$ nhỏ ở đây nhưng lớn dần khi tốc độ tăng cao.

Note: the test must be written on β_2 (log scale), not as $H_0: \beta_2 = 0.08$. The gap between 0.08 and $\ln 1.08$ is small here but widens as growth rates rise.

5. (d) Năm 2026 ứng với $t = 11$: $\ln \hat{Y} = 3,7901 + 0,07942 \cdot 11 = 4,6636$; $\hat{Y} = \exp(4,6636) = 106,0$ tỷ đồng. Đây là dự báo ngoài mẫu một bước, giả định xu thế giữ nguyên.

(d) 2026 corresponds to $t = 11$: $\ln \hat{Y} = 3.7901 + 0.07942 \cdot 11 = 4.6636$; $\hat{Y} = \exp(4.6636) = \text{VND } 106.0 \text{ billion}$. This is a one-step out-of-sample forecast that assumes the trend continues.

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

Câu 1. Trình bày ba cách kết luận một kiểm định t về hệ số hồi quy (miền bác bỏ, khoảng tin cậy, p-value). Minh họa bằng kiểm định $H_0: \beta_2 = -2,5$ trên bộ A và cho biết khi nào ba cách cho cùng kết luận.

Question 1. Explain the three ways of concluding a t-test on a regression coefficient (rejection region, confidence interval, p-value). Illustrate with the test of $H_0: \beta_2 = -2.5$ on data set A and state when the three agree.

- Miền bác bỏ: tính $t = (\beta_2 - \beta^*)/\text{se}(\beta_2) = 1,831$, so với $t_{0,025}^{(8)} = 2,306$; $|t| < 2,306$ nên chưa đủ cơ sở bác bỏ H_0 .
Rejection region: $t = (\beta_2 - \beta^)/\text{se}(\beta_2) = 1.831$, compared with $t_{0.025}^{(8)} = 2.306$; $|t| < 2.306$, so there is insufficient evidence to reject H_0 .*
- Khoảng tin cậy 95% của β_2 là $(-2,6004; -1,6251)$, chứa $-2,5$ nên chưa đủ cơ sở bác bỏ H_0 .
The 95% confidence interval for β_2 is $(-2.6004; -1.6251)$; it contains -2.5 , so H_0 is not rejected.
- p-value = $2 \cdot P(T_8 > |t|) = 0,1044 > 0,05$ nên chưa đủ cơ sở bác bỏ H_0 .
p-value = $2 \cdot P(T_8 > |t|) = 0.1044 > 0.05$, so H_0 is not rejected.
- Ba cách cho cùng kết luận khi dùng cùng mức ý nghĩa α , cùng kiểm định hai phía, cùng bậc tự do $n - 2$. Kiểm định một phía thì so p-value một phía với α , hoặc dùng khoảng tin cậy một phía.
The three agree when they use the same α , the same two-sided alternative and the same $n - 2$ degrees of freedom. For a one-sided test, compare the one-sided p-value with α or use a one-sided interval.
- p-value cho thêm thông tin: biết được kết luận ở mọi mức α mà không phải tra bảng lại; khoảng tin cậy cho biết cùng lúc mọi giá trị β^* không bị bác bỏ.
The p-value adds information: it gives the verdict at every α without a new table lookup; the interval shows at once every β^ that would not be rejected.*

Câu 2. Phân biệt dự báo giá trị trung bình và dự báo giá trị cá biệt trong hồi quy hai biến. Viết công thức phương sai của mỗi loại, giải thích vì sao khoảng dự báo cá biệt rộng hơn và vì sao cả hai khoảng hẹp nhất tại $X_0 = \bar{X}$.

Question 2. Distinguish the forecast of the mean value from the forecast of an individual value in

two-variable regression. Give the variance of each, explain why the individual interval is wider and why both are narrowest at $X_0 = \bar{X}$.

- Dự báo trung bình ước lượng $E(Y | X_0) = \beta_1 + \beta_2 X_0$: mức trung bình của mọi quan sát có $X = X_0$. Dự báo cá biệt ước lượng một giá trị $Y_0 = \beta_1 + \beta_2 X_0 + u_0$ cụ thể.
The mean forecast estimates $E(Y | X_0) = \beta_1 + \beta_2 X_0$, the average over all observations with $X = X_0$. The individual forecast estimates one particular value $Y_0 = \beta_1 + \beta_2 X_0 + u_0$.
- $\text{Var}(\hat{Y}_0) = \sigma^2[1/n + (X_0 - \bar{X})^2/\Sigma x^2]$; $\text{Var}(Y_0 - \hat{Y}_0) = \sigma^2[1 + 1/n + (X_0 - \bar{X})^2/\Sigma x^2]$.
 $\text{Var}(\hat{Y}_0) = \sigma^2[1/n + (X_0 - \bar{X})^2/\Sigma x^2]$; $\text{Var}(Y_0 - \hat{Y}_0) = \sigma^2[1 + 1/n + (X_0 - \bar{X})^2/\Sigma x^2]$.
- Khoảng cá biệt rộng hơn vì cộng thêm σ^2 : ngoài sai số khi ước lượng đường hồi quy còn có sai số ngẫu nhiên u_0 của chính quan sát mới; phần này không giảm khi n tăng.
The individual interval is wider because σ^2 is added: besides the error in estimating the line there is the disturbance u_0 of the new observation itself, which does not shrink as n grows.
- Số hạng $(X_0 - \bar{X})^2$ bằng 0 tại $X_0 = \bar{X}$ và tăng khi X_0 ra xa trung bình, nên dải dự báo có dạng hai đường cong loe ra hai phía; dự báo ngoài khoảng dữ liệu kém tin cậy.
The term $(X_0 - \bar{X})^2$ is zero at $X_0 = \bar{X}$ and grows as X_0 moves away from the mean, so the bands flare out on both sides; forecasts outside the data range are less reliable.
- Bộ A tại $X_0 = 30$: $\hat{Y}_0 = 40,429$; khoảng trung bình (37,803; 43,056), khoảng cá biệt (32,545; 48,314) nghìn lốc.
Data set A at $X_0 = 30$: $\hat{Y}_0 = 40.429$; mean interval (37.803; 43.056), individual interval (32.545; 48.314) thousand packs.

Câu 3. So sánh các dạng hàm tuyến tính, log-log, log-lin, lin-log và nghịch đảo: cách viết mô hình, ý nghĩa của β_2 , hệ số co giãn. Có được dùng R^2 để chọn giữa mô hình tuyến tính và mô hình log-log không?

Question 3. Compare the linear, log-log, log-lin, lin-log and reciprocal forms: how each model is written, the meaning of β_2 and the elasticity. May R^2 be used to choose between the linear and the log-log model?

- Tuyến tính $Y = \beta_1 + \beta_2 X$: X tăng 1 đơn vị thì Y thay đổi β_2 đơn vị; co giãn $\beta_2 X/Y$ thay đổi dọc đường hồi quy.
Linear $Y = \beta_1 + \beta_2 X$: a one-unit rise in X changes Y by β_2 units; the elasticity $\beta_2 X/Y$ varies along the line.
- Log-log $\ln Y = \beta_1 + \beta_2 \ln X$: β_2 là hệ số co giãn không đổi; X tăng 1% thì Y thay đổi $\beta_2\%$.
Log-log $\ln Y = \beta_1 + \beta_2 \ln X$: β_2 is a constant elasticity; a 1% rise in X changes Y by $\beta_2\%$.
- Log-lin $\ln Y = \beta_1 + \beta_2 X$: X tăng 1 đơn vị thì Y thay đổi $100 \cdot \beta_2\%$ (chính xác $100 \cdot (\exp(\beta_2) - 1)\%$); khi X là thời gian, β_2 là tốc độ tăng trưởng.
Log-lin $\ln Y = \beta_1 + \beta_2 X$: a one-unit rise in X changes Y by $100 \cdot \beta_2\%$ (exactly $100 \cdot (\exp(\beta_2) - 1)\%$); when X is time, β_2 is the growth rate.
- Lin-log $Y = \beta_1 + \beta_2 \ln X$: X tăng 1% thì Y thay đổi $\beta_2/100$ đơn vị. Nghịch đảo $Y = \beta_1 + \beta_2/X$: Y tiến dần tới β_1 khi X tăng, dùng cho chi phí bình quân, đường Phillips.
Lin-log $Y = \beta_1 + \beta_2 \ln X$: a 1% rise in X changes Y by $\beta_2/100$ units. Reciprocal $Y = \beta_1 + \beta_2/X$: Y approaches β_1 as X grows, used for average cost and the Phillips curve.

- Không so R^2 giữa hai mô hình có biến phụ thuộc khác nhau (Y và $\ln Y$): R^2 đo tỷ lệ biến động của hai đại lượng khác nhau. Bộ A: R^2 tuyến tính 0,9258, R^2 log-log 0,8752, nhưng không vì vậy mà kết luận mô hình tuyến tính tốt hơn. Chọn dạng hàm theo lý thuyết kinh tế, dấu hệ số và kiểm định chỉ định (Chương 9).

R^2 must not be compared across models with different dependent variables (Y and $\ln Y$): each measures the share of variation in a different quantity. Data set A: linear R^2 0.9258, log-log R^2 0.8752, yet this does not show that the linear model is better. Choose the form from economic theory, coefficient signs and specification tests (Chapter 9).

3.14. Câu hỏi trắc nghiệm — Multiple-choice questions

17. Bộ A: $\beta_2 = -2,1127$, $se(\beta_2) = 0,2115$. Thống kê t của kiểm định $H_0: \beta_2 = 0$ là:

Data set A: $\beta_2 = -2.1127$, $se(\beta_2) = 0.2115$. The t statistic for $H_0: \beta_2 = 0$ is:

- A. -2,113 / -2.113
- B. 2,306 / 2.306
- C. -0,100 / -0.100
- D. -9,99 / -9.99

18. Bộ A ($n = 10$): $\beta_2 = -2,1127$, $se(\beta_2) = 0,2115$. Khoảng tin cậy 95% của β_2 là:

Data set A ($n = 10$): $\beta_2 = -2.1127$, $se(\beta_2) = 0.2115$. The 95% confidence interval for β_2 is:

- A. (-2,506; -1,719) / (-2.506; -1.719)
- B. (-2,600; -1,625) / (-2.600; -1.625)
- C. (-2,584; -1,642) / (-2.584; -1.642)
- D. (1,625; 2,600) / (1.625; 2.600)

19. Bộ A có $R^2 = 0,9258$, $n = 10$. Thống kê F của kiểm định sự phù hợp của mô hình là:

Data set A has $R^2 = 0.9258$, $n = 10$. The F statistic for overall significance is:

- A. 9,99 / 9.99
- B. 0,9258 / 0.9258
- C. 99,81 / 99.81
- D. 5,318 / 5.318

20. Với cùng X_0 và cùng độ tin cậy, khoảng dự báo giá trị cá biệt rộng hơn khoảng dự báo giá trị trung bình vì:

For the same X_0 and confidence level, the interval for an individual value is wider than that for the mean because:

- A. phương sai sai số dự báo cá biệt cộng thêm σ^2 của sai số ngẫu nhiên u_0 / *the individual forecast-error variance adds σ^2 of the disturbance u_0*
- B. dự báo cá biệt dùng bậc tự do n thay cho $n - 2$ / *the individual forecast uses n instead of $n - 2$ degrees of freedom*
- C. dự báo cá biệt dùng giá trị tới hạn z thay cho t / *the individual forecast uses z instead of t*
- D. $\hat{Y}_0$ của dự báo cá biệt lớn hơn $\bar{Y}_0$ của dự báo trung bình / *the individual $\hat{Y}_0$ is larger than the mean $\bar{Y}_0$*

- 21.** Bộ A ước lượng dạng log-log: $\ln \hat{Y} = 8,147 - 1,320 \cdot \ln X$ (X: giá, Y: lượng bán). Diễn giải đúng của hệ số góc là:
Data set A in log-log form: $\ln \hat{Y} = 8.147 - 1.320 \cdot \ln X$ (X: price, Y: quantity). The correct reading of the slope is:
- A.** giá tăng 1% thì lượng bán trung bình giảm khoảng 1,32% / *a 1% price rise lowers mean quantity by about 1.32%*
 - B.** giá tăng 1 nghìn đồng thì lượng bán giảm 1,32 nghìn lốc / *a price rise of 1 thousand VND lowers quantity by 1.32 thousand packs*
 - C.** giá tăng 1% thì lượng bán giảm 132% / *a 1% price rise lowers quantity by 132%*
 - D.** giá tăng 1% thì lượng bán giảm 0,0132 nghìn lốc / *a 1% price rise lowers quantity by 0.0132 thousand packs*
- 22.** Trong hồi quy hai biến có hệ số chặn, đổi đơn vị X từ nghìn đồng sang đồng ($X^* = 1.000X$), Y giữ nguyên. Kết quả nào đúng?
In a two-variable regression with an intercept, X is re-expressed from thousand VND to VND ($X^ = 1,000X$) and Y is unchanged. Which statement is correct?*
- A.** β_2 nhân 1.000, thống kê t của β_2 không đổi / *β_2 is multiplied by 1,000 and its t statistic is unchanged*
 - B.** β_2 và $se(\beta_2)$ cùng chia 1.000; thống kê t và R^2 không đổi / *β_2 and $se(\beta_2)$ are both divided by 1,000; the t statistic and R^2 are unchanged*
 - C.** β_1 chia 1.000, β_2 không đổi / *β_1 is divided by 1,000 and β_2 is unchanged*
 - D.** R^2 giảm vì X lớn hơn 1.000 lần / *R^2 falls because X is 1,000 times larger*
- 23.** Kiểm định ý kiến “giá tăng 1 nghìn đồng thì lượng bán giảm 2,5 nghìn lốc” trên bộ A: $t = 1,831$, $p\text{-value} = 0,1044$. Kết luận ở mức ý nghĩa 5% là:
Testing the claim “a price rise of 1 thousand VND lowers sales by 2.5 thousand packs” on data set A: $t = 1.831$, $p\text{-value} = 0.1044$. The conclusion at the 5% level is:
- A.** bác bỏ H_0 vì $p\text{-value}$ lớn hơn 0,05 / *reject H_0 because the p-value exceeds 0.05*
 - B.** bác bỏ H_0 vì thống kê t dương / *reject H_0 because the t statistic is positive*
 - C.** không kết luận được vì $p\text{-value}$ nằm giữa 0,05 và 0,1 / *no conclusion because the p-value lies between 0.05 and 0.1*
 - D.** chưa đủ cơ sở bác bỏ H_0 : số liệu không mâu thuẫn với ý kiến trên / *insufficient evidence to reject H_0 : the data do not contradict the claim*
- 24.** Bộ A ($n = 10$), kiểm định “giảm giá có làm tăng lượng bán không”, $H_0: \beta_2 \geq 0$, $H_1: \beta_2 < 0$, $\alpha = 5\%$. Quy tắc bác bỏ H_0 là:
Data set A ($n = 10$), testing “does a price cut raise sales”, $H_0: \beta_2 \geq 0$, $H_1: \beta_2 < 0$, $\alpha = 5\%$. The rule for rejecting H_0 is:
- A.** $|t| > 2,306$ / *$|t| > 2.306$*
 - B.** $t > 1,860$ / *$t > 1.860$*
 - C.** $t < -1,860$ / *$t < -1.860$*
 - D.** $t < -1,812$ / *$t < -1.812$*

Chương 4. Hồi quy bội — Chapter 4. Multiple regression

Mục tiêu chương — Chapter objectives

- Viết mô hình hồi quy k biến ở dạng thông thường và dạng ma trận; diễn giải hệ số hồi quy riêng với điều kiện các biến khác không đổi.
Write the k -variable regression model in scalar and matrix form; interpret partial regression coefficients with the other regressors held constant.
- Tính $\beta = (X'X)^{-1}X'Y$, σ^2 , $\text{Var}(\beta)$ cho mô hình ba biến bằng ma trận và bằng công thức độ lệch.
Compute $\beta = (X'X)^{-1}X'Y$, σ^2 and $\text{Var}(\beta)$ for a three-variable model with matrices and with deviation formulas.
- Tính R^2 , R^2 hiệu chỉnh; kiểm định t từng hệ số, kiểm định F sự phù hợp, kiểm định Wald thu hẹp và kiểm định ràng buộc tuyến tính; lập khoảng tin cậy và khoảng dự báo.
Compute R^2 and adjusted R^2 ; carry out t -tests, the overall F -test, the Wald test for dropping variables and tests of linear restrictions; build confidence and prediction intervals.
- Ước lượng hàm Cobb–Douglas dạng log và kiểm định hiệu quả không đổi theo quy mô.
Estimate a log-linear Cobb–Douglas function and test constant returns to scale.
- Đọc bảng kết quả EViews của hồi quy bội và điền các ô bị khuyết như câu 2 của đề thi.
Read an EViews multiple-regression output and fill in missing cells as in exam question 2.

GHI CHÚ / NOTE — Dữ liệu dùng trong chương. Bộ B (dùng chung cả sách): 36 cửa hàng cà phê Hạt Nâu, tháng 9/2026; Y doanh thu tháng (triệu đồng), GIA giá bình quân một ly (nghìn đồng), QC chi quảng cáo trên mạng xã hội (triệu đồng/tháng), DT diện tích phục vụ (m^2). Các mục dùng bộ B chỉ đưa vào ba biến định lượng; biến TT và KV vào mô hình ở Chương 5. Các bộ số nhỏ khác là dữ kiện đề bài, cho ngay tại ví dụ.

Data used. Data set B (shared across the book): 36 Hạt Nâu coffee shops, September 2026; Y monthly revenue (million VND), GIA average price per cup (thousand VND), QC social-media advertising (million VND/month), DT serving area (m^2). Wherever data set B is used, only the three quantitative regressors enter; TT and KV enter in Chapter 5. The other small data sets are given with each example.

4.1. Mô hình hồi quy bội và ý nghĩa hệ số hồi quy riêng — The multiple regression model and partial regression coefficients

Hồi quy bội đưa nhiều biến giải thích vào cùng một phương trình. Doanh thu của một cửa hàng cà phê phụ thuộc đồng thời vào giá, quảng cáo, diện tích; hồi quy doanh thu theo riêng giá sẽ gán cho giá cả phần tác động của các biến bị bỏ ra ngoài nếu chúng tương quan với giá. Hàm hồi quy tổng thể với k hệ số ($k - 1$ biến giải thích và một hệ số chặn):

Multiple regression puts several explanatory variables into one equation. A coffee shop's revenue depends on price, advertising and floor area at the same time; regressing revenue on price alone attributes to price the effect of omitted variables whenever they are correlated with

price. The population regression function with k coefficients ($k - 1$ regressors and an intercept) is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i, \quad i = 1, \dots, n \quad (4.1)$$

Hàm hồi quy mẫu và phần dư:

The sample regression function and the residuals:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}; \quad e_i = Y_i - \hat{Y}_i \quad (4.2)$$

Quy ước trong sách: k là số hệ số phải ước lượng, **kể cả hệ số chặn**. Mô hình Y theo GIA, QC, DT có $k = 4$. Một số giáo trình dùng k là số biến giải thích; khi đó bậc tự do viết $n - k - 1$. Đọc kỹ đề để dùng đúng một quy ước.

Convention used in this book: k is the number of coefficients to estimate, including the intercept. The model of Y on GIA, QC and DT has $k = 4$. Some textbooks let k be the number of regressors and write the degrees of freedom as $n - k - 1$. Read the question and use one convention consistently.

Hệ số hồi quy riêng β_j đo mức thay đổi của giá trị trung bình của Y khi X_j tăng 1 đơn vị với điều kiện các biến giải thích khác không đổi. Hệ số chặn β_1 là giá trị trung bình của Y khi mọi X bằng 0; chỉ diễn giải khi điểm $X = 0$ nằm gần vùng số liệu. Cách đọc một hệ số luôn gồm ba phần: chiều (tăng/giảm), độ lớn kèm đơn vị của Y và của X_j , và cụm "các yếu tố khác không đổi".

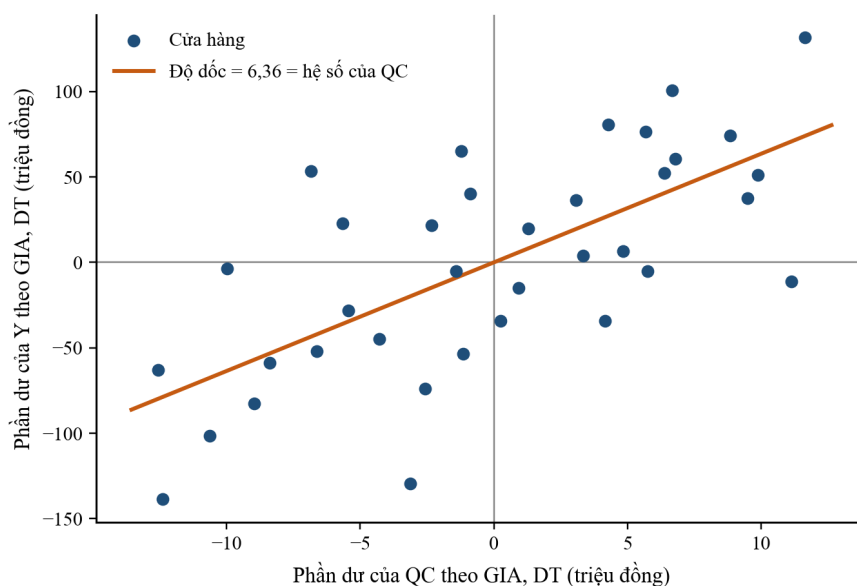
A partial regression coefficient β_j measures the change in the mean of Y when X_j rises by one unit, holding the other regressors constant. The intercept β_1 is the mean of Y when every X is zero and is interpreted only if $X = 0$ lies near the data. A coefficient reading always has three parts: direction (rise/fall), size with the units of Y and X_j , and the phrase 'other things equal'.

Giả thiết của mô hình hồi quy tuyến tính cổ điển ở Chương 2 vẫn giữ nguyên (kỳ vọng sai số bằng 0, phương sai không đổi, không tự tương quan, X không ngẫu nhiên hoặc không tương quan với u). Hồi quy bội thêm hai điều kiện: **không có đa cộng tuyến hoàn hảo** giữa các X (không biến nào là tổ hợp tuyến tính đúng của các biến khác) và $n > k$. Thiếu một trong hai điều kiện thì $X'X$ không khả nghịch, OLS không cho nghiệm duy nhất.

*The classical linear regression assumptions of Chapter 2 still hold (zero-mean disturbances, constant variance, no autocorrelation, X non-random or uncorrelated with u). Multiple regression adds two conditions: **no perfect multicollinearity** among the regressors (none is an exact linear combination of the others) and $n > k$. If either fails, $X'X$ is singular and OLS has no unique solution.*

VÍ DỤ / EXAMPLE — Hồi quy đơn và hồi quy bội cho hai câu trả lời khác nhau. Hồi quy doanh thu Hạt Nâu theo riêng QC cho hệ số 7,122 triệu đồng; hồi quy theo GIA, QC, DT cho hệ số của QC bằng 6,356. Hệ số trong hồi quy bội bằng độ dốc của đường hồi quy giữa hai chuỗi phần dư: phần dư của Y sau khi hồi quy theo GIA, DT và phần dư của QC sau khi hồi quy theo GIA, DT (định lý Frisch–Waugh–Lovell, Hình 4.1). Hai chuỗi phần dư là phần của Y và của QC không giải thích được bằng giá và diện tích; độ dốc giữa chúng là tác động của QC khi giá và diện tích đã được giữ cố định.

Simple and multiple regression give different answers. Regressing Hạt Nâu revenue on QC alone gives a coefficient of 7.122; regressing on GIA, QC and DT gives a QC coefficient of 6.356. The multiple-regression coefficient equals the slope between two residual series: the residuals of Y on GIA and DT and the residuals of QC on GIA and DT (the Frisch–Waugh–Lovell theorem, Figure 4.1). Both series are what price and area cannot explain; the slope between them is the effect of QC with price and area held fixed.



Hình 4.1. Hệ số của QC trong hồi quy bội là độ dốc giữa phần dư của Y và phần dư của QC sau khi loại tác động của GIA và DT (bộ B)

Figure 4.1. The QC coefficient in the multiple regression is the slope between the residuals of Y and of QC after removing the effects of GIA and DT (data set B)

Nội dung <i>Item</i>	Hồi quy hai biến <i>Simple regression</i>	Hồi quy bội (k hệ số) <i>Multiple regression (k coefficients)</i>
Ý nghĩa hệ số góc <i>Slope meaning</i>	Tác động tổng của X <i>Total effect of X</i>	Tác động riêng, các X khác không đổi <i>Partial effect, other X fixed</i>
Bậc tự do của RSS <i>Residual df</i>	$n - 2$	$n - k$
σ^2	$RSS/(n - 2)$	$RSS/(n - k)$
Kiểm định F <i>F-test</i>	$F(1, n - 2) = t^2$	$F(k - 1, n - k)$, khác t^2 <i>$F(k - 1, n - k)$, not t^2</i>
So sánh mô hình <i>Model comparison</i>	R^2	R^2 hiệu chỉnh <i>Adjusted R^2</i>
Điều kiện thêm <i>Extra condition</i>	X có biến thiên <i>X varies</i>	Không đa cộng tuyến hoàn hảo, $n > k$ <i>No perfect multicollinearity, $n > k$</i>

4.2. Dạng ma trận và ước lượng bình phương nhỏ nhất — *Matrix form and least-squares estimation*

Viết n phương trình (4.1) chồng lên nhau được dạng ma trận. Y là vector $n \times 1$ các quan sát của biến phụ thuộc; X là ma trận $n \times k$, cột đầu toàn số 1 (ứng với hệ số chặn), mỗi cột sau là một biến giải thích; β là vector $k \times 1$ các hệ số; U là vector $n \times 1$ các sai số ngẫu nhiên.

Stacking the n equations (4.1) gives the matrix form. Y is the $n \times 1$ vector of observations on the dependent variable; X is the $n \times k$ matrix whose first column is all ones (for the intercept) and whose other columns are the regressors; β is the $k \times 1$ coefficient vector; U is the $n \times 1$ vector of disturbances.

$$Y = X\beta + U \quad (4.3)$$

$$X = \begin{bmatrix} 1 & X_{21} & X_{31} & \dots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{2n} & X_{3n} & \dots & X_{kn} \end{bmatrix}$$

OLS chọn β để tổng bình phương phần dư $e'e = (Y - X\beta)'(Y - X\beta)$ nhỏ nhất. Lấy đạo hàm theo β và cho bằng 0 được hệ phương trình chuẩn $X'X\beta = X'Y$. Khi $X'X$ khả nghịch: *OLS chooses β to minimise the residual sum of squares $e'e = (Y - X\beta)'(Y - X\beta)$. Setting the derivative with respect to β to zero gives the normal equations $X'X\beta = X'Y$. When $X'X$ is invertible:*

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (4.4)$$

Với giả thiết cổ điển, $\hat{\beta}$ không chệch và là ước lượng tuyến tính không chệch tốt nhất (BLUE, định lý Gauss–Markov). Ma trận hiệp phương sai của $\hat{\beta}$ và ước lượng của σ^2 :

Under the classical assumptions $\hat{\beta}$ is unbiased and is the best linear unbiased estimator (BLUE, Gauss–Markov theorem). The covariance matrix of $\hat{\beta}$ and the estimator of σ^2 are:

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}; \quad \hat{\sigma}^2 = e'e/(n - k) = \text{RSS}/(n - k) \quad (4.5)$$

$$\text{RSS} = Y'Y - \hat{\beta}'X'Y; \quad \text{TSS} = Y'Y - n\bar{Y}^2; \quad \text{ESS} = \hat{\beta}'X'Y - n\bar{Y}^2 \quad (4.6)$$

Phần tử thứ j trên đường chéo của $\sigma^2(X'X)^{-1}$ là $\text{Var}(\hat{\beta}_j)$, căn bậc hai là $\text{se}(\hat{\beta}_j)$. Phần tử ngoài đường chéo ở hàng i , cột j là $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$, cần cho kiểm định ràng buộc ở mục 4.10.

The j -th diagonal element of $\sigma^2(X'X)^{-1}$ is $\text{Var}(\hat{\beta}_j)$ and its square root is $\text{se}(\hat{\beta}_j)$. The off-diagonal element in row i , column j is $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$, needed for the restriction tests in section 4.10.

Bài tính bằng ma trận: chuỗi điện thoại Ánh Dương — *Worked matrix example: the Ánh Dương phone chain*

Dữ kiện đề bài: số liệu tháng 9/2026 của 7 cửa hàng thuộc chuỗi điện thoại di động Ánh Dương. Y là doanh số (chục triệu đồng/tháng), X_2 là số nhân viên bán hàng (người), X_3 là chi quảng cáo (triệu đồng/tháng). Ước lượng $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ bằng ma trận; tính R^2 , R^2 hiệu chỉnh, se và kiểm định ở các mục 4.3–4.6.

Problem data: September 2026 figures for 7 shops of the Ánh Dương mobile-phone chain. Y is sales (tens of millions of VND a month), X2 the number of sales staff, X3 advertising spending (million VND a month). Estimate $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ with matrices; R^2 , adjusted R^2 , standard errors and tests follow in sections 4.3–4.6.

Cửa hàng Shop	Y	X2	X3	Ŷ	e	e ²
1	34	7	6	34,50 34.50	-0,50 -0.50	0,2500 0.2500
2	27	6	4	29,00 29.00	-2,00 -2.00	4,0000 4.0000
3	33	6	7	32,60 32.60	0,40 0.40	0,1600 0.1600
4	25	5	4	25,90 25.90	-0,90 -0.90	0,8100 0.8100
5	31	6	5	30,20 30.20	0,80 0.80	0,6400 0.6400
6	24	4	4	22,80 22.80	1,20 1.20	1,4400 1.4400
7	35	8	3	34,00 34.00	1,00 1.00	1,0000 1.0000
Tổng Sum	209	42	33	209,00 209.00	0,00 0.00	8,3000 8.3000

1. Lập $X'X$ (tổng các tích chéo) và $X'Y$. Hàng 1 của $X'X$ là $(n; \sum X_2; \sum X_3)$, hàng 2 là $(\sum X_2; \sum X_2^2; \sum X_2 X_3)$, hàng 3 là $(\sum X_3; \sum X_2 X_3; \sum X_3^2)$; $X'Y = (\sum Y; \sum X_2 Y; \sum X_3 Y)'$.
Form $X'X$ (sums of cross-products) and $X'Y$. Row 1 of $X'X$ is $(n; \sum X_2; \sum X_3)$, row 2 $(\sum X_2; \sum X_2^2; \sum X_2 X_3)$, row 3 $(\sum X_3; \sum X_2 X_3; \sum X_3^2)$; $X'Y = (\sum Y; \sum X_2 Y; \sum X_3 Y)'$.

$$X'X = \begin{bmatrix} 7 & 42 & 33 \\ 42 & 262 & 198 \\ 33 & 198 & 167 \end{bmatrix} \quad X'Y = (209; 1285; 999)'$$

2. Định thức, khai triển theo hàng 1: $\det(X'X) = 7 \cdot 4550 - 42 \cdot 480 + 33 \cdot (-330) = 800$.
 Định thức khác 0 nên $X'X$ khả nghịch.
Determinant by the first row: $\det(X'X) = 7 \cdot 4550 - 42 \cdot 480 + 33 \cdot (-330) = 800$. It is non-zero, so $X'X$ is invertible.
3. Ma trận nghịch đảo $(X'X)^{-1} = (1/\det) \cdot$ ma trận phụ hợp; ma trận phụ hợp là chuyển vị của ma trận các phần bù đại số. $X'X$ đối xứng nên $(X'X)^{-1}$ cũng đối xứng.
Inverse $(X'X)^{-1} = (1/\det) \cdot$ adjugate; the adjugate is the transpose of the cofactor matrix. $X'X$ is symmetric, so its inverse is symmetric too.

$$(X'X)^{-1} = \frac{1}{800} \begin{bmatrix} 4550 & -480 & -330 \\ -480 & 80 & 0 \\ -330 & 0 & 70 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 5,6875 & -0,6000 & -0,4125 \\ -0,6000 & 0,1000 & 0,0000 \\ -0,4125 & 0,0000 & 0,0875 \end{bmatrix}$$

4. $\beta = (X'X)^{-1}X'Y$. Hàng 1: $5,6875 \cdot 209 - 0,6 \cdot 1285 - 0,4125 \cdot 999 = 5,6$. Tương tự hàng 2, hàng 3.

$\beta = (X'X)^{-1}X'Y$. Row 1: $5.6875 \cdot 209 - 0.6 \cdot 1285 - 0.4125 \cdot 999 = 5.6$. Rows 2 and 3 likewise.

$$\hat{\beta} = \begin{bmatrix} 5,6 \\ 3,1 \\ 1,2 \end{bmatrix} \quad \hat{Y} = 5,6 + 3,1X_2 + 1,2X_3$$

5. Diễn giải: giữ chi quảng cáo không đổi, thêm 1 nhân viên bán hàng thì doanh số trung bình tăng 3,1 chục triệu đồng (31 triệu đồng)/tháng; giữ số nhân viên không đổi, chi thêm 1 triệu đồng quảng cáo thì doanh số trung bình tăng 1,2 chục triệu đồng (12 triệu đồng)/tháng. Hệ số chặn 5,6 không có ý nghĩa kinh tế: không có cửa hàng nào hoạt động với 0 nhân viên.

Interpretation: holding advertising constant, one more salesperson raises mean sales by 3.1 tens of millions (VND 31 million) a month; holding staff constant, VND 1 million more advertising raises mean sales by 1.2 tens of millions (VND 12 million) a month. The intercept 5.6 has no economic meaning: no shop operates with zero staff.

6. Tổng bình phương: $Y'Y = 6.361$; $\beta'X'Y = 6.352,7$; $RSS = 6.361 - 6.352,7 = 8,3$ (khớp tổng cột e^2 trong bảng). $\sigma^2 = 8,3/(7 - 3) = 2,0750$.

Sums of squares: $Y'Y = 6,361$; $\beta'X'Y = 6,352.7$; $RSS = 6,361 - 6,352.7 = 8.3$ (matching the e^2 column). $\sigma^2 = 8.3/(7 - 3) = 2.0750$.

7. Ma trận hiệp phương sai ước lượng: $\sigma^2(X'X)^{-1}$. Lấy căn các phần tử đường chéo được se.

Estimated covariance matrix $\sigma^2(X'X)^{-1}$; the square roots of its diagonal are the standard errors.

$$\text{Var}(\hat{\beta}) = \begin{bmatrix} 11,8016 & -1,2450 & -0,8559 \\ -1,2450 & 0,2075 & 0,0000 \\ -0,8559 & 0,0000 & 0,1816 \end{bmatrix} \quad \text{se} = (3,4353; 0,4555; 0,4261)$$

Phần tử ở hàng 2, cột 3 bằng 0: trong mẫu này $\sum X_2X_3 - n\bar{X}_2\bar{X}_3 = 0$, tức X_2 và X_3 không tương quan, nên β_2 và β_3 không tương quan. Đây là trường hợp đặc biệt của số liệu tự ra; số liệu thực tế hầu như luôn có $\text{Cov}(\beta_2, \beta_3) \neq 0$.

The element in row 2, column 3 is 0: here $\sum X_2X_3 - n\bar{X}_2\bar{X}_3 = 0$, so X_2 and X_3 are uncorrelated and so are β_2 and β_3 . This is a feature of the constructed data; real data almost always give $\text{Cov}(\beta_2, \beta_3) \neq 0$.

Công thức dạng độ lệch cho mô hình ba biến — *Deviation-form formulas for the three-variable model*

Đề thi thường cho sẵn các tổng độ lệch (chữ thường: $x_{2i} = X_{2i} - \bar{X}_2$, $x_{3i} = X_{3i} - \bar{X}_3$, $y_i = Y_i - \bar{Y}$) và yêu cầu tính tay. Giải hệ phương trình chuẩn sau khi khử hệ số chặn được:

Exams often give deviation sums (lower case: $x_{2i} = X_{2i} - \bar{X}_2$, $x_{3i} = X_{3i} - \bar{X}_3$, $y_i = Y_i - \bar{Y}$) and ask for a hand calculation. Solving the normal equations after eliminating the intercept gives:

$$\hat{\beta}_2 = (\sum yx_2 \cdot \sum x_3^2 - \sum yx_3 \cdot \sum x_2x_3) / (\sum x_2^2 \cdot \sum x_3^2 - (\sum x_2x_3)^2) \quad (4.7)$$

$$\hat{\beta}_3 = (\sum yx_3 \cdot \sum x_2^2 - \sum yx_2 \cdot \sum x_2x_3) / (\sum x_2^2 \cdot \sum x_3^2 - (\sum x_2x_3)^2) \quad (4.8)$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}_2 - \hat{\beta}_3\bar{X}_3 \quad (4.9)$$

$$\text{Var}(\hat{\beta}_2) = \sigma^2 \sum x_3^2 / D; \quad \text{Var}(\hat{\beta}_3) = \sigma^2 \sum x_2^2 / D; \quad D = \sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2 \quad (4.10)$$

VÍ DỤ / EXAMPLE — Kiểm lại bài Ánh Dương bằng công thức độ lệch. $\sum x_2^2 = 10$; $\sum x_3^2 = 11,4286$; $\sum x_2x_3 = 0$; $\sum x_2y = 31$; $\sum x_3y = 13,7143$. Vì $\sum x_2x_3 = 0$ nên (4.7) rút gọn thành $\hat{\beta}_2 = \sum x_2y / \sum x_2^2 = 31/10 = 3,1$ và $\hat{\beta}_3 = \sum x_3y / \sum x_3^2 = 1,2$, trùng kết quả ma trận. $\hat{\beta}_1 = 29,8571 - 3,1 \cdot 6 - 1,2 \cdot 4,7143 = 5,6$. $\text{Var}(\hat{\beta}_2) = 2,0750/10 = 0,2075$, đúng phần tử (2, 2) của ma trận $\text{Var}(\hat{\beta})$.

Checking Ánh Dương with deviation formulas. $\sum x_2^2 = 10$; $\sum x_3^2 = 11.4286$; $\sum x_2x_3 = 0$; $\sum x_2y = 31$; $\sum x_3y = 13.7143$. Since $\sum x_2x_3 = 0$, (4.7) reduces to $\hat{\beta}_2 = \sum x_2y / \sum x_2^2 = 31/10 = 3.1$ and $\hat{\beta}_3 = \sum x_3y / \sum x_3^2 = 1.2$, the matrix result. $\hat{\beta}_1 = 29.8571 - 3.1 \cdot 6 - 1.2 \cdot 4.7143 = 5.6$. $\text{Var}(\hat{\beta}_2) = 2.0750/10 = 0.2075$, the (2, 2) element of $\text{Var}(\hat{\beta})$.

LỖI HAY GẶP / COMMON MISTAKE — Nhân ma trận sai thứ tự: viết $\hat{\beta} = X'Y(X'X)^{-1}$ hoặc $(X'X)^{-1} = (X')^{-1}X^{-1}$. X không vuông nên không có X^{-1} ; phép nhân ma trận không giao hoán. Luôn kiểm cỡ: $(k \times k)(k \times n)(n \times 1) = k \times 1$.

Multiplying in the wrong order: writing $\hat{\beta} = X'Y(X'X)^{-1}$ or $(X'X)^{-1} = (X')^{-1}X^{-1}$. X is not square, so X^{-1} does not exist, and matrix multiplication is not commutative. Always check sizes: $(k \times k)(k \times n)(n \times 1) = k \times 1$.

4.3. Hệ số xác định R^2 và R^2 hiệu chỉnh — *R-squared and adjusted R-squared*

Phân tích $TSS = ESS + RSS$ giữ nguyên như hồi quy hai biến. R^2 là tỷ lệ biến động của Y được các biến giải thích (cùng nhau) giải thích:

The decomposition $TSS = ESS + RSS$ is the same as in simple regression. R^2 is the share of the variation in Y explained jointly by the regressors:

$$R^2 = ESS/TSS = 1 - RSS/TSS; \quad \text{ba biến: } ESS = \hat{\beta}_2 \sum yx_2 + \hat{\beta}_3 \sum yx_3 \quad (4.11)$$

R^2 không giảm khi thêm biến, kể cả biến vô nghĩa: OLS với thêm một cột trong X luôn đạt RSS nhỏ hơn hoặc bằng (nếu hệ số của biến mới bằng 0 thì RSS giữ nguyên). Vì vậy R^2 không dùng được để so sánh các mô hình có số biến khác nhau. R^2 hiệu chỉnh (ký hiệu $\bar{R}^2$) chia RSS và TSS cho bậc tự do tương ứng:

R^2 never falls when a regressor is added, even a meaningless one: OLS with an extra column in X always reaches an RSS no larger (if the new coefficient were zero, RSS would be unchanged). So R^2 cannot compare models with different numbers of regressors. Adjusted R^2 (written $\bar{R}^2$) divides RSS and TSS by their degrees of freedom:

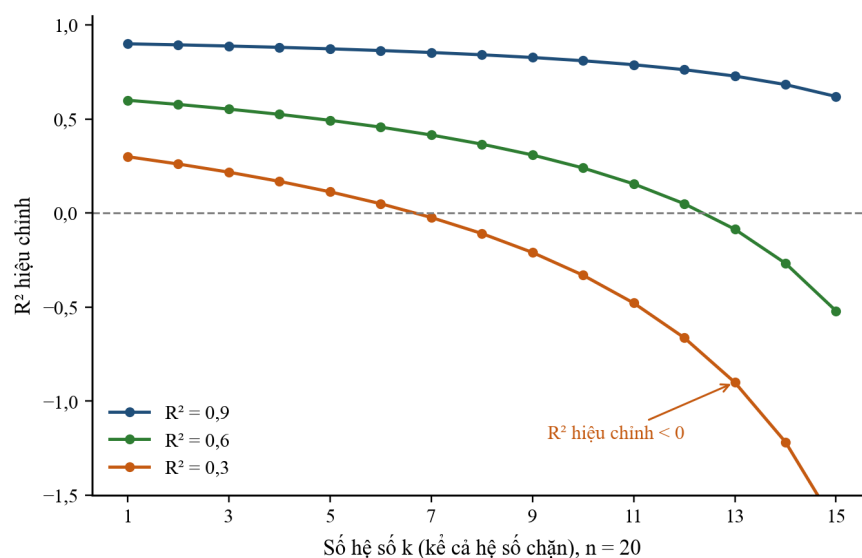
$$R^2_{\text{adj}} = 1 - [\text{RSS}/(n - k)] / [\text{TSS}/(n - 1)] = 1 - (1 - R^2)(n - 1)/(n - k) \quad (4.12)$$

Tính chất: $\bar{R}^2 \leq R^2$; thêm một biến làm $\bar{R}^2$ tăng khi và chỉ khi $|t|$ của biến đó lớn hơn 1; $\bar{R}^2$ có thể âm khi R^2 nhỏ và k lớn so với n (Hình 4.2). $\bar{R}^2$ không phải "tỷ lệ giải thích", chỉ là thước đo để so sánh mô hình có cùng biến phụ thuộc.

Properties: adjusted $R^2 \leq R^2$; adding a regressor raises adjusted R^2 if and only if its $|t|$ exceeds 1; adjusted R^2 can be negative when R^2 is small and k is large relative to n (Figure 4.2). It is not an 'explained share', only a yardstick for comparing models with the same dependent variable.

VÍ DỤ / EXAMPLE — Ánh Dương. $\bar{Y} = 29,8571$; $\text{TSS} = Y'Y - n\bar{Y}^2 = 6.361 - 7 \cdot 29,8571^2 = 120,8571$; $\text{ESS} = \text{TSS} - \text{RSS} = 112,5571$. $R^2 = 1 - 8,3/120,8571 = \mathbf{0,9313}$: số nhân viên và chi quảng cáo giải thích 93,13% biến động doanh số giữa các cửa hàng. $\bar{R}^2 = 1 - (1 - 0,9313) \cdot 6/4 = \mathbf{0,8970}$. Với $n = 7$, mỗi hệ số thêm vào làm mất một trong bốn bậc tự do còn lại, nên khoảng cách $R^2 - \bar{R}^2$ (0,0343) lớn hơn so với mẫu lớn.

Ánh Dương. $\bar{Y} = 29.8571$; $\text{TSS} = Y'Y - n\bar{Y}^2 = 6.361 - 7 \cdot 29.8571^2 = 120.8571$; $\text{ESS} = \text{TSS} - \text{RSS} = 112.5571$. $R^2 = 1 - 8.3/120.8571 = \mathbf{0.9313}$: staff numbers and advertising explain 93.13% of the variation in sales across shops. Adjusted $R^2 = 1 - (1 - 0.9313) \cdot 6/4 = \mathbf{0.8970}$. With $n = 7$ each extra coefficient uses one of only four remaining degrees of freedom, so the gap $R^2 - \text{adjusted } R^2$ (0.0343) is wider than in a large sample.



Hình 4.2. R^2 hiệu chỉnh giảm khi số hệ số tăng mà R^2 giữ nguyên ($n = 20$); với $R^2 = 0,3$ và $k \geq 8$, R^2 hiệu chỉnh âm

Figure 4.2. Adjusted R^2 falls as the number of coefficients rises with R^2 unchanged ($n = 20$); with $R^2 = 0.3$ and $k \geq 8$ it is negative

Tiêu chí <i>Criterion</i>	R^2	R^2 hiệu chỉnh <i>Adjusted R^2</i>
Công thức <i>Formula</i>	$1 - \text{RSS}/\text{TSS}$	$1 - (1 - R^2)(n - 1)/(n - k)$
Miền giá trị <i>Range</i>	$[0; 1]$ khi có hệ số chặn <i>$[0, 1]$ with an intercept</i>	$\leq R^2$, có thể âm <i>$\leq R^2$, may be negative</i>
Thêm biến <i>Adding a regressor</i>	Không giảm <i>Never falls</i>	Tăng khi $ t > 1$, giảm khi $ t < 1$ <i>Rises if $t > 1$, falls if $t < 1$</i>
Diễn giải <i>Meaning</i>	Tỷ lệ biến động Y được giải thích <i>Share of variation in Y explained</i>	Không diễn giải là tỷ lệ <i>Not a share</i>
Dùng để <i>Used for</i>	Mô tả độ phù hợp, tính F <i>Describing fit, computing F</i>	So sánh mô hình khác số biến <i>Comparing models with different k</i>

LỖI HAY GẶP / COMMON MISTAKE — So sánh R^2 của mô hình Y với mô hình $\ln Y$, hoặc của hai mô hình ước lượng trên hai mẫu khác nhau. TSS khác nhau nên hai R^2 không cùng thước đo. Cũng sai khi viết $\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1)$ trong khi k đã gồm hệ số chặn.

Comparing the R^2 of a model in Y with one in $\ln Y$, or of models estimated on different samples. TSS differs, so the two R^2 values are not on the same scale. It is also wrong to write adjusted $R^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1)$ when k already counts the intercept.

4.4. Kiểm định t từng hệ số và khoảng tin cậy — *t-tests on individual coefficients and confidence intervals*

Khi sai số có phân phối chuẩn, $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ có phân phối t với $n - k$ bậc tự do. Kiểm định giả thuyết $H_0: \beta_j = \beta^*$ dùng thống kê:

When the disturbances are normal, $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ follows a t distribution with $n - k$ degrees of freedom. The test of $H_0: \beta_j = \beta^$ uses:*

$$t = (\hat{\beta}_j - \beta^*) / \text{se}(\hat{\beta}_j) \sim t(n - k) \quad (4.13)$$

Cặp giả thuyết <i>Hypotheses</i>	Bác bỏ H_0 khi <i>Reject H_0 if</i>	Ví dụ câu hỏi <i>Typical question</i>
$H_0: \beta_j = \beta^*; H_1: \beta_j \neq \beta^*$	$ t > t(\alpha/2; n - k)$	X_j có ảnh hưởng tới Y không? <i>Does X_j affect Y?</i>
$H_0: \beta_j \leq \beta^*; H_1: \beta_j > \beta^*$	$t > t(\alpha; n - k)$	Tăng X_j có làm Y tăng không? <i>Does raising X_j raise Y?</i>
$H_0: \beta_j \geq \beta^*; H_1: \beta_j < \beta^*$	$t < -t(\alpha; n - k)$	Tăng giá có làm lượng bán giảm không? <i>Does a higher price lower sales?</i>

Khoảng tin cậy $1 - \alpha$ cho β_j :

The $1 - \alpha$ confidence interval for β_j :

$$\hat{\beta}_j - t_{\alpha/2}^{(n-k)} \cdot \text{se}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2}^{(n-k)} \cdot \text{se}(\hat{\beta}_j) \quad (4.14)$$

VÍ DỤ / EXAMPLE — Ảnh Dương, $\alpha = 5\%$. Bậc tự do $n - k = 7 - 3 = 4$; $t(0,025; 4) = 2,776$. $H_0: \beta_2 = 0$: $t = 3,1/0,4555 = 6,805 > 2,776$, bác bỏ H_0 , số nhân viên có ảnh hưởng tới doanh số. $H_0: \beta_3 = 0$: $t = 1,2/0,4261 = 2,816 > 2,776$, bác bỏ H_0 . Khoảng tin cậy 95%: $\beta_2 \in (1,8355; 4,3645)$; $\beta_3 \in (0,0171; 2,3829)$ chục triệu đồng. Với hệ số chặn: $t = 1,630 < 2,776$, chưa đủ cơ sở bác bỏ H_0 : $\beta_1 = 0$, nhưng không vì thế mà bỏ hệ số chặn khỏi mô hình.

Ảnh Dương, $\alpha = 5\%$. Degrees of freedom $n - k = 7 - 3 = 4$; $t(0.025; 4) = 2.776$. $H_0: \beta_2 = 0$: $t = 3.1/0.4555 = 6.805 > 2.776$; reject H_0 , staff numbers affect sales. $H_0: \beta_3 = 0$: $t = 1.2/0.4261 = 2.816 > 2.776$; reject H_0 . 95% intervals: $\beta_2 \in (1.8355; 4.3645)$; $\beta_3 \in (0.0171; 2.3829)$ tens of millions. For the intercept $t = 1.630 < 2.776$: $H_0: \beta_1 = 0$ cannot be rejected, but that is no reason to drop the intercept.

Ba cách kết luận cho cùng đáp án: so $|t|$ với giá trị tới hạn; xem khoảng tin cậy có chứa β^* không; so p-value (cột Prob. của phần mềm) với α . Prob. trong bảng EViews là p-value hai phía; kiểm định một phía đúng chiều thì lấy Prob./2.

Three routes give the same answer: compare $|t|$ with the critical value; check whether the confidence interval contains β^ ; compare the p-value (the software's Prob. column) with α . EViews prints two-sided p-values; for a one-sided test in the right direction use Prob./2.*

LỖI HAY GẶP / COMMON MISTAKE — Dùng bậc tự do $n - 2$ (thói quen của hồi quy hai biến) hoặc $n - 1$ cho hồi quy bội. Với $n = 7$, $k = 3$: $t(0,025; 4) = 2,776$, còn $t(0,025; 5) = 2,571$; lấy sai bậc tự do làm khoảng tin cậy và kết luận sai theo.

Using $n - 2$ (the simple-regression habit) or $n - 1$ degrees of freedom in multiple regression. With $n = 7$, $k = 3$: $t(0.025; 4) = 2.776$, while $t(0.025; 5) = 2.571$; the wrong df distorts the interval and possibly the conclusion.

4.5. Kiểm định F sự phù hợp của mô hình — *The F-test of overall significance*

Mô hình "phù hợp" khi ít nhất một biến giải thích có ảnh hưởng tới Y . Cặp giả thuyết: $H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$ (tương đương $R^2 = 0$ trong tổng thể); H_1 : có ít nhất một $\beta_j \neq 0$ ($j = 2, \dots, k$). Thống kê kiểm định:

A model is 'significant' when at least one regressor affects Y . Hypotheses: $H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$ (equivalently a population R^2 of zero); H_1 : at least one $\beta_j \neq 0$ ($j = 2, \dots, k$). The statistic is:

$$F = [ESS/(k - 1)] / [RSS/(n - k)] = [R^2/(k - 1)] / [(1 - R^2)/(n - k)] \sim F(k - 1, n - k) \quad (4.15)$$

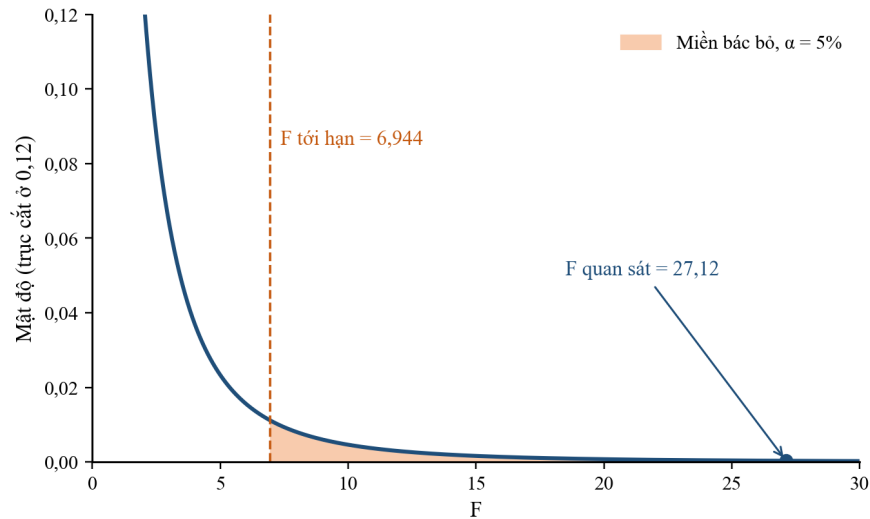
Bác bỏ H_0 khi $F > F(\alpha; k - 1, n - k)$ hoặc $\text{Prob}(F\text{-statistic}) < \alpha$. Dạng theo R^2 cho phép tính F khi đề chỉ cho R^2 , n , k . F tăng khi R^2 tăng: kiểm định F cũng là kiểm định R^2 tổng thể có khác 0 không.

Reject H_0 when $F > F(\alpha; k - 1, n - k)$ or $\text{Prob}(F\text{-statistic}) < \alpha$. The R^2 form lets you compute F when only R^2 , n and k are given. F rises with R^2 : the F -test is also a test of whether the population R^2 is zero.

Nguồn biến động <i>Source</i>	Tổng bình phương <i>Sum of squares</i>	Bậc tự do <i>df</i>	Bình phương trung bình <i>Mean square</i>	F
Hồi quy (ESS) <i>Regression (ESS)</i>	112,5571 <i>112.5571</i>	2	56,2786 <i>56.2786</i>	27,122 <i>27.122</i>
Phần dư (RSS) <i>Residual (RSS)</i>	8,3000 <i>8.3000</i>	4	2,0750 <i>2.0750</i>	
Tổng (TSS) <i>Total (TSS)</i>	120,8571 <i>120.8571</i>	6		

VÍ DỤ / EXAMPLE — Ảnh Dương. $F = (112,5571/2)/(8,3/4) = 27,122$; theo R^2 : $(0,9313/2)/((1 - 0,9313)/4) = 27,122$. $F(0,05; 2; 4) = 6,944$. $F > 6,944$ nên bác bỏ H_0 : mô hình phù hợp, số nhân viên và chi quảng cáo cùng giải thích được doanh số ($p\text{-value} = 0,0047$; Hình 4.3).

Ảnh Dương. $F = (112.5571/2)/(8.3/4) = 27.122$; from R^2 : $(0.9313/2)/((1 - 0.9313)/4) = 27.122$. $F(0.05; 2, 4) = 6.944$. Since $F > 6.944$, reject H_0 : the model is significant; staff and advertising jointly explain sales ($p\text{-value} = 0.0047$; Figure 4.3).



Hình 4.3. Phân phối $F(2; 4)$, miền bác bỏ 5% và giá trị F quan sát của mô hình Ảnh Dương
Figure 4.3. The $F(2, 4)$ distribution, the 5% rejection region and the observed F of the Ảnh Dương model

Tiêu chí <i>Criterion</i>	Kiểm định t <i>t-test</i>	Kiểm định F tổng thể <i>Overall F-test</i>
H_0	$\beta_j = 0$ (một hệ số) $\beta_j = 0$ (one coefficient)	$\beta_2 = \dots = \beta_k = 0$ (mọi hệ số góc) $\beta_2 = \dots = \beta_k = 0$ (all slopes)
Phân phối <i>Distribution</i>	$t(n - k)$	$F(k - 1, n - k)$
Một phía được không <i>One-sided allowed</i>	Được <i>Yes</i>	Không, luôn đuôi phải <i>No, always right tail</i>
Câu hỏi trả lời <i>Question answered</i>	X_j có tác động riêng? <i>Does X_j have a partial effect?</i>	Các X cùng giải thích được Y ? <i>Do the X jointly explain Y?</i>
Quan hệ <i>Relation</i>	$k = 2: t^2 = F$	$k \geq 3$: không suy được từ các t $k \geq 3$: not implied by the t values

LỖI HAY GẶP / COMMON MISTAKE — Lấy $F(\alpha; k, n - k)$ hoặc $F(\alpha; k - 1, n - 1)$ làm giá trị tới hạn. Bậc tự do tử số là số hệ số góc ($k - 1$), mẫu số là bậc tự do của RSS ($n - k$). Với Ảnh Dương: $F(2; 4)$, không phải $F(3; 4)$.

Using $F(\alpha; k, n - k)$ or $F(\alpha; k - 1, n - 1)$ as the critical value. The numerator df is the number of slopes ($k - 1$); the denominator df is that of RSS ($n - k$). For Ảnh Dương: $F(2, 4)$, not $F(3, 4)$.

4.6. Dự báo trong hồi quy bội — Forecasting with multiple regression

Cho vector giá trị của các biến giải thích $X_0 = (1; X_{02}; \dots; X_{0k})'$. Dự báo điểm và phương sai của dự báo trung bình $E(Y|X_0)$, của dự báo giá trị cá biệt Y_0 :

Let $X_0 = (1; X_{02}; \dots; X_{0k})'$ be given regressor values. The point forecast and the variances of the mean forecast $E(Y|X_0)$ and of the individual forecast Y_0 are:

$$\hat{Y}_0 = X_0' \hat{\beta}; \quad \text{Var}(\hat{Y}_0) = \hat{\sigma}^2 \cdot X_0' (X'X)^{-1} X_0 \quad (4.16)$$

$$\text{Var}(Y_0 - \hat{Y}_0) = \hat{\sigma}^2 \cdot [1 + X_0'(X'X)^{-1}X_0] \quad (4.17)$$

Khoảng dự báo: $\hat{Y}_0 \pm t(\alpha/2; n - k) \cdot \text{se}$. Khoảng cá biệt rộng hơn vì cộng thêm phương sai σ^2 của sai số của chính quan sát mới. Dự báo càng xa điểm trung bình ($\bar{X}_2, \dots, \bar{X}_k$) thì $X_0'(X'X)^{-1}X_0$ càng lớn và khoảng càng rộng.

Prediction interval: $\hat{Y}_0 \pm t(\alpha/2; n - k) \cdot \text{se}$. The individual interval is wider because it adds the variance σ^2 of the new observation's own disturbance. The farther X_0 lies from the mean point ($\bar{X}_2, \dots, \bar{X}_k$), the larger $X_0'(X'X)^{-1}X_0$ and the wider the interval.

VÍ DỤ / EXAMPLE — Ánh Dương: cửa hàng mới với 6 nhân viên, quảng cáo 5 triệu đồng/tháng. $X_0 = (1; 6; 5)'$. $\hat{Y}_0 = 5,6 + 3,1 \cdot 6 + 1,2 \cdot 5 = 30,2$ chục triệu đồng.

$X_0'(X'X)^{-1}X_0 = 0,1500$. $\text{se}(\hat{Y}_0) = \sqrt{(2,0750 \cdot 0,1500)} = 0,5579$; khoảng tin cậy 95% cho doanh số trung bình: $30,2 \pm 2,776 \cdot 0,5579 = (28,65; 31,75)$. se cá biệt = $\sqrt{(2,0750 \cdot (1 + 0,1500))} = 1,5447$; khoảng dự báo cho doanh số của một cửa hàng: **(25,91; 34,49)** chục triệu đồng.

*Ánh Dương: a new shop with 6 staff and VND 5 million of advertising a month. $X_0 = (1; 6; 5)'$. $\hat{Y}_0 = 5.6 + 3.1 \cdot 6 + 1.2 \cdot 5 = 30.2$ tens of millions. $X_0'(X'X)^{-1}X_0 = 0.1500$. $\text{se}(\hat{Y}_0) = \sqrt{(2.0750 \cdot 0.1500)} = 0.5579$; 95% interval for mean sales: $30.2 \pm 2.776 \cdot 0.5579 = (28.65; 31.75)$. Individual se = $\sqrt{(2.0750 \cdot (1 + 0.1500))} = 1.5447$; prediction interval for one shop: **(25.91; 34.49)** tens of millions.*

LỖI HAY GẶP / COMMON MISTAKE — Dùng công thức dự báo của hồi quy hai biến $1/n + (X_0 - \bar{X})^2/\Sigma x^2$ cho hồi quy bội. Với nhiều biến, phải dùng $X_0'(X'X)^{-1}X_0$ vì các hệ số ước lượng có hiệp phương sai với nhau.

Using the simple-regression term $1/n + (X_0 - \bar{X})^2/\Sigma x^2$ in a multiple regression. With several regressors $X_0'(X'X)^{-1}X_0$ must be used because the coefficient estimates covary.

4.7. Hồi quy doanh thu Hạt Nâu theo giá, quảng cáo và diện tích — Hạt Nâu revenue regressed on price, advertising and area

Mô hình $Y = \beta_1 + \beta_2 \text{GIA} + \beta_3 \text{QC} + \beta_4 \text{DT} + u$ ước lượng trên 36 cửa hàng của bộ B. Kỳ vọng dấu theo lý thuyết: $\beta_2 < 0$ (giá cao hơn, ít khách hơn, với cầu co giãn), $\beta_3 > 0$, $\beta_4 > 0$. Trung bình mẫu: $Y = 360,99$ triệu đồng, $\text{GIA} = 43,0$ nghìn đồng, $\text{QC} = 17,90$ triệu đồng, $\text{DT} = 101,14 \text{ m}^2$. Kết quả EViews:

The model $Y = \beta_1 + \beta_2 \text{GIA} + \beta_3 \text{QC} + \beta_4 \text{DT} + u$ is estimated on the 36 shops of data set B. Expected signs: $\beta_2 < 0$ (higher prices, fewer customers, with elastic demand), $\beta_3 > 0$, $\beta_4 > 0$. Sample means: $Y = 360.99$ million VND, $\text{GIA} = 43.0$ thousand VND, $\text{QC} = 17.90$ million VND, $\text{DT} = 101.14 \text{ m}^2$. EViews output:

Thông tin mô hình

Model information

Dependent Variable: Y

Method: Least Squares

Included observations: 36

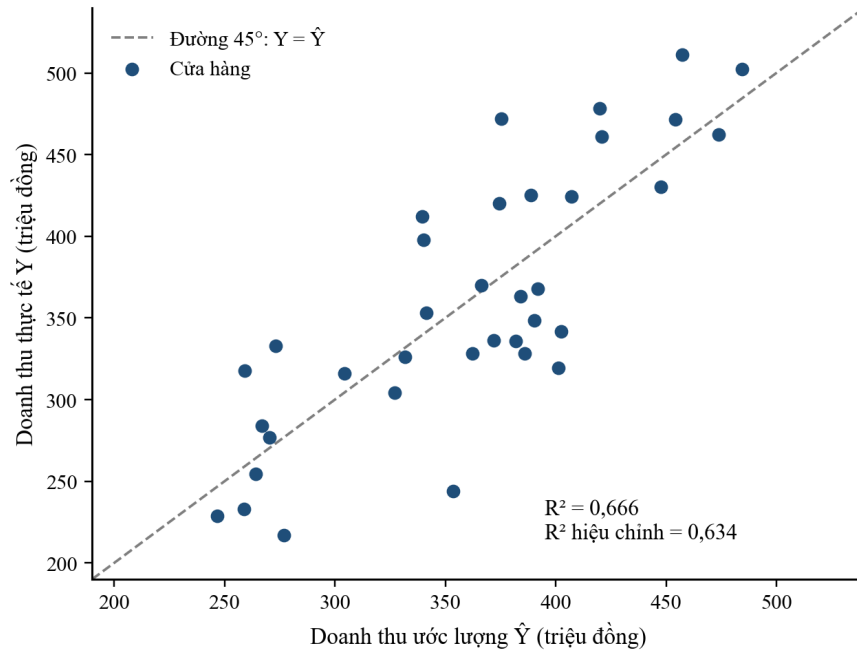
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	289.6837	60.8501	4.7606	0.0000
GIA	-3.4593	1.0712	-3.2292	0.0029
QC	6.3563	1.1948	5.3200	0.0000
DT	1.0510	0.2802	3.7508	0.0007

Thống kê Statistic	Giá trị Value	Thống kê Statistic	Giá trị Value
R-squared	0.6655	Mean dependent var	360.9917
Adjusted R-squared	0.6342	S.D. dependent var	80.8812
S.E. of regression	48.9188	Akaike info criterion	10.7226
Sum squared resid	76577.4529	Schwarz criterion	10.8986
Log likelihood	-189.0075	Hannan-Quinn criter.	10.7840
F-statistic	21.2260	Durbin-Watson stat	1.5652
Prob(F-statistic)	0.0000		

Ô trong bảng Cell	Đọc và kiểm Reading and check
Coefficient	$\beta: \hat{Y} = 289,68 - 3,4593GIA + 6,3563QC + 1,0510DT$ $\beta: \hat{Y} = 289.68 - 3.4593GIA + 6.3563QC + 1.0510DT$
Std. Error	$se(\beta_j) = \sqrt{[\sigma^2(X'X)^{-1}]_{jj}}$
t-Statistic	= Coefficient/Std. Error, vd. $6,3563/1,1948 = 5,3200$ = <i>Coefficient/Std. Error, e.g. $6.3563/1.1948 = 5.3200$</i>
Prob.	p-value hai phía của $t(32)$ <i>Two-sided p-value of $t(32)$</i>
R-squared / Adjusted	0,6655 / 0,6342 theo (4.11), (4.12) <i>0.6655 / 0.6342 by (4.11), (4.12)</i>
S.E. of regression	$\hat{\sigma} = \sqrt{(RSS/32)} = 48,9188$ $\hat{\sigma} = \sqrt{(RSS/32)} = 48.9188$
Sum squared resid	$RSS = 76.577,45$ <i>$RSS = 76,577.45$</i>
S.D. dependent var	Độ lệch chuẩn mẫu của Y; $TSS = 35 \cdot 80,8812^2 = 228.961,95$ <i>Sample s.d. of Y; $TSS = 35 \cdot 80.8812^2 = 228,961.95$</i>

Ô trong bảng Cell	Đọc và kiểm Reading and check
F-statistic / Prob	Kiểm định F (4.15): 21,2260, p-value 0,0000 <i>F-test (4.15): 21.2260, p-value 0.0000</i>
Durbin-Watson, AIC, SC	Dùng ở Chương 8 và 9 <i>Used in Chapters 8 and 9</i>

- Ý nghĩa hệ số. Giữ QC và DT không đổi, giá một ly tăng 1 nghìn đồng thì doanh thu tháng trung bình giảm 3,46 triệu đồng. Giữ GIA và DT không đổi, chi thêm 1 triệu đồng quảng cáo thì doanh thu trung bình tăng 6,36 triệu đồng. Giữ GIA và QC không đổi, thêm 1 m² diện tích thì doanh thu trung bình tăng 1,05 triệu đồng. Ba dấu khớp kỳ vọng.
Coefficients. Holding QC and DT constant, a VND 1,000 higher price per cup lowers mean monthly revenue by VND 3.46 million. Holding GIA and DT constant, VND 1 million more advertising raises mean revenue by VND 6.36 million. Holding GIA and QC constant, one more m² raises mean revenue by VND 1.05 million. All three signs match expectations.
- Kiểm định t, $\alpha = 5\%$, $t(0,025; 32) = 2,037$: $|t|$ của GIA, QC, DT lần lượt 3,229; 5,320; 3,751 đều lớn hơn 2,037 (Prob. 0,0029; 0,0000; 0,0007 < 0,05): bác bỏ H₀ với từng hệ số, cả ba biến có ảnh hưởng tới doanh thu.
t-tests, $\alpha = 5\%$, $t(0.025; 32) = 2.037$: the $|t|$ values of GIA, QC and DT are 3.229, 5.320 and 3.751, all above 2.037 (Prob. 0.0029, 0.0000, 0.0007 < 0.05): reject H₀ for each; all three affect revenue.
- Khoảng tin cậy 95%: $\beta(\text{GIA}) \in (-5,6414; -1,2772)$; $\beta(\text{QC}) \in (3,9225; 8,7901)$; $\beta(\text{DT}) \in (0,4802; 1,6218)$.
95% confidence intervals: $\beta(\text{GIA}) \in (-5.6414; -1.2772)$; $\beta(\text{QC}) \in (3.9225; 8.7901)$; $\beta(\text{DT}) \in (0.4802; 1.6218)$.
- Kiểm định một phía. Giám đốc marketing cho rằng mỗi triệu đồng quảng cáo đem về hơn 5 triệu đồng doanh thu. H₀: $\beta(\text{QC}) \leq 5$; H₁: $\beta(\text{QC}) > 5$. $t = (6,3563 - 5)/1,1948 = 1,135 < t(0,05; 32) = 1,694$: chưa đủ cơ sở bác bỏ H₀, số liệu chưa ủng hộ nhận định này dù ước lượng điểm lớn hơn 5.
One-sided test. The marketing director claims each million of advertising brings more than 5 million of revenue. H₀: $\beta(\text{QC}) \leq 5$; H₁: $\beta(\text{QC}) > 5$. $t = (6.3563 - 5)/1.1948 = 1.135 < t(0.05; 32) = 1.694$: H₀ cannot be rejected; the data do not support the claim even though the point estimate exceeds 5.
- Kiểm định F: $F = (0,6655/3)/((1 - 0,6655)/32) = 21,226 > F(0,05; 3; 32) = 2,901$: bác bỏ H₀, mô hình phù hợp. $R^2 = 0,666$: ba biến giải thích 66,6% biến động doanh thu giữa các cửa hàng; phần còn lại thuộc các yếu tố chưa đưa vào, trong đó có vị trí mặt tiền và khu vực (Chương 5 đưa thêm hai yếu tố này, R^2 lên 0,936).
F-test: $F = (0.6655/3)/((1 - 0.6655)/32) = 21.226 > F(0.05; 3, 32) = 2.901$: reject H₀; the model is significant. $R^2 = 0.666$: the three regressors explain 66.6% of the variation in revenue; the rest belongs to omitted factors, including a central frontage and the region (Chapter 5 adds them and R^2 rises to 0.936).



Hình 4.4. Doanh thu thực tế và doanh thu ước lượng từ mô hình GIA, QC, DT (36 cửa hàng Hạt Nâu)

Figure 4.4. Actual and fitted revenue from the GIA, QC, DT model (36 Hạt Nâu shops)

VÍ DỤ / EXAMPLE — Dự báo cho cửa hàng mới. Hạt Nâu dự kiến mở cửa hàng giá bình quân 42 nghìn đồng/ly, quảng cáo 20 triệu đồng/tháng, diện tích 110 m². $\hat{Y}_0 = 289,6837 - 3,4593 \cdot 42 + 6,3563 \cdot 20 + 1,0510 \cdot 110 = 387,13$ triệu đồng. Phần mềm cho $X_0'(X'X)^{-1}X_0 = 0,0325$; $se(\hat{Y}_0) = 8,8236$, se cá biệt = 49,7082. Khoảng tin cậy 95% cho doanh thu trung bình của các cửa hàng có cùng đặc điểm: (369,16; 405,10); khoảng dự báo cho doanh thu của chính cửa hàng này: (285,87; 488,39) triệu đồng.

Forecast for a new shop. Hạt Nâu plans a shop with an average price of VND 42,000 a cup, VND 20 million of advertising a month and 110 m². $\hat{Y}_0 = 289.6837 - 3.4593 \cdot 42 + 6.3563 \cdot 20 + 1.0510 \cdot 110 = 387.13$ million VND. The software gives $X_0'(X'X)^{-1}X_0 = 0.0325$; $se(\hat{Y}_0) = 8.8236$, individual $se = 49.7082$. 95% interval for the mean revenue of shops with these features: (369.16; 405.10); prediction interval for this shop: (285.87; 488.39) million VND.

Câu lệnh trên các phần mềm — Commands in common software

```

EViews:  ls Y c GIA QC DT
          wald c(3)=0, c(4)=0          (kiểm định Wald, mục 4.8)
Stata:    regress Y GIA QC DT
          test QC DT
R:        m <- lm(Y ~ GIA + QC + DT, data = b); summary(m)
          car::linearHypothesis(m, c("QC = 0", "DT = 0"))
Python:   m = smf.ols("Y ~ GIA + QC + DT", data=b).fit(); print(m.summary())
          m.f_test("QC = 0, DT = 0")
    
```

4.8. Kiểm định Wald thu hẹp: bỏ một nhóm biến — The Wald test for dropping a group of regressors

Câu hỏi thường gặp: có thể bỏ cùng lúc m biến khỏi mô hình không? Mô hình đầy đủ gọi là mô hình không ràng buộc (U), có k hệ số; mô hình sau khi bỏ m biến là mô hình ràng

buộc (R). H_0 : hệ số của m biến đều bằng 0; H_1 : có ít nhất một hệ số khác 0. Bỏ biến làm RSS tăng; nếu RSS tăng ít thì các biến bỏ đi không cần thiết.

A common question: can m regressors be dropped together? The full model is the unrestricted model (U) with k coefficients; the model without the m regressors is the restricted model (R). H_0 : all m coefficients are zero; H_1 : at least one is not. Dropping regressors raises RSS; if it rises only a little, the dropped variables are not needed.

$$F = [(RSS_R - RSS_U)/m] / [RSS_U/(n - k)] \sim F(m, n - k) \quad (4.18)$$

$$F = [(R^2_U - R^2_R)/m] / [(1 - R^2_U)/(n - k)] \quad (\text{cùng biến phụ thuộc}) \quad (4.19)$$

Kiểm định F sự phù hợp ở mục 4.5 là trường hợp riêng: bỏ mọi biến giải thích, mô hình ràng buộc chỉ còn hệ số chặn, $RSS(R) = TSS$, $R^2(R) = 0$, $m = k - 1$. Luôn có $RSS(R) \geq RSS(U)$; ra $RSS(R) < RSS(U)$ là dấu hiệu đã đảo hai mô hình.

The overall F-test of section 4.5 is a special case: drop every regressor, the restricted model keeps only the intercept, $RSS(R) = TSS$, $R^2(R) = 0$ and $m = k - 1$. Always $RSS(R) \geq RSS(U)$; getting $RSS(R) < RSS(U)$ means the models have been swapped.

VÍ DỤ / EXAMPLE — Bộ B: có thể bỏ cùng lúc QC và DT không? Mô hình U: Y theo GIA, QC, DT; $RSS(U) = 76.577,45$; $R^2(U) = 0,6655$. Mô hình R: Y theo GIA; $RSS(R) = 189.961,56$; $R^2(R) = 0,1703$. $m = 2$, $n - k = 32$. $F = [(189.961,56 - 76.577,45)/2]/[76.577,45/32] = 23,690$; theo R^2 : $[(0,6655 - 0,1703)/2]/[(1 - 0,6655)/32] = 23,690$. $F(0,05; 2, 32) = 3,295$. $F > 3,295$: bác bỏ H_0 , không được bỏ QC và DT (p -value = 0,0000).

Data set B: can QC and DT be dropped together? Model U: Y on GIA, QC, DT; $RSS(U) = 76,577.45$; $R^2(U) = 0.6655$. Model R: Y on GIA; $RSS(R) = 189,961.56$; $R^2(R) = 0.1703$. $m = 2$, $n - k = 32$. $F = [(189,961.56 - 76,577.45)/2]/[76,577.45/32] = 23.690$; from R^2 : $[(0.6655 - 0.1703)/2]/[(1 - 0.6655)/32] = 23.690$. $F(0.05; 2, 32) = 3.295$. Since $F > 3.295$, reject H_0 : QC and DT cannot be dropped (p -value = 0.0000).

Kiểm định Test	Thống kê Statistic	Giá trị Value	Xác suất Probability	Giá trị Value
Wald Test: C(3)=0, C(4)=0	F-statistic	23.6903	Prob. F(2,32)	0.0000
	Chi-square	47.3807	Prob. Chi-square(2)	0.0000

Dòng F-statistic là kiểm định (4.18); dòng Chi-square = $m \cdot F$ là dạng tiệm cận, dùng khi mẫu lớn. Đọc Prob. của dòng F: nhỏ hơn α thì bác bỏ H_0 .

The F-statistic row is test (4.18); the Chi-square row, $m \cdot F$, is the asymptotic version for large samples. Read the Prob. of the F row: below α , reject H_0 .

LỖI HAY GẶP / COMMON MISTAKE — Thấy một biến có t nhỏ rồi bỏ luôn cả nhóm biến có t nhỏ. Từng t nhỏ không có nghĩa nhóm biến cùng bằng 0; phải kiểm định đồng thời bằng (4.18). Cũng sai khi dùng (4.19) cho hai mô hình khác biến phụ thuộc (Y và Y/L).

Seeing small t -values and dropping the whole group. Small individual t -values do not mean the

coefficients are jointly zero; test them together with (4.18). It is also wrong to apply (4.19) to models with different dependent variables (Y and Y/L).

4.9. Hàm sản xuất Cobb–Douglas — The Cobb–Douglas production function

Hàm Cobb–Douglas mô tả sản lượng Q theo vốn K và lao động L. Hàm phi tuyến theo biến nhưng lấy logarit hai vế thì tuyến tính theo tham số, ước lượng được bằng OLS:

The Cobb–Douglas function relates output Q to capital K and labour L. It is non-linear in the variables, but taking logs of both sides makes it linear in the parameters, so OLS applies:

$$Q_i = A \cdot K_i^{\beta_2} \cdot L_i^{\beta_3} \cdot \exp(u_i) \quad (4.20)$$

$$\ln Q_i = \beta_1 + \beta_2 \ln K_i + \beta_3 \ln L_i + u_i, \quad \beta_1 = \ln A \quad (4.21)$$

β_2 là hệ số co giãn của sản lượng theo vốn: giữ lao động không đổi, vốn tăng 1% thì sản lượng tăng $\beta_2\%$. β_3 là hệ số co giãn theo lao động. Tổng $\beta_2 + \beta_3$ đo hiệu quả theo quy mô: tăng mọi đầu vào 1% thì sản lượng tăng $(\beta_2 + \beta_3)\%$.

β_2 is the output elasticity of capital: holding labour constant, a 1% rise in capital raises output by $\beta_2\%$. β_3 is the output elasticity of labour. The sum $\beta_2 + \beta_3$ measures returns to scale: raising every input by 1% raises output by $(\beta_2 + \beta_3)\%$.

$\beta_2 + \beta_3$	Hiệu quả theo quy mô <i>Returns to scale</i>	Tăng mọi đầu vào gấp đôi <i>Doubling every input</i>
= 1	Không đổi <i>Constant</i>	Sản lượng gấp đôi <i>Output doubles</i>
> 1	Tăng <i>Increasing</i>	Sản lượng tăng hơn gấp đôi <i>Output more than doubles</i>
< 1	Giảm <i>Decreasing</i>	Sản lượng tăng chưa tới gấp đôi <i>Output less than doubles</i>

Dữ kiện đề bài: 12 doanh nghiệp may, năm 2025. Q là giá trị sản xuất (tỷ đồng/năm), K là vốn cố định (tỷ đồng), L là lao động (chục người).

Problem data: 12 garment firms, 2025. Q is gross output (VND billion a year), K fixed capital (VND billion), L labour (tens of workers).

Doanh nghiệp <i>Firm</i>	1	2	3	4	5	6	7	8	9	10	11	12
K	48	38	44	10	40	59	28	11	58	42	16	42
L	149	47	131	111	49	44	144	64	77	105	77	20
Q	229,3 229.3	99,6 99.6	206,0 206.0	114,1 114.1	94,6 94.6	109,6 109.6	189,6 189.6	69,4 69.4	145,9 145.9	182,9 182.9	99,0 99.0	53,4 53.4

Thông tin mô hình
Model information

Dependent Variable: LOG(Q)

Method: Least Squares

Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.4759	0.1570	3.0312	0.0142
LOG(K)	0.3609	0.0257	14.0366	0.0000
LOG(L)	0.7162	0.0266	26.9050	0.0000

Thống kê <i>Statistic</i>	Giá trị <i>Value</i>	Thống kê <i>Statistic</i>	Giá trị <i>Value</i>
R-squared	0.9891	Mean dependent var	4.7998
Adjusted R-squared	0.9867	S.D. dependent var	0.4502
S.E. of regression	0.0520	Akaike info criterion	-2.8634
Sum squared resid	0.0243	Schwarz criterion	-2.7422
Log likelihood	20.1803	Hannan-Quinn criter.	-2.9083
F-statistic	407.9083	Durbin-Watson stat	2.3036
Prob(F-statistic)	0.0000		

VÍ DỤ / EXAMPLE — Đọc kết quả. $\ln \hat{Q} = 0,4759 + 0,3609 \ln K + 0,7162 \ln L$; $R^2 = 0,9891$. Giữ lao động không đổi, vốn tăng 1% thì giá trị sản xuất tăng khoảng 0,361%; giữ vốn không đổi, lao động tăng 1% thì giá trị sản xuất tăng khoảng 0,716%. Hai hệ số co giãn đều có ý nghĩa ($t = 14,04$ và $26,90$, so với $t(0,025; 9) = 2,262$). $\hat{A} = \exp(0,4759) = 1,610$. Tổng $\beta_2 + \beta_3 = 1,0772$: ước lượng điểm lớn hơn 1, nhưng muốn kết luận hiệu quả tăng theo quy mô phải kiểm định (mục 4.10).

Reading the output. $\ln \hat{Q} = 0.4759 + 0.3609 \ln K + 0.7162 \ln L$; $R^2 = 0.9891$. Holding labour constant, 1% more capital raises output by about 0.361%; holding capital constant, 1% more labour raises output by about 0.716%. Both elasticities are significant ($t = 14.04$ and 26.90 against $t(0.025; 9) = 2.262$). $\hat{A} = \exp(0.4759) = 1.610$. The sum $\beta_2 + \beta_3 = 1.0772$: the point estimate exceeds 1, but increasing returns can be claimed only after a test (section 4.10).

LỖI HAY GẶP / COMMON MISTAKE — Đọc hệ số log-log như hệ số tuyến tính: "vốn tăng 1 tỷ đồng thì sản lượng tăng 0,36 tỷ đồng". Trong mô hình log-log, hệ số là độ co giãn, đơn vị là %/% và không phụ thuộc đơn vị đo của K, Q.

Reading a log–log coefficient as a linear one: ‘one more billion of capital adds 0.36 billion of output’. In a log–log model the coefficient is an elasticity, measured in %/% and free of the units of K and Q.

4.10. Kiểm định ràng buộc tuyến tính giữa các hệ số — *Testing linear restrictions on coefficients*

Ràng buộc tuyến tính là đẳng thức giữa các hệ số, ví dụ $\beta_2 + \beta_3 = 1$ (hiệu quả không đổi theo quy mô) hay $\beta_2 = \beta_3$ (hai biến có tác động như nhau). Có hai cách kiểm định một ràng buộc.

A linear restriction is an equality among coefficients, e.g. $\beta_2 + \beta_3 = 1$ (constant returns to scale) or $\beta_2 = \beta_3$ (two regressors have equal effects). A single restriction can be tested in two ways.

Cách 1: kiểm định t dùng hiệp phương sai — *Method 1: a t-test using the covariance*

$$se(\hat{\beta}_2 \pm \hat{\beta}_3) = \sqrt{[Var(\hat{\beta}_2) + Var(\hat{\beta}_3) \pm 2Cov(\hat{\beta}_2, \hat{\beta}_3)]} \quad (4.22)$$

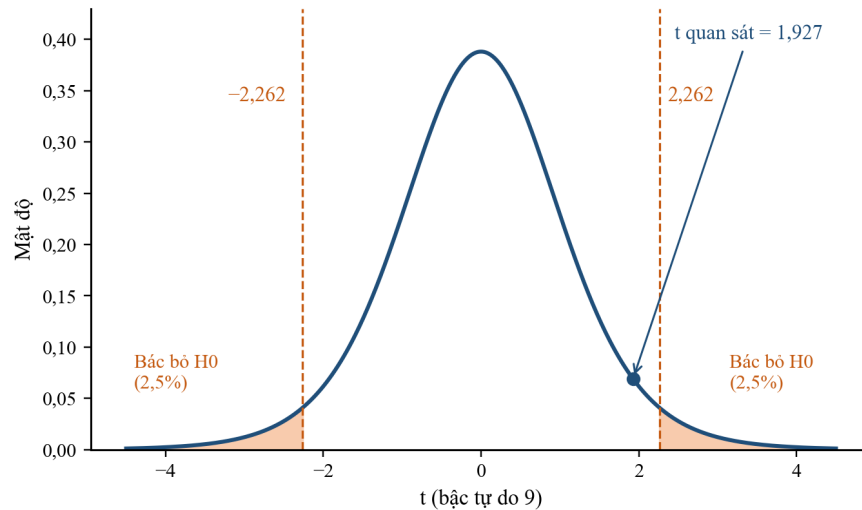
$$t = (\hat{\beta}_2 + \hat{\beta}_3 - 1) / se(\hat{\beta}_2 + \hat{\beta}_3); \quad t = (\hat{\beta}_2 - \hat{\beta}_3) / se(\hat{\beta}_2 - \hat{\beta}_3) \quad (4.23)$$

Var và Cov lấy từ ma trận hiệp phương sai của hệ số (EViews: View/Covariance Matrix). Cả hai thống kê có phân phối $t(n - k)$ dưới H_0 .

Var and Cov come from the coefficient covariance matrix (EViews: View/Covariance Matrix). Both statistics are $t(n - k)$ under H_0 .

$$Var(\hat{\beta}) = \begin{bmatrix} 0,024651 & -0,002783 & -0,003449 \\ -0,002783 & 0,000661 & 0,000117 \\ -0,003449 & 0,000117 & 0,000709 \end{bmatrix} \quad \text{thứ tự: C, LOG(K), LOG(L)}$$

VÍ DỤ / EXAMPLE — Cobb–Douglas, $H_0: \beta_2 + \beta_3 = 1$, $H_1: \beta_2 + \beta_3 \neq 1$. $Var(\beta_2) = 0,000661$; $Var(\beta_3) = 0,000709$; $Cov = 0,000117$. $se(\beta_2 + \beta_3) = \sqrt{(0,000661 + 0,000709 + 2 \cdot 0,000117)} = 0,0400$. $t = (1,0772 - 1)/0,0400 = 1,927$. $|t| < t(0,025; 9) = 2,262$: chưa đủ cơ sở bác bỏ H_0 ; số liệu phù hợp với hiệu quả không đổi theo quy mô (p-value = 0,0861; Hình 4.5). **Cobb–Douglas, $H_0: \beta_2 + \beta_3 = 1$, $H_1: \beta_2 + \beta_3 \neq 1$.** $Var(\beta_2) = 0,000661$; $Var(\beta_3) = 0,000709$; $Cov = 0,000117$. $se(\beta_2 + \beta_3) = \sqrt{(0,000661 + 0,000709 + 2 \cdot 0,000117)} = 0,0400$. $t = (1,0772 - 1)/0,0400 = 1,927$. $|t| < t(0,025; 9) = 2,262$: H_0 cannot be rejected; the data are consistent with constant returns to scale (p-value = 0.0861; Figure 4.5).



Hình 4.5. Kiểm định t hai phía cho ràng buộc $\beta_2 + \beta_3 = 1$ của hàm Cobb–Douglas: t quan sát nằm ngoài miền bác bỏ

Figure 4.5. Two-sided t-test of the restriction $\beta_2 + \beta_3 = 1$ in the Cobb–Douglas function: the observed t lies outside the rejection region

VÍ DỤ / EXAMPLE — $H_0: \beta_2 = \beta_3$ (co giãn theo vốn bằng co giãn theo lao động). $se(\beta_2 - \beta_3) = \sqrt{(0,000661 + 0,000709 - 2 \cdot 0,000117)} = 0,0337$. $t = (0,3609 - 0,7162)/0,0337 = -10,539$; $|t| > 2,262$: bác bỏ H_0 , sản lượng co giãn theo lao động nhiều hơn theo vốn. Nếu quên hiệp phương sai: $se = 0,0370$; ở bài này Cov nhỏ nên sai lệch ít, nhưng với số liệu có Cov lớn kết luận có thể đổi (Bài tập 3).

$H_0: \beta_2 = \beta_3$ (equal elasticities of capital and labour). $se(\beta_2 - \beta_3) = \sqrt{(0,000661 + 0,000709 - 2 \cdot 0,000117)} = 0,0337$. $t = (0,3609 - 0,7162)/0,0337 = -10,539$; $|t| > 2,262$: reject H_0 ; output responds more to labour than to capital. Ignoring the covariance gives $se = 0,0370$; the covariance is small here, but with a large covariance the conclusion can change (Exercise 3).

Cách 2: ước lượng mô hình ràng buộc và dùng F — Method 2: estimate the restricted model and use F

Thế $\beta_3 = 1 - \beta_2$ vào (4.21): $\ln Q - \ln L = \beta_1 + \beta_2(\ln K - \ln L) + u$. Mô hình ràng buộc: Substitute $\beta_3 = 1 - \beta_2$ into (4.21): $\ln Q - \ln L = \beta_1 + \beta_2(\ln K - \ln L) + u$. The restricted model is:

$$\ln(Q_i/L_i) = \beta_1 + \beta_2 \ln(K_i/L_i) + u_i \quad (4.24)$$

VÍ DỤ / EXAMPLE — Hồi quy (4.24) được $RSS(R) = 0,034357$; mô hình (4.21) có $RSS(U) = 0,024323$. $m = 1$, $n - k = 9$. $F = [(0,034357 - 0,024323)/1]/[0,024323/9] = 3,713 < F(0,05; 1; 9) = 5,117$: chưa đủ cơ sở bác bỏ H_0 . Kiểm tra: $t^2 = 1,927^2 = 3,713 = F$ và $t(0,025; 9)^2 = 5,117 \approx F(0,05; 1; 9)$. Chỉ dùng dạng RSS ở đây: biến phụ thuộc của (4.24) là $\ln(Q/L)$, khác $\ln Q$, nên không so sánh R^2 của hai mô hình.

Regression (4.24) gives $RSS(R) = 0.034357$; model (4.21) has $RSS(U) = 0.024323$. $m = 1$, $n - k = 9$. $F = [(0.034357 - 0.024323)/1]/[0.024323/9] = 3.713 < F(0.05; 1, 9) = 5.117$: H_0 cannot be rejected. Check: $t^2 = 1.927^2 = 3.713 = F$ and $t(0.025; 9)^2 = 5.117 \approx F(0.05; 1, 9)$. Only the RSS form applies: the dependent variable of (4.24) is $\ln(Q/L)$, not $\ln Q$, so the two R^2 values are not comparable.

Kiểm định <i>Test</i>	Thống kê <i>Statistic</i>	Giá trị <i>Value</i>	Xác suất <i>Probability</i>	Giá trị <i>Value</i>
Wald Test: $C(2)+C(3)=1$	t-statistic	1.9269	Prob. t(9)	0.0861
	F-statistic	3.7130	Prob. F(1,9)	0.0861
	Chi-square	3.7130	Prob. Chi-square(1)	0.0540

Tiêu chí <i>Criterion</i>	Wald thu hẹp <i>Dropping variables</i>	Ràng buộc tuyến tính <i>Linear restriction</i>
H0	Các hệ số của m biến = 0 <i>m coefficients = 0</i>	Đẳng thức, vd. $\beta_2 + \beta_3 = 1$ <i>An equality, e.g. $\beta_2 + \beta_3 = 1$</i>
Mô hình R <i>Model R</i>	Bỏ m biến <i>Drop the m regressors</i>	Thế ràng buộc, có thể đổi biến phụ thuộc <i>Substitute the restriction; Y may change</i>
Thống kê <i>Statistic</i>	F theo RSS hoặc R^2 <i>F from RSS or R^2</i>	F theo RSS; m = 1 dùng thêm t với Cov <i>F from RSS; for m = 1 also t with Cov</i>
Lệnh EViews <i>EViews command</i>	wald c(3)=0, c(4)=0	wald c(2)+c(3)=1

LỖI HAY GẶP / COMMON MISTAKE — Tính $se(\beta_2 - \beta_3) = se(\beta_2) - se(\beta_3)$ hoặc $\sqrt{Var \beta_2 + Var \beta_3}$ bỏ quên $2Cov$; hoặc dùng dấu trừ trước $2Cov$ khi kiểm định tổng $\beta_2 + \beta_3$. Nhớ: $Var(a \pm b) = Var a + Var b \pm 2Cov(a, b)$, dấu của Cov đi theo dấu của phép tính.
Computing $se(\beta_2 - \beta_3)$ as $se(\beta_2) - se(\beta_3)$, or as $\sqrt{Var \beta_2 + Var \beta_3}$ without $2Cov$, or using a minus before $2Cov$ when testing the sum $\beta_2 + \beta_3$. Remember $Var(a \pm b) = Var a + Var b \pm 2Cov(a, b)$: the sign of Cov follows the operation.

4.11. Bài thi mẫu: bảng EViews bị khuyết ô — *Model exam question: an EViews table with missing cells*

Đề (dạng câu 2, 4 điểm). Khảo sát 22 hộ gia đình ở Cần Thơ năm 2026 (dữ kiện đề bài): Y là chi tiêu cho giáo dục (triệu đồng/năm), X2 là thu nhập của hộ (triệu đồng/tháng), X3 là số con đang đi học (người). Kết quả hồi quy Y theo X2, X3 bằng EViews bị mất một số ô (dấu ...). (a) Điền các ô còn thiếu. (b) Nêu ý nghĩa các hệ số hồi quy riêng. (c) Với mức ý nghĩa 5%, số con đi học có ảnh hưởng tới chi tiêu giáo dục không? (d) Tìm khoảng tin cậy 95% của β_2, β_3 . (e) Kiểm định sự phù hợp của mô hình. (f) Có ý kiến cho rằng mỗi con đi học làm chi tiêu giáo dục tăng 2 triệu đồng/năm; kiểm định ý kiến đó.

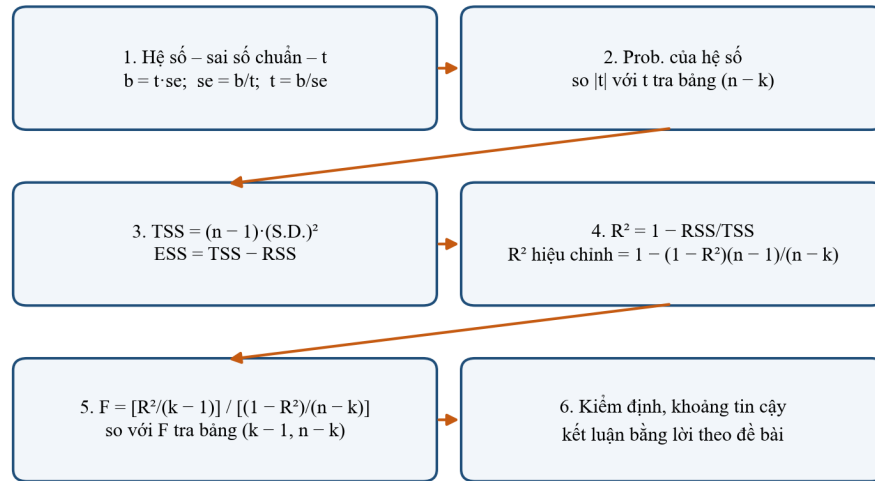
Question (type 2, 4 marks). A 2026 survey of 22 households in Cần Thơ: Y is education spending (million VND/year), X2 household income (million VND/month), X3 the number of children in school. Some cells of the EViews output for the regression of Y on X2 and X3 are missing (...).

(a) Fill in the missing cells. (b) Interpret the partial regression coefficients. (c) At the 5% level, does the number of schoolchildren affect education spending? (d) Find 95% confidence intervals for β_2 and β_3 . (e) Test the overall significance. (f) Someone claims each child in school adds VND 2 million a year to education spending; test the claim.

Thông tin mô hình <i>Model information</i>
Dependent Variable: Y
Method: Least Squares
Included observations: 22

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	...	1.4821	1.5022	0.1495
X2	0.3554	...	7.2846	0.0000
X3	2.9559	0.3903	...	...

Thống kê <i>Statistic</i>	Giá trị <i>Value</i>	Thống kê <i>Statistic</i>	Giá trị <i>Value</i>
R-squared	...	Mean dependent var	15.5227
Adjusted R-squared	...	S.D. dependent var	5.0130
S.E. of regression	2.1064	Akaike info criterion	4.4539
Sum squared resid	84.2990	Schwarz criterion	4.6027
Log likelihood	-45.9932	Hannan-Quinn criter.	4.4890
F-statistic	...	Durbin-Watson stat	2.6244
Prob(F-statistic)	...		



Hình 4.6. Thứ tự điền ô của bảng EViews bị khuyết
Figure 4.6. Order for filling missing cells in an EViews table

Lời giải — Solution

- (a) Ô của hệ số: $b = t \cdot se$, $se = b/t$, $t = b/se$. C: $\beta_1 = 1,5022 \cdot 1,4821 = \mathbf{2,2264}$. X_2 : $se(\beta_2) = 0,3554/7,2846 = \mathbf{0,0488}$. X_3 : $t = 2,9559/0,3903 = \mathbf{7,5734}$.
(a) Coefficient cells: $b = t \cdot se$, $se = b/t$, $t = b/se$. C: $\beta_1 = 1.5022 \cdot 1.4821 = \mathbf{2.2264}$. X_2 : $se(\beta_2) = 0.3554/7.2846 = \mathbf{0.0488}$. X_3 : $t = 2.9559/0.3903 = \mathbf{7.5734}$.
- Prob. của X_3 : không tính tay được giá trị chính xác, nhưng $|t| = 7,573$ lớn hơn $t(0,005; 19) = 2,861$ nên Prob. $< 0,01$; máy in 0,0000.
Prob. for X_3 : the exact value cannot be found by hand, but $|t| = 7.573$ exceeds $t(0.005; 19) = 2.861$, so Prob. < 0.01 ; the software prints 0.0000.
- $TSS = (n - 1) \cdot (S.D. \text{ dependent var})^2 = 21 \cdot 5,0130^2 = 527,7335$. $R^2 = 1 - RSS/TSS = 1 - 84,2990/527,7335 = \mathbf{0,8403}$; $ESS = 443,4345$. $\bar{R}^2 = 1 - (1 - 0,8403) \cdot 21/19 = \mathbf{0,8234}$.
 $TSS = (n - 1) \cdot (S.D. \text{ dependent var})^2 = 21 \cdot 5.0130^2 = 527.7335$. $R^2 = 1 - RSS/TSS = 1 - 84.2990/527.7335 = \mathbf{0.8403}$; $ESS = 443.4345$. Adjusted $R^2 = 1 - (1 - 0.8403) \cdot 21/19 = \mathbf{0.8234}$.
- $F = [R^2/(k - 1)] / [(1 - R^2)/(n - k)] = (0,8403/2) / ((1 - 0,8403)/19) = \mathbf{49,9725}$. Prob(F-statistic): F lớn hơn nhiều so với $F(0,05; 2; 19) = 3,522$ nên Prob $\approx 0,0000$.
 $F = [R^2/(k - 1)] / [(1 - R^2)/(n - k)] = (0.8403/2) / ((1 - 0.8403)/19) = \mathbf{49.9725}$. Prob(F-statistic): F far exceeds $F(0.05; 2, 19) = 3.522$, so Prob ≈ 0.0000 .
- (b) Giữ số con đi học không đổi, thu nhập hộ tăng 1 triệu đồng/tháng thì chi tiêu giáo dục trung bình tăng 0,355 triệu đồng/năm. Giữ thu nhập không đổi, thêm một con đi học thì chi tiêu giáo dục trung bình tăng 2,956 triệu đồng/năm. $R^2 = 0,840$: hai biến giải thích 84,0% biến động chi tiêu giáo dục giữa các hộ.
(b) Holding the number of schoolchildren constant, VND 1 million more monthly income raises mean education spending by VND 0.355 million a year. Holding income constant, one more child in school raises it by VND 2.956 million a year. $R^2 = 0.840$: the two regressors explain 84.0% of the variation across households.

6. (c) $H_0: \beta_3 = 0$; $H_1: \beta_3 \neq 0$. $t = 7,573 > t(0,025; 19) = 2,093$: bác bỏ H_0 , số con đi học có ảnh hưởng tới chi tiêu giáo dục.
 (c) $H_0: \beta_3 = 0$; $H_1: \beta_3 \neq 0$. $t = 7.573 > t(0.025; 19) = 2.093$: reject H_0 ; the number of schoolchildren affects education spending.
7. (d) $\beta_2 \in 0,3554 \pm 2,093 \cdot 0,0488 = (0,2533; 0,4575)$; $\beta_3 \in 2,9559 \pm 2,093 \cdot 0,3903 = (2,1390; 3,7728)$ triệu đồng/năm.
 (d) $\beta_2 \in 0.3554 \pm 2.093 \cdot 0.0488 = (0.2533; 0.4575)$; $\beta_3 \in 2.9559 \pm 2.093 \cdot 0.3903 = (2.1390; 3.7728)$ million VND a year.
8. (e) $H_0: \beta_2 = \beta_3 = 0$. $F = 49,972 > F(0,05; 2; 19) = 3,522$: bác bỏ H_0 , mô hình phù hợp.
 (e) $H_0: \beta_2 = \beta_3 = 0$. $F = 49.972 > F(0.05; 2, 19) = 3.522$: reject H_0 ; the model is significant.
9. (f) $H_0: \beta_3 = 2$; $H_1: \beta_3 \neq 2$. $t = (2,9559 - 2)/0,3903 = 2,449 > 2,093$: bác bỏ H_0 , mức tăng thực tế khác 2 triệu đồng/năm (khoảng tin cậy ở (d) không chứa 2, cùng kết luận).
 (f) $H_0: \beta_3 = 2$; $H_1: \beta_3 \neq 2$. $t = (2.9559 - 2)/0.3903 = 2.449 > 2.093$: reject H_0 ; the increase differs from VND 2 million a year (the interval in (d) excludes 2, the same conclusion).

LỖI HAY GẶP / COMMON MISTAKE — Tính $TSS = n \cdot (S.D.)^2$ thay vì $(n - 1) \cdot (S.D.)^2$. S.D. dependent var trong EViews là độ lệch chuẩn mẫu (chia $n - 1$). Sai ở bước này kéo theo R^2 , $\bar{R}^2$, F đều sai. Lỗi thứ hai: lấy Prob. của hệ số chặn để kết luận về mô hình; mô hình được đánh giá bằng Prob(F-statistic).
 Computing $TSS = n \cdot (S.D.)^2$ instead of $(n - 1) \cdot (S.D.)^2$. The EViews S.D. dependent var is the sample standard deviation (divisor $n - 1$). An error here spoils R^2 , adjusted R^2 and F . A second error: using the intercept's Prob. to judge the model; the model is judged by Prob(F-statistic).

4.12. Tóm tắt và thuật ngữ — Summary and glossary

PHẢI THUỘC / MEMORISE — Mô hình: $Y = X\beta + U$; $\beta' = (X'X)^{-1}X'Y$; $\text{Var}(\beta) = \sigma^2(X'X)^{-1}$; $\sigma^2 = \text{RSS}/(n - k)$, k gồm hệ số chặn. Ba biến, dạng độ lệch: $\beta_2 = (\sum yx_2 \sum x_3^2 - \sum yx_3 \sum x_2 x_3)/D$, $D = \sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2$; $\text{Var}(\beta_2) = \sigma^2 \sum x_3^2 / D$. $R^2 = 1 - \text{RSS}/\text{TSS}$; $\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k)$. $t = (\hat{\beta}_j - \beta^*)/se \sim t(n - k)$; khoảng tin cậy $\hat{\beta}_j \pm t(\alpha/2; n - k) \cdot se$. F tổng thể = $[R^2/(k - 1)]/[(1 - R^2)/(n - k)] \sim F(k - 1, n - k)$. Wald: $F = [(RSS(R) - RSS(U))/m]/[RSS(U)/(n - k)]$. Ràng buộc một đẳng thức: t với $se = \sqrt{(\text{Var} + \text{Var} \pm 2\text{Cov})}$, $t^2 = F$. Bảng khuyết ô: $b = t \cdot se$; $TSS = (n - 1) \cdot S.D.^2$; R^2 từ RSS, TSS; F từ R^2 . Cobb–Douglas: hệ số là co giãn, $\beta_2 + \beta_3$ so với 1.
 Model: $Y = X\beta + U$; $\beta' = (X'X)^{-1}X'Y$; $\text{Var}(\beta) = \sigma^2(X'X)^{-1}$; $\sigma^2 = \text{RSS}/(n - k)$, k includes the intercept. Three variables, deviation form: $\beta_2 = (\sum yx_2 \sum x_3^2 - \sum yx_3 \sum x_2 x_3)/D$, $D = \sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2$; $\text{Var}(\beta_2) = \sigma^2 \sum x_3^2 / D$. $R^2 = 1 - \text{RSS}/\text{TSS}$; adjusted $R^2 = 1 - (1 - R^2)(n - 1)/(n - k)$. $t = (\hat{\beta}_j - \beta^*)/se \sim t(n - k)$; interval $\hat{\beta}_j \pm t(\alpha/2; n - k) \cdot se$. Overall $F = [R^2/(k - 1)]/[(1 - R^2)/(n - k)] \sim F(k - 1, n - k)$. Wald: $F = [(RSS(R) - RSS(U))/m]/[RSS(U)/(n - k)]$. One equality restriction: t with $se = \sqrt{(\text{Var} + \text{Var} \pm 2\text{Cov})}$, $t^2 = F$. Missing cells: $b = t \cdot se$; $TSS = (n - 1) \cdot S.D.^2$; R^2 from RSS and TSS; F from R^2 . Cobb–Douglas: coefficients are elasticities; compare $\beta_2 + \beta_3$ with 1.

Các bước của một kiểm định trong bài thi — Steps of a test in an exam answer

Bước <i>Step</i>	Nội dung <i>Content</i>	Ví dụ (bộ B, hệ số của DT) <i>Example (data set B, DT coefficient)</i>
1	Viết H_0 , H_1 bằng ký hiệu và bằng lời <i>State H_0 and H_1 in symbols and words</i>	$H_0: \beta_4 = 0$ (DT không ảnh hưởng); $H_1: \beta_4 \neq 0$ <i>$H_0: \beta_4 = 0$ (no effect); $H_1: \beta_4 \neq 0$</i>
2	Chọn thống kê, ghi phân phối và bậc tự do <i>Choose the statistic, its distribution and df</i>	$t \sim t(32)$
3	Thế số tính giá trị quan sát <i>Plug in the numbers</i>	$t = 1,0510/0,2802 = 3,751$ <i>$t = 1.0510/0.2802 = 3.751$</i>
4	Tra bảng giá trị tới hạn <i>Look up the critical value</i>	$t(0,025; 32) = 2,037$ <i>$t(0.025; 32) = 2.037$</i>
5	So sánh, kết luận thống kê <i>Compare and decide</i>	$ t > t$ tới hạn: bác bỏ H_0 <i>Reject H_0</i>
6	Kết luận bằng lời theo câu hỏi <i>State the answer in words</i>	Diện tích có ảnh hưởng tới doanh thu <i>Floor area affects revenue</i>

Lỗi hay gặp khác — Other common mistakes

LỖI HAY GẶP / COMMON MISTAKE — Diễn giải hệ số hồi quy riêng mà thiếu cụm "khi các biến khác không đổi", hoặc so sánh độ lớn hệ số của các biến có đơn vị khác nhau để nói biến nào "quan trọng hơn" (1 nghìn đồng giá và 1 m² diện tích không cùng thước đo).
Interpreting a partial coefficient without 'other things equal', or comparing the sizes of coefficients measured in different units to decide which variable is 'more important' (VND 1,000 of price and 1 m² of area are not on the same scale).

LỖI HAY GẶP / COMMON MISTAKE — Viết "chấp nhận H_0 " hoặc " $\beta_3 = 0$ " khi $|t|$ nhỏ. Kết luận đúng: chưa đủ cơ sở bác bỏ H_0 ở mức ý nghĩa đã chọn; số liệu không chứng minh được hệ số bằng 0.
Writing 'accept H_0 ' or ' $\beta_3 = 0$ ' when $|t|$ is small. The correct wording: H_0 cannot be rejected at the chosen level; the data do not prove the coefficient is zero.

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>
Hồi quy bội	Multiple regression	Hồi quy Y theo hai biến giải thích trở lên <i>Regression of Y on two or more regressors</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Hệ số hồi quy riêng	Partial regression coefficient	Tác động của X_j khi các X khác không đổi <i>Effect of X_j, other X fixed</i>
Ma trận thiết kế	Design matrix (X)	Ma trận $n \times k$, cột đầu toàn số 1 <i>$n \times k$ matrix, first column ones</i>
Hệ phương trình chuẩn	Normal equations	$X'X\beta = X'Y$
Ma trận hiệp phương sai	Covariance matrix	$\sigma^2(X'X)^{-1}$: Var trên đường chéo, Cov ngoài <i>$\sigma^2(X'X)^{-1}$: Var on the diagonal, Cov off it</i>
Bậc tự do	Degrees of freedom	$n - k$ với RSS <i>$n - k$ for RSS</i>
Hệ số xác định	Coefficient of determination (R^2)	Tỷ lệ biến động Y được giải thích <i>Share of variation in Y explained</i>
R^2 hiệu chỉnh	Adjusted R-squared	R^2 đã phạt số hệ số <i>R^2 penalised for k</i>
Kiểm định t	t-test	Kiểm định một hệ số <i>Test on one coefficient</i>
Kiểm định F	F-test	Kiểm định đồng thời nhiều hệ số <i>Joint test on several coefficients</i>
Sự phù hợp của mô hình	Overall significance	H_0 : mọi hệ số góc bằng 0 <i>H_0: all slopes are zero</i>
Kiểm định Wald	Wald test	Kiểm định ràng buộc trên hệ số <i>Test of restrictions on coefficients</i>
Mô hình ràng buộc	Restricted model	Mô hình đã áp H_0 <i>Model with H_0 imposed</i>
Mô hình không ràng buộc	Unrestricted model	Mô hình đầy đủ <i>Full model</i>
Ràng buộc tuyến tính	Linear restriction	Đẳng thức tuyến tính giữa các β <i>Linear equality among the β</i>
Khoảng tin cậy	Confidence interval	$\beta \pm t \cdot se$
Khoảng dự báo	Prediction interval	Khoảng cho một giá trị Y mới <i>Interval for a new Y value</i>
Hàm Cobb–Douglas	Cobb–Douglas function	$\ln Q = \ln A + \beta_2 \ln K + \beta_3 \ln L$
Hệ số co giãn	Elasticity	% thay đổi Y khi X tăng 1% <i>% change in Y per 1% in X</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Hiệu quả theo quy mô	Returns to scale	So sánh $\beta_2 + \beta_3$ với 1 <i>Compare $\beta_2 + \beta_3$ with 1</i>
Đa cộng tuyến hoàn hảo	Perfect multicollinearity	Một X là tổ hợp tuyến tính đúng của các X khác <i>One X is an exact linear combination of others</i>
Giá trị tới hạn	Critical value	Ngưỡng tra bảng để bác bỏ H_0 <i>Table threshold for rejecting H_0</i>

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

Bài 1. Hồi quy ba biến tính tay theo dạng độ lệch — *Exercise 1. A three-variable regression by hand in deviation form*

Số liệu của 10 hợp tác xã trồng lúa (dữ kiện đề bài): Y là năng suất (tạ/ha), X₂ là lượng phân bón (tạ/ha), X₃ là công lao động (chục ngày công/ha). Y = (44; 45; 57; 54; 64; 44; 58; 48; 52; 54); X₂ = (2; 4; 5; 6; 8; 3; 7; 5; 4; 6); X₃ = (3; 2; 4; 3; 5; 1; 4; 2; 3; 3). Với chữ thường là độ lệch so với trung bình, đã tính được $\sum x_2^2 = 30$; $\sum x_3^2 = 12$; $\sum x_2x_3 = 13$; $\sum x_2y = 99$; $\sum x_3y = 62$; $\sum y^2 = 406$. (a) Ước lượng hàm hồi quy Y theo X₂, X₃ và nêu ý nghĩa các hệ số. (b) Tính R² và R² hiệu chỉnh. (c) Với mức ý nghĩa 5%, công lao động có ảnh hưởng tới năng suất không? (d) Tìm khoảng tin cậy 95% của β_3 . (e) Kiểm định sự phù hợp của mô hình.

Data for 10 rice cooperatives: Y is yield (quintals/ha), X₂ fertiliser (quintals/ha), X₃ labour (tens of workdays/ha). Y = (44; 45; 57; 54; 64; 44; 58; 48; 52; 54); X₂ = (2; 4; 5; 6; 8; 3; 7; 5; 4; 6); X₃ = (3; 2; 4; 3; 5; 1; 4; 2; 3; 3). With lower-case letters denoting deviations from the mean: $\sum x_2^2 = 30$; $\sum x_3^2 = 12$; $\sum x_2x_3 = 13$; $\sum x_2y = 99$; $\sum x_3y = 62$; $\sum y^2 = 406$. (a) Estimate the regression of Y on X₂ and X₃ and interpret the coefficients. (b) Compute R² and adjusted R². (c) At the 5% level, does labour affect yield? (d) Find a 95% confidence interval for β_3 . (e) Test the overall significance of the model.

1. Trung bình: $\bar{Y} = 52$; $\bar{X}_2 = 5$; $\bar{X}_3 = 3$. Mẫu số chung: $\sum x_2^2 \cdot \sum x_3^2 - (\sum x_2x_3)^2 = 30 \cdot 12 - 13^2 = 191$.

Means: $\bar{Y} = 52$; $\bar{X}_2 = 5$; $\bar{X}_3 = 3$. Common denominator: $\sum x_2^2 \cdot \sum x_3^2 - (\sum x_2x_3)^2 = 30 \cdot 12 - 13^2 = 191$.

2. (a) $\beta_2 = (99 \cdot 12 - 62 \cdot 13)/191 = 2$; $\beta_3 = (62 \cdot 30 - 99 \cdot 13)/191 = 3$; $\beta_1 = 52 - 2 \cdot 5 - 3 \cdot 3 = 33$. $\hat{Y} = 33 + 2X_2 + 3X_3$. Giữ công lao động không đổi, bón thêm 1 tạ phân/ha thì năng suất trung bình tăng 2 tạ/ha; giữ lượng phân không đổi, thêm 10 ngày công/ha thì năng suất trung bình tăng 3 tạ/ha.

(a) $\beta_2 = (99 \cdot 12 - 62 \cdot 13)/191 = 2$; $\beta_3 = (62 \cdot 30 - 99 \cdot 13)/191 = 3$; $\beta_1 = 52 - 2 \cdot 5 - 3 \cdot 3 = 33$. Holding labour constant, one more quintal of fertiliser per hectare raises mean yield by 2 quintals/ha; holding fertiliser constant, ten more workdays per hectare raise mean yield by 3 quintals/ha.

3. (b) $ESS = \beta_2 \sum x_2 y + \beta_3 \sum x_3 y = 2 \cdot 99 + 3 \cdot 62 = 384$; $TSS = \sum y^2 = 406$; $RSS = 22$. $R^2 = 384/406 = \mathbf{0,9458}$; $\bar{R}^2 = 1 - (1 - 0,9458) \cdot 9/7 = \mathbf{0,9303}$.
 (b) $ESS = \beta_2 \sum x_2 y + \beta_3 \sum x_3 y = 2 \cdot 99 + 3 \cdot 62 = 384$; $TSS = \sum y^2 = 406$; $RSS = 22$. $R^2 = 384/406 = \mathbf{0.9458}$; $adjusted\ R^2 = 1 - (1 - 0.9458) \cdot 9/7 = \mathbf{0.9303}$.
4. (c) $\sigma^2 = 22/(10 - 3) = 3,1429$. $Var(\beta_3) = \sigma^2 \cdot \sum x_3^2 / 191 = 3,1429 \cdot 30 / 191 = 0,4936$, $se(\beta_3) = 0,7026$. $H_0: \beta_3 = 0$, $H_1: \beta_3 \neq 0$; $t = 3/0,7026 = \mathbf{4,270} > t(0,025; 7) = 2,365$: bác bỏ H_0 , công lao động có ảnh hưởng tới năng suất.
 (c) $\sigma^2 = 22/(10 - 3) = 3.1429$. $Var(\beta_3) = \sigma^2 \cdot \sum x_3^2 / 191 = 3.1429 \cdot 30 / 191 = 0.4936$, $se(\beta_3) = 0.7026$. $H_0: \beta_3 = 0$, $H_1: \beta_3 \neq 0$; $t = 3/0.7026 = \mathbf{4.270} > t(0.025; 7) = 2.365$: *reject H_0 ; labour affects yield.*
5. (d) $\beta_3 \in 3 \pm 2,365 \cdot 0,7026 = \mathbf{(1,3384; 4,6616)}$ tạ/ha cho mỗi 10 ngày công. Khoảng không chứa 0, khớp kết luận ở (c).
 (d) $\beta_3 \in 3 \pm 2.365 \cdot 0.7026 = \mathbf{(1.3384; 4.6616)}$ quintals/ha per ten workdays. The interval excludes 0, consistent with (c).
6. (e) $H_0: \beta_2 = \beta_3 = 0$. $F = (384/2)/(22/7) = \mathbf{61,09} > F(0,05; 2; 7) = 4,737$: bác bỏ H_0 , mô hình phù hợp.
 (e) $H_0: \beta_2 = \beta_3 = 0$. $F = (384/2)/(22/7) = \mathbf{61.09} > F(0.05; 2, 7) = 4.737$: *reject H_0 ; the model is significant.*

Bài 2. Đọc phương trình hồi quy kèm t-statistic — Exercise 2. Reading a regression equation reported with t-statistics

Với số liệu 20 đại lý gạo, ước lượng được: $\hat{Y} = 15,4 + 0,82X_2 - 1,35X_3$; giá trị t ghi dưới hệ số lần lượt (3,12), (4,56), (-2,48); $R^2 = 0,78$. Y: lượng bán (tấn/tháng), X_2 : thu nhập bình quân đầu người của khu vực (triệu đồng/tháng), X_3 : giá bán (nghìn đồng/kg). (a) Nêu ý nghĩa các hệ số hồi quy riêng. (b) Tìm khoảng tin cậy 95% của β_2 và β_3 . (c) Có thể cho rằng giá tăng làm lượng bán giảm không ($\alpha = 5\%$)? (d) Mô hình có phù hợp không? (e) Tính R^2 hiệu chỉnh.

Using data for 20 rice dealers: $\hat{Y} = 15.4 + 0.82X_2 - 1.35X_3$; t-values below the coefficients are (3.12), (4.56), (-2.48); $R^2 = 0.78$. Y: sales (tonnes/month), X_2 : average per-capita income in the area (million VND/month), X_3 : price (thousand VND/kg). (a) Interpret the partial regression coefficients. (b) Find 95% confidence intervals for β_2 and β_3 . (c) Can we say that a higher price lowers sales ($\alpha = 5\%$)? (d) Is the model significant? (e) Compute adjusted R^2 .

1. (a) Giữ giá không đổi, thu nhập khu vực tăng 1 triệu đồng/tháng thì lượng bán trung bình tăng 0,82 tấn/tháng. Giữ thu nhập không đổi, giá tăng 1 nghìn đồng/kg thì lượng bán trung bình giảm 1,35 tấn/tháng. Hệ số chặn 15,4 không có ý nghĩa kinh tế vì không có khu vực thu nhập 0 và giá 0.
 (a) Holding price constant, a rise of VND 1 million a month in area income raises mean sales by 0.82 tonnes a month. Holding income constant, a rise of VND 1,000/kg in price lowers mean sales by 1.35 tonnes a month. The intercept has no economic meaning: no area has zero income and zero price.
2. (b) $se = \text{hệ số}/t$: $se(\beta_2) = 0,82/4,56 = 0,1798$; $se(\beta_3) = -1,35/(-2,48) = 0,5444$. Với $t(0,025; 17) = 2,110$: $\beta_2 \in \mathbf{(0,4406; 1,1994)}$; $\beta_3 \in \mathbf{(-2,4986; -0,2014)}$.

(b) $se = \text{coefficient}/t$: $se(\beta_2) = 0.82/4.56 = 0.1798$; $se(\beta_3) = -1.35/(-2.48) = 0.5444$. With $t(0.025; 17) = 2.110$: $\beta_2 \in (0.4406; 1.1994)$; $\beta_3 \in (-2.4986; -0.2014)$.

3. (c) Kiểm định một phía: $H_0: \beta_3 \geq 0$, $H_1: \beta_3 < 0$. $t = -2.48 < -t(0.05; 17) = -1.740$: bác bỏ H_0 , giá tăng làm lượng bán giảm.
(c) *One-sided test: $H_0: \beta_3 \geq 0$, $H_1: \beta_3 < 0$. $t = -2.48 < -t(0.05; 17) = -1.740$: reject H_0 ; a higher price lowers sales.*
4. (d) $H_0: \beta_2 = \beta_3 = 0$. $F = (0.78/2)/((1 - 0.78)/17) = 30.14 > F(0.05; 2; 17) = 3.592$: bác bỏ H_0 , mô hình phù hợp.
(d) *$H_0: \beta_2 = \beta_3 = 0$. $F = (0.78/2)/((1 - 0.78)/17) = 30.14 > F(0.05; 2, 17) = 3.592$: reject H_0 ; the model is significant.*
5. (e) $\bar{R}^2 = 1 - (1 - 0.78) \cdot 19/17 = 0.7541$.
(e) *Adjusted $R^2 = 1 - (1 - 0.78) \cdot 19/17 = 0.7541$.*

Bài 3. Hai kênh quảng cáo có hiệu quả như nhau không? — Exercise 3. Are two advertising channels equally effective?

Một chuỗi siêu thị điện máy ước lượng doanh thu tháng Y (triệu đồng) theo chi quảng cáo truyền hình X_2 và chi quảng cáo mạng xã hội X_3 (triệu đồng) với 30 tháng số liệu: $\beta_2 = 2.40$, $se(\beta_2) = 0.50$; $\beta_3 = 3.60$, $se(\beta_3) = 0.45$; $Cov(\beta_2, \beta_3) = -0.06$ (dữ kiện đề bài). (a) Kiểm định giả thuyết hai kênh có hiệu quả như nhau, $\alpha = 5\%$. (b) Phòng marketing cho rằng mạng xã hội hiệu quả hơn truyền hình; kiểm định ý kiến này. (c) Nếu bỏ quên hiệp phương sai thì kết quả thay đổi thế nào?

An electronics retail chain regresses monthly revenue Y (million VND) on TV advertising X_2 and social-media advertising X_3 (million VND) over 30 months: $\beta_2 = 2.40$, $se(\beta_2) = 0.50$; $\beta_3 = 3.60$, $se(\beta_3) = 0.45$; $Cov(\beta_2, \beta_3) = -0.06$. (a) Test the hypothesis that the two channels are equally effective, $\alpha = 5\%$. (b) Marketing claims social media is more effective than TV; test the claim. (c) How does the result change if the covariance is ignored?

1. (a) $H_0: \beta_3 - \beta_2 = 0$, $H_1: \beta_3 - \beta_2 \neq 0$. $Var(\beta_3 - \beta_2) = 0.45^2 + 0.50^2 - 2 \cdot (-0.06) = 0.5725$; $se = 0.7566$.
(a) *$H_0: \beta_3 - \beta_2 = 0$, $H_1: \beta_3 - \beta_2 \neq 0$. $Var(\beta_3 - \beta_2) = 0.45^2 + 0.50^2 - 2 \cdot (-0.06) = 0.5725$; $se = 0.7566$.*
2. $t = (3.60 - 2.40)/0.7566 = 1.586 < t(0.025; 27) = 2.052$: chưa đủ cơ sở bác bỏ H_0 ; số liệu chưa cho thấy hai kênh khác nhau về hiệu quả.
 $t = (3.60 - 2.40)/0.7566 = 1.586 < t(0.025; 27) = 2.052$: H_0 cannot be rejected; the data do not show a difference between the channels.
3. (b) $H_0: \beta_3 \leq \beta_2$, $H_1: \beta_3 > \beta_2$. $t = 1.586 < t(0.05; 27) = 1.703$: chưa đủ cơ sở bác bỏ H_0 ; ý kiến của phòng marketing chưa được số liệu ủng hộ ở mức 5%.
(b) *$H_0: \beta_3 \leq \beta_2$, $H_1: \beta_3 > \beta_2$. $t = 1.586 < t(0.05; 27) = 1.703$: H_0 cannot be rejected; the claim is not supported at the 5% level.*
4. (c) Bỏ quên hiệp phương sai: $se = \sqrt{(0.45^2 + 0.50^2)} = 0.6727$, $t = 1.784$. Kiểm định hai phía ở (a) không đổi kết luận ($1.784 < 2.052$), nhưng kiểm định một phía ở (b) đổi kết luận: $1.784 > 1.703$, tức sẽ bác bỏ H_0 và ủng hộ sai ý kiến của phòng marketing. Cov âm làm phương sai của hiệu tăng; bỏ Cov làm se nhỏ đi và t lớn lên.

(c) Ignoring the covariance: $se = \sqrt{(0.45^2 + 0.50^2)} = 0.6727$, $t = 1.784$. The two-sided test in (a) keeps its conclusion ($1.784 < 2.052$), but the one-sided test in (b) flips: $1.784 > 1.703$, so H_0 would be rejected and the marketing claim wrongly supported. A negative covariance raises the variance of a difference; dropping it shrinks se and inflates t .

5. Ghi chú: kiểm định tổng $\beta_2 + \beta_3$ dùng dấu cộng trước $2Cov$: $se(\beta_2 + \beta_3) = \sqrt{(0.50^2 + 0.45^2 + 2 \cdot (-0.06))} = 0.5766$.

Note: a test on the sum $\beta_2 + \beta_3$ uses a plus sign before $2Cov$: $se(\beta_2 + \beta_3) = \sqrt{(0.50^2 + 0.45^2 + 2 \cdot (-0.06))} = 0.5766$.

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

Câu 1. Vì sao phải dùng R^2 hiệu chỉnh khi so sánh các mô hình hồi quy bội có số biến khác nhau? R^2 hiệu chỉnh có thể âm không?

Question 1. Why should adjusted R^2 be used when comparing multiple regressions with different numbers of regressors? Can adjusted R^2 be negative?

- $R^2 = 1 - RSS/TSS$ không giảm khi thêm biến: OLS với nhiều biến hơn luôn đạt RSS nhỏ hơn hoặc bằng, kể cả khi biến thêm vào không liên quan tới Y .

$R^2 = 1 - RSS/TSS$ never falls when a regressor is added: OLS with more regressors always attains an RSS no larger, even when the new variable is unrelated to Y .

- $\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k)$ chia RSS và TSS cho bậc tự do tương ứng; thêm biến làm $n - k$ giảm nên $\bar{R}^2$ chỉ tăng khi RSS giảm đủ bù phần bậc tự do mất đi (khi $|t|$ của biến mới lớn hơn 1).

Adjusted $R^2 = 1 - (1 - R^2)(n - 1)/(n - k)$ divides RSS and TSS by their degrees of freedom; adding a variable lowers $n - k$, so adjusted R^2 rises only if RSS falls enough to offset the lost degree of freedom (when the new variable's $|t|$ exceeds 1).

- $\bar{R}^2 \leq R^2$, bằng nhau khi $k = 1$ hoặc $R^2 = 1$. $\bar{R}^2$ âm khi $R^2 < (k - 1)/(n - 1)$, ví dụ $n = 20$, $k = 8$, $R^2 = 0.3$.

Adjusted $R^2 \leq R^2$, with equality when $k = 1$ or $R^2 = 1$. It is negative when $R^2 < (k - 1)/(n - 1)$, for example $n = 20$, $k = 8$, $R^2 = 0.3$.

- Chỉ so sánh R^2 hay $\bar{R}^2$ giữa các mô hình có cùng biến phụ thuộc (cùng dạng: Y với Y , $\ln Y$ với $\ln Y$) và cùng mẫu.

R^2 or adjusted R^2 can be compared only across models with the same dependent variable (Y with Y , $\ln Y$ with $\ln Y$) and the same sample.

- $\bar{R}^2$ cao chưa đủ để chọn mô hình: còn phải xét dấu hệ số theo lý thuyết kinh tế, ý nghĩa thống kê và các kiểm định chỉ định ở Chương 6–9.

A high adjusted R^2 is not enough to choose a model: coefficient signs must match economic theory, coefficients must be significant and the specification tests of Chapters 6–9 must be passed.

Câu 2. Phân biệt kiểm định t từng hệ số và kiểm định F sự phù hợp của mô hình trong hồi quy bội. Có thể xảy ra trường hợp các t đều không có ý nghĩa nhưng F có ý nghĩa không?

Question 2. Distinguish the t-test on individual coefficients from the F-test of overall significance in multiple regression. Can all t-statistics be insignificant while F is significant?

- Kiểm định t: $H_0: \beta_j = 0$ (hoặc $\beta_j = \beta^*$), thống kê $t = (\hat{\beta}_j - \beta^*)/se(\hat{\beta}_j) \sim t(n - k)$; kiểm tra tác động riêng của một biến khi các biến khác giữ nguyên.
t-test: $H_0: \beta_j = 0$ (or $\beta_j = \beta^$), statistic $t = (\hat{\beta}_j - \beta^*)/se(\hat{\beta}_j) \sim t(n - k)$; it checks the partial effect of one regressor holding the others fixed.*
- Kiểm định F: $H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$, $F = [R^2/(k - 1)]/[(1 - R^2)/(n - k)] \sim F(k - 1, n - k)$; kiểm tra các biến giải thích có cùng giải thích được Y hay không.
F-test: $H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$, $F = [R^2/(k - 1)]/[(1 - R^2)/(n - k)] \sim F(k - 1, n - k)$; it checks whether the regressors jointly explain Y.
- Hồi quy hai biến ($k = 2$): $F = t^2$ nên hai kiểm định cho cùng kết luận. Từ $k \geq 3$ hai kiểm định trả lời hai câu hỏi khác nhau.
With one regressor ($k = 2$): $F = t^2$, so both tests reach the same conclusion. From $k \geq 3$ they answer different questions.
- Có thể xảy ra: khi các biến giải thích tương quan cao với nhau (đa cộng tuyến, Chương 6), se của từng hệ số lớn nên t nhỏ, nhưng nhóm biến vẫn giải thích phần lớn biến động của Y nên R^2 và F lớn.
Yes: when regressors are highly correlated (multicollinearity, Chapter 6), each coefficient's se is large and t is small, yet the group still explains most of the variation in Y, so R^2 and F are large.
- Kết luận đúng trong trường hợp đó: các biến cùng có ảnh hưởng tới Y nhưng số liệu không tách được tác động riêng của từng biến; không được bỏ cùng lúc mọi biến có t nhỏ.
The correct reading is that the regressors jointly affect Y but the data cannot separate their individual effects; do not drop all low-t variables at once.

Câu 3. Trình bày các bước kiểm định Wald thu hẹp và kiểm định một ràng buộc tuyến tính (ví dụ $\beta_2 + \beta_3 = 1$). Khi nào hai cách cho cùng kết quả?

Question 3. Set out the steps of the Wald test for dropping a group of variables and of the test of a linear restriction (for example $\beta_2 + \beta_3 = 1$). When do the two approaches give the same result?

- Ước lượng mô hình không ràng buộc (U), lấy $RSS(U)$ và bậc tự do $n - k$; viết ràng buộc dưới H_0 , đếm số ràng buộc m.
Estimate the unrestricted model (U) and record $RSS(U)$ and $n - k$; write the restrictions under H_0 and count them, m.
- Ước lượng mô hình ràng buộc (R): bỏ nhóm biến khi H_0 cho các hệ số bằng 0; thế ràng buộc vào mô hình khi H_0 là đẳng thức tuyến tính, ví dụ $\beta_3 = 1 - \beta_2$ dẫn tới hồi quy $\ln(Q/L)$ theo $\ln(K/L)$.
Estimate the restricted model (R): drop the group when H_0 sets the coefficients to zero;

substitute the restriction when H_0 is a linear equality, e.g. $\beta_3 = 1 - \beta_2$ leads to regressing $\ln(Q/L)$ on $\ln(K/L)$.

- $F = [(RSS(R) - RSS(U))/m]/[RSS(U)/(n - k)]$; so với $F(\alpha; m, n - k)$; F lớn hơn thì bác bỏ H_0 . Dạng theo R^2 chỉ dùng được khi hai mô hình cùng biến phụ thuộc.

$F = [(RSS(R) - RSS(U))/m]/[RSS(U)/(n - k)]$; compare with $F(\alpha; m, n - k)$; reject H_0 if F is larger. The R^2 form is valid only when both models share the dependent variable.

- Một ràng buộc ($m = 1$) còn kiểm định được bằng t với hiệp phương sai: $t = (\beta_2 + \beta_3 - 1)/\sqrt{(\text{Var } \beta_2 + \text{Var } \beta_3 + 2\text{Cov})}$ và $t^2 = F$.

A single restriction ($m = 1$) can also be tested by t using the covariance: $t = (\beta_2 + \beta_3 - 1)/\sqrt{(\text{Var } \beta_2 + \text{Var } \beta_3 + 2\text{Cov})}$, and $t^2 = F$.

- Hai cách cho cùng kết quả khi $m = 1$; với $m \geq 2$ chỉ dùng F (hoặc lệnh Wald của phần mềm).

The two approaches agree when $m = 1$; with $m \geq 2$ only the F form (or the software's Wald command) applies.

4.15. Câu hỏi trắc nghiệm — Multiple-choice questions

25. Hồi quy doanh thu 36 cửa hàng Hạt Nâu theo GIA, QC, DT (có hệ số chặn). Kiểm định t cho hệ số của QC dùng phân phối t với bậc tự do bằng bao nhiêu?

Revenue of 36 Hạt Nâu shops is regressed on GIA, QC and DT with an intercept. How many degrees of freedom does the t -test on the QC coefficient use?

- A. 35
- B. 33
- C. 32
- D. 34

26. Mô hình có $n = 36$, $k = 4$ (kể cả hệ số chặn), $R^2 = 0,6655$. R^2 hiệu chỉnh xấp xỉ:

A model has $n = 36$, $k = 4$ (including the intercept) and $R^2 = 0.6655$. Adjusted R^2 is about:

- A. 0,616 / 0.616
- B. 0,634 / 0.634
- C. 0,666 / 0.666
- D. 0,645 / 0.645