
ECON3371

Applied Econometric Methods and Data Analysis

Notes

Built exclusively from the uploaded ECON3371 course materials: Lectures 1 to 11, Homework 1 to 4 solutions, Test 1 to 4 solutions, Test 2 to 4 preparation reviews, and Test 3 to 4 practice papers.

Every formula, definition, assumption, interpretation, and worked figure in this document is traceable to a specific location in those files. Where a point is not supported by the materials, it is flagged explicitly rather than invented. Lectures 10 (Limited Dependent Variables) and 11 (LASSO) appear in the lecture set but are not assessed by any uploaded test; they are included in full and flagged accordingly.

- Test 1** Lectures 1 to 2, Week 4
- Test 2** Lectures 3 to 4, Week 8
- Test 3** Lectures 5 to 7, Week 11
- Test 4** Lectures 8 to 9, Week 13

Contents

How to use these notes		3
1 1. Simple and Multiple Linear Regression	Test 1	3
1.1 Concept Overview		3
1.2 Core Definitions		3
1.3 Equations and Formula Sheet		4
1.4 Assumptions and Why They Matter		4
1.5 Functional Forms and Interpretation		4
1.6 Step-by-Step Procedure		5
1.7 Worked Example (Lecture 1, HW1 Q5, Test 1 Q2)		5
1.8 Common Mistakes		5
1.9 Exam Answer Blueprint		6
2 2. Hypothesis Testing	Test 1	6
2.1 Core Definitions		6
2.2 Equations and Formula Sheet		6
2.3 Interpretation Framework		6
2.4 Worked Examples (Lecture 2, Test 1 Q4)		7
2.5 Common Mistakes and Exam Blueprint		7
3 3. Pooled Cross-Sections and Difference-in-Differences	Test 2	7
3.1 Concept and Pooled Cross-Section Model		7
3.2 Difference-in-Differences		7
3.3 Worked Example: Garbage Incinerator		8
3.4 Common Mistakes		8
4 4. Panel Data: Fixed Effects, First Differencing, Random Effects, Hausman	Test 2	8
4.1 The Fixed Effects Model		8
4.2 Removing the Fixed Effect		9
4.3 Random Effects and the Hausman Test		9
4.4 Worked Examples (Lectures 3 to 4, HW2)		9
4.5 Common Mistakes and Exam Blueprint		9
5 5. Endogeneity and Instrumental Variables	Test 3	10
5.1 Concept and Core Definitions		10
5.2 The IV Estimator and Its Variance		10
5.3 Multiple Regression and Identification		10
6 6. Two-Stage Least Squares and IV Testing	Test 3	11
6.1 2SLS Algorithm and Variance		11
6.2 Testing Endogeneity (Hausman / control function)		11

6.3	Testing Overidentifying Restrictions (Sargan)	11
7	7. Simultaneous Equations Models	Test 3 12
7.1	Concept, Reduced Form, Identification	12
7.2	Common Mistakes and Exam Blueprint	12
8	8. Regression Discontinuity Design	Test 4 12
8.1	Concept and Setup	12
8.2	Estimands and Identification	13
8.3	Estimation	13
8.4	Diagnostics	13
8.5	Worked Examples (MLDA, HW4, Test 4)	13
9	9. Matching and Propensity Scores	Test 4 14
9.1	Concept and Core Definitions	14
9.2	Matching Estimators (all weighted averages)	14
9.3	Balance Diagnostics (SMD)	15
9.4	Workflow, Refinements, and What Goes Wrong	15
9.5	Worked Example (Test 4 Q3) and Blueprint	15
10	10. Limited Dependent Variable Models	Lecture only 15
10.1	Why OLS Fails and the Binary Models	15
10.2	Tobit, Poisson, and Heckman	16
11	11. High-Dimensional Econometrics (LASSO)	Lecture only 16
11.1	The LASSO Estimator	16
11.2	Valid Inference for a Target Effect	16
12	Homework Intelligence	17
13	Test Intelligence	17
14	High-Yield Revision: the 20 percent that earns 80 percent	18
15	One-Night-Before-Exam Guide	19
16	Ultimate Cheat Sheet	19

How to use these notes

The course is organised as four assessed blocks, each tested by a 40 minute, 100 point test that allows one double-sided A4 cheat sheet and a calculator. The mapping that the test solutions reveal is below.

Test	Week	Lectures	Core machinery assessed
Test 1	4	1 to 2	OLS feasibility, omitted variable bias sign, t-tests, confidence intervals, F-tests, interaction terms, R squared
Test 2	8	3 to 4	Pooled OLS bias, first differencing, within transformation, DiD, parallel trends, FE vs RE, Hausman
Test 3	11	5 to 7	Endogeneity, IV conditions, first stage, 2SLS, endogeneity test, overidentification test, simultaneous equations
Test 4	13	8 to 9	Sharp and fuzzy RDD, jumps and slopes, McCrary, ATT, unconfoundedness, propensity score, SMD balance

Lectures 10 (Limited Dependent Variables) and 11 (LASSO) appear in the lecture set but no test in the uploaded files assesses them. They are included in full for completeness and likely final-examination relevance, but treat the four test blocks above as the proven assessment core. Badges such as **Test 3** flag where a result has actually been examined in the uploaded papers.

1. Simple and Multiple Linear Regression

Test 1

The foundation of the whole unit. Everything later (panel, IV, RDD, matching, LDV, LASSO) is built on the OLS machinery and the exogeneity assumption introduced here.

Concept Overview

Econometrics bridges economic questions and data. It applies statistical models to data to answer quantitative economic problems. The central distinction the unit returns to repeatedly is correlation versus causality. Correlation is the degree to which two variables move together; causality is the direct effect of one variable on another holding other factors fixed. Correlation does not imply causality. Data are rarely ideal because there is usually no natural random experiment (students who attend college often have better family backgrounds) and there are measurement issues (rounding error, misreporting). Three data types: cross-sectional (different units at one time), time series (one unit over many periods), and panel data (multiple units over multiple periods).

Core Definitions

- **Dependent variable** y_i : the outcome to be explained. **Regressor** x_i : the variable used to explain it.
- **Error term** u_i : an unobserved variable capturing all other factors.
- **Coefficients** β_0, β_1 : population quantities to be estimated; estimates are $\hat{\beta}_0, \hat{\beta}_1$.
- **Fitted value** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. **Residual** $\hat{u}_i = y_i - \hat{y}_i$.
- **Marginal effect**: with other factors fixed ($\Delta u = 0$), $\Delta y = \beta_1 \Delta x$, so β_1 is the change in y per one-unit change in x .
- **Dummy variable**: takes only 1 or 0; its coefficient is a **treatment effect**.

Equations and Formula Sheet

Models. Simple: $y_i = \beta_0 + \beta_1 x_i + u_i$. Multiple: $y = \beta_0 + \beta_1 x_1 + \dots + \beta_K x_K + u$.
OLS (least squares). $SSR = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$; OLS minimises SSR:

$$(\hat{\beta}_0, \dots, \hat{\beta}_K) = \arg \min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_K x_{iK})^2.$$

Closed-form slope (stated in the IV lecture as the OLS comparison):

$$\hat{\beta}_1^{OLS} = \frac{\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_i (x_i - \bar{x})^2}.$$

Standard errors. Simple: $SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \cdot \frac{\sigma^2}{\text{Var}(x_i)}}$. Multiple:

$$SE(\hat{\beta}_k) = \sqrt{\frac{1}{n} \cdot \frac{\sigma^2}{\text{Var}(x_{ik}) (1 - R_k^2)}}, \quad \text{VIF}_k = \frac{1}{1 - R_k^2}.$$

Goodness of fit. $SST = \sum (y_i - \bar{y})^2$, $R^2 = SSE/SST = 1 - SSR/SST$,

$$\bar{R}^2 = 1 - \frac{SSR/(n - K - 1)}{SST/(n - 1)}.$$

Omitted variable bias. True $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, omit x_2 :

$$\text{Bias} = \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)}, \quad \text{sgn}(\text{Bias}) = \text{sgn}(\beta_2) \cdot \text{sgn}(\text{corr}(x_1, x_2)).$$

Assumptions and Why They Matter

Assumption	Statement	Role / consequence if violated
1. Exogeneity	$E[u x] = 0$, so $\text{Cov}(u, x_j) = 0$	Required for unbiasedness. If violated the regressor is endogenous and OLS is biased and inconsistent. Adding regressors makes this more realistic.
2. Homoskedasticity	$\text{Var}(u x) = \sigma^2$	Error variance does not depend on x . If it does, heteroskedasticity affects the standard error formula.
3. No perfect multicollinearity	No regressor is an exact linear combination of the others	Ensures $R_k^2 \neq 1$ so coefficients are separately identified.

Properties of OLS. The regression line passes through the sample means; residuals sum to zero; under Assumptions 1 to 2 OLS is unbiased, $E[\hat{\beta}_k] = \beta_k$. The variance of $\hat{\beta}_k$ is smaller with larger n , smaller σ^2 , larger $\text{Var}(x)$, and smaller R_k^2 .

Functional Forms and Interpretation

Form	Model	Interpretation of β_1
Level-level	$y = \beta_0 + \beta_1 x + u$	One unit higher x changes y by β_1 units.