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

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

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

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

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

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

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

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

$$\mathrm{Bias} = \beta_2 \frac{\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)}, \qquad \mathrm{sgn}(\mathrm{Bias}) = \mathrm{sgn}(\beta_2) \cdot \mathrm{sgn}\left(\mathrm{corr}(x_1, x_2)\right).$$

Assumptions and Why They Matter

Assumption	Statement	Role / consequence if violated
1. Exogeneity	$E[u \mid x] = 0$, so $\mathrm{Cov}(u, x_j) = 0$	Required for unbiasedness. If violated the regressor is endogenous and OLS is biased and inconsistent. Adding regressors makes this more realistic.
2. Homoskedasticity	$\mathrm{Var}(u \mid x) = \sigma^2$	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 $\mathrm{Var}(x)$, 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.