
Contents

List of Figures	xvii
Preface	xix
1 Finite-Sample Properties of OLS	3
1.1 The Classical Linear Regression Model	3
The Linearity Assumption	4
Matrix Notation	6
The Strict Exogeneity Assumption	7
Implications of Strict Exogeneity	8
Strict Exogeneity in Time-Series Models	9
Other Assumptions of the Model	10
The Classical Regression Model for Random Samples	12
“Fixed” Regressors	13
1.2 The Algebra of Least Squares	15
OLS Minimizes the Sum of Squared Residuals	15
Normal Equations	16
Two Expressions for the OLS Estimator	18
More Concepts and Algebra	18
Influential Analysis (optional)	21
A Note on the Computation of OLS Estimates	23
1.3 Finite-Sample Properties of OLS	27
Finite-Sample Distribution of \mathbf{b}	27
Finite-Sample Properties of s^2	30
Estimate of $\text{Var}(\mathbf{b} \mid \mathbf{X})$	31
1.4 Hypothesis Testing under Normality	33
Normally Distributed Error Terms	33
Testing Hypotheses about Individual Regression Coefficients	35
Decision Rule for the t -Test	37
Confidence Interval	38

	<i>p</i> -Value	38
	Linear Hypotheses	39
	The <i>F</i> -Test	40
	A More Convenient Expression for <i>F</i>	42
	<i>t</i> versus <i>F</i>	43
	An Example of a Test Statistic Whose Distribution Depends on X	45
1.5	Relation to Maximum Likelihood	47
	The Maximum Likelihood Principle	47
	Conditional versus Unconditional Likelihood	47
	The Log Likelihood for the Regression Model	48
	ML via Concentrated Likelihood	48
	Cramer-Rao Bound for the Classical Regression Model	49
	The <i>F</i> -Test as a Likelihood Ratio Test	52
	Quasi-Maximum Likelihood	53
1.6	Generalized Least Squares (GLS)	54
	Consequence of Relaxing Assumption 1.4	55
	Efficient Estimation with Known V	55
	A Special Case: Weighted Least Squares (WLS)	58
	Limiting Nature of GLS	58
1.7	Application: Returns to Scale in Electricity Supply	60
	The Electricity Supply Industry	60
	The Data	60
	Why Do We Need Econometrics?	61
	The Cobb-Douglas Technology	62
	How Do We Know Things Are Cobb-Douglas?	63
	Are the OLS Assumptions Satisfied?	64
	Restricted Least Squares	65
	Testing the Homogeneity of the Cost Function	65
	Detour: A Cautionary Note on R^2	67
	Testing Constant Returns to Scale	67
	Importance of Plotting Residuals	68
	Subsequent Developments	68
	Problem Set	71
	Answers to Selected Questions	84
2	Large-Sample Theory	88
2.1	Review of Limit Theorems for Sequences of Random Variables	88
	Various Modes of Convergence	89
	Three Useful Results	92

Viewing Estimators as Sequences of Random Variables	94
Laws of Large Numbers and Central Limit Theorems	95
2.2 Fundamental Concepts in Time-Series Analysis	97
Need for Ergodic Stationarity	97
Various Classes of Stochastic Processes	98
Different Formulation of Lack of Serial Dependence	106
The CLT for Ergodic Stationary Martingale Differences Sequences	106
2.3 Large-Sample Distribution of the OLS Estimator	109
The Model	109
Asymptotic Distribution of the OLS Estimator	113
s^2 Is Consistent	115
2.4 Hypothesis Testing	117
Testing Linear Hypotheses	117
The Test Is Consistent	119
Asymptotic Power	120
Testing Nonlinear Hypotheses	121
2.5 Estimating $E(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$ Consistently	123
Using Residuals for the Errors	123
Data Matrix Representation of \mathbf{S}	125
Finite-Sample Considerations	125
2.6 Implications of Conditional Homoskedasticity	126
Conditional versus Unconditional Homoskedasticity	126
Reduction to Finite-Sample Formulas	127
Large-Sample Distribution of t and F Statistics	128
Variations of Asymptotic Tests under Conditional Homoskedasticity	129
2.7 Testing Conditional Homoskedasticity	131
2.8 Estimation with Parameterized Conditional Heteroskedasticity (optional)	133
The Functional Form	133
WLS with Known α	134
Regression of e_i^2 on \mathbf{z}_i Provides a Consistent Estimate of α	135
WLS with Estimated α	136
OLS versus WLS	137
2.9 Least Squares Projection	137
Optimally Predicting the Value of the Dependent Variable	138
Best Linear Predictor	139
OLS Consistently Estimates the Projection Coefficients	140

2.10	Testing for Serial Correlation	141
	Box-Pierce and Ljung-Box	142
	Sample Autocorrelations Calculated from Residuals	144
	Testing with Predetermined, but Not Strictly Exogenous, Regressors	146
	An Auxiliary Regression-Based Test	147
2.11	Application: Rational Expectations Econometrics	150
	The Efficient Market Hypotheses	150
	Testable Implications	152
	Testing for Serial Correlation	153
	Is the Nominal Interest Rate the Optimal Predictor? R_t Is Not Strictly Exogenous	156
	Subsequent Developments	159
2.12	Time Regressions	160
	The Asymptotic Distribution of the OLS Estimator	161
	Hypothesis Testing for Time Regressions	163
	Appendix 2.A: Asymptotics with Fixed Regressors	164
	Appendix 2.B: Proof of Proposition 2.10	165
	Problem Set	168
	Answers to Selected Questions	183
3	Single-Equation GMM	186
3.1	Endogeneity Bias: Working's Example	187
	A Simultaneous Equations Model of Market Equilibrium	187
	Endogeneity Bias	188
	Observable Supply Shifters	189
3.2	More Examples	193
	A Simple Macroeconometric Model	193
	Errors-in-Variables	194
	Production Function	196
3.3	The General Formulation	198
	Regressors and Instruments	198
	Identification	200
	Order Condition for Identification	202
	The Assumption for Asymptotic Normality	202
3.4	Generalized Method of Moments Defined	204
	Method of Moments	205
	Generalized Method of Moments	206
	Sampling Error	207

3.5	Large-Sample Properties of GMM	208
	Asymptotic Distribution of the GMM Estimator	209
	Estimation of Error Variance	210
	Hypothesis Testing	211
	Estimation of S	212
	Efficient GMM Estimator	212
	Asymptotic Power	214
	Small-Sample Properties	215
3.6	Testing Overidentifying Restrictions	217
	Testing Subsets of Orthogonality Conditions	218
3.7	Hypothesis Testing by the Likelihood-Ratio Principle	222
	The LR Statistic for the Regression Model	223
	Variable Addition Test (optional)	224
3.8	Implications of Conditional Homoskedasticity	225
	Efficient GMM Becomes 2SLS	226
	J Becomes Sargan's Statistic	227
	Small-Sample Properties of 2SLS	229
	Alternative Derivations of 2SLS	229
	When Regressors Are Predetermined	231
	Testing a Subset of Orthogonality Conditions	232
	Testing Conditional Homoskedasticity	234
	Testing for Serial Correlation	234
3.9	Application: Returns from Schooling	236
	The NLS-Y Data	236
	The Semi-Log Wage Equation	237
	Omitted Variable Bias	238
	IQ as the Measure of Ability	239
	Errors-in-Variables	239
	2SLS to Correct for the Bias	242
	Subsequent Developments	243
	Problem Set	244
	Answers to Selected Questions	254
4	Multiple-Equation GMM	258
4.1	The Multiple-Equation Model	259
	Linearity	259
	Stationarity and Ergodicity	260
	Orthogonality Conditions	261
	Identification	262

The Assumption for Asymptotic Normality	264
Connection to the “Complete” System of Simultaneous Equations	265
4.2 Multiple-Equation GMM Defined	265
4.3 Large-Sample Theory	268
4.4 Single-Equation versus Multiple-Equation Estimation	271
When Are They “Equivalent”?	272
Joint Estimation Can Be Hazardous	273
4.5 Special Cases of Multiple-Equation GMM: FIVE, 3SLS, and SUR	274
Conditional Homoskedasticity	274
Full-Information Instrumental Variables Efficient (FIVE)	275
Three-Stage Least Squares (3SLS)	276
Seemingly Unrelated Regressions (SUR)	279
SUR versus OLS	281
4.6 Common Coefficients	286
The Model with Common Coefficients	286
The GMM Estimator	287
Imposing Conditional Homoskedasticity	288
Pooled OLS	290
Beautifying the Formulas	292
The Restriction That Isn’t	293
4.7 Application: Interrelated Factor Demands	296
The Translog Cost Function	296
Factor Shares	297
Substitution Elasticities	298
Properties of Cost Functions	299
Stochastic Specifications	300
The Nature of Restrictions	301
Multivariate Regression Subject to Cross-Equation Restrictions	302
Which Equation to Delete?	304
Results	305
Problem Set	308
Answers to Selected Questions	320
5 Panel Data	323
5.1 The Error-Components Model	324
Error Components	324
Group Means	327
A Reparameterization	327
5.2 The Fixed-Effects Estimator	330

The Formula	330
Large-Sample Properties	331
Digression: When η_i Is Spherical	333
Random Effects versus Fixed Effects	334
Relaxing Conditional Homoskedasticity	335
5.3 Unbalanced Panels (optional)	337
“Zeroing Out” Missing Observations	338
Zeroing Out versus Compression	339
No Selectivity Bias	340
5.4 Application: International Differences in Growth Rates	342
Derivation of the Estimation Equation	342
Appending the Error Term	343
Treatment of α_i	344
Consistent Estimation of Speed of Convergence	345
Appendix 5.A: Distribution of Hausman Statistic	346
Problem Set	349
Answers to Selected Questions	363
6 Serial Correlation	365
6.1 Modeling Serial Correlation: Linear Processes	365
MA(q)	366
MA(∞) as a Mean Square Limit	366
Filters	369
Inverting Lag Polynomials	372
6.2 ARMA Processes	375
AR(1) and Its MA(∞) Representation	376
Autocovariances of AR(1)	378
AR(p) and Its MA(∞) Representation	378
ARMA(p, q)	380
ARMA(p, q) with Common Roots	382
Invertibility	383
Autocovariance-Generating Function and the Spectrum	383
6.3 Vector Processes	387
6.4 Estimating Autoregressions	392
Estimation of AR(1)	392
Estimation of AR(p)	393
Choice of Lag Length	394
Estimation of VARs	397
Estimation of ARMA(p, q)	398

6.5	Asymptotics for Sample Means of Serially Correlated Processes	400
	LLN for Covariance-Stationary Processes	401
	Two Central Limit Theorems	402
	Multivariate Extension	404
6.6	Incorporating Serial Correlation in GMM	406
	The Model and Asymptotic Results	406
	Estimating \mathbf{S} When Autocovariances Vanish after Finite Lags	407
	Using Kernels to Estimate \mathbf{S}	408
	VARHAC	410
6.7	Estimation under Conditional Homoskedasticity (Optional)	413
	Kernel-Based Estimation of \mathbf{S} under Conditional Homoskedasticity	413
	Data Matrix Representation of Estimated Long-Run Variance	414
	Relation to GLS	415
6.8	Application: Forward Exchange Rates as Optimal Predictors	418
	The Market Efficiency Hypothesis	419
	Testing Whether the Unconditional Mean Is Zero	420
	Regression Tests	423
	Problem Set	428
	Answers to Selected Questions	441
7	Extremum Estimators	445
7.1	Extremum Estimators	446
	“Measurability” of $\hat{\theta}$	446
	Two Classes of Extremum Estimators	447
	Maximum Likelihood (ML)	448
	Conditional Maximum Likelihood	450
	Invariance of ML	452
	Nonlinear Least Squares (NLS)	453
	Linear and Nonlinear GMM	454
7.2	Consistency	456
	Two Consistency Theorems for Extremum Estimators	456
	Consistency of M-Estimators	458
	Concavity after Reparameterization	461
	Identification in NLS and ML	462
	Consistency of GMM	467
7.3	Asymptotic Normality	469
	Asymptotic Normality of M-Estimators	470
	Consistent Asymptotic Variance Estimation	473
	Asymptotic Normality of Conditional ML	474

Two Examples	476
Asymptotic Normality of GMM	478
GMM versus ML	481
Expressing the Sampling Error in a Common Format	483
7.4 Hypothesis Testing	487
The Null Hypothesis	487
The Working Assumptions	489
The Wald Statistic	489
The Lagrange Multiplier (LM) Statistic	491
The Likelihood Ratio (LR) Statistic	493
Summary of the Trinity	494
7.5 Numerical Optimization	497
Newton-Raphson	497
Gauss-Newton	498
Writing Newton-Raphson and Gauss-Newton in a Common Format	498
Equations Nonlinear in Parameters Only	499
Problem Set	501
Answers to Selected Questions	505
8 Examples of Maximum Likelihood	507
8.1 Qualitative Response (QR) Models	507
Score and Hessian for Observation t	508
Consistency	509
Asymptotic Normality	510
8.2 Truncated Regression Models	511
The Model	511
Truncated Distributions	512
The Likelihood Function	513
Reparameterizing the Likelihood Function	514
Verifying Consistency and Asymptotic Normality	515
Recovering Original Parameters	517
8.3 Censored Regression (Tobit) Models	518
Tobit Likelihood Function	518
Reparameterization	519
8.4 Multivariate Regressions	521
The Multivariate Regression Model Restated	522
The Likelihood Function	523
Maximizing the Likelihood Function	524

	Consistency and Asymptotic Normality	525
8.5	FIML	526
	The Multiple-Equation Model with Common Instruments Restated	526
	The Complete System of Simultaneous Equations	529
	Relationship between $(\mathbf{\Gamma}_0, \mathbf{B}_0)$ and δ_0	530
	The FIML Likelihood Function	531
	The FIML Concentrated Likelihood Function	532
	Testing Overidentifying Restrictions	533
	Properties of the FIML Estimator	533
	ML Estimation of the SUR Model	535
8.6	LIML	538
	LIML Defined	538
	Computation of LIML	540
	LIML versus 2SLS	542
8.7	Serially Correlated Observations	543
	Two Questions	543
	Unconditional ML for Dependent Observations	545
	ML Estimation of AR(1) Processes	546
	Conditional ML Estimation of AR(1) Processes	547
	Conditional ML Estimation of AR(p) and VAR(p) Processes	549
	Problem Set	551
9	Unit-Root Econometrics	557
9.1	Modeling Trends	557
	Integrated Processes	558
	Why Is It Important to Know if the Process Is I(1)?	560
	Which Should Be Taken as the Null, I(0) or I(1)?	562
	Other Approaches to Modeling Trends	563
9.2	Tools for Unit-Root Econometrics	563
	Linear I(0) Processes	563
	Approximating I(1) by a Random Walk	564
	Relation to ARMA Models	566
	The Wiener Process	567
	A Useful Lemma	570
9.3	Dickey-Fuller Tests	573
	The AR(1) Model	573
	Deriving the Limiting Distribution under the I(1) Null	574
	Incorporating the Intercept	577
	Incorporating Time Trend	581

9.4	Augmented Dickey-Fuller Tests	585
	The Augmented Autoregression	585
	Limiting Distribution of the OLS Estimator	586
	Deriving Test Statistics	590
	Testing Hypotheses about ζ	591
	What to Do When p Is Unknown?	592
	A Suggestion for the Choice of $p_{max}(T)$	594
	Including the Intercept in the Regression	595
	Incorporating Time Trend	597
	Summary of the DF and ADF Tests and Other Unit-Root Tests	599
9.5	Which Unit-Root Test to Use?	601
	Local-to-Unity Asymptotics	602
	Small-Sample Properties	602
9.6	Application: Purchasing Power Parity	603
	The Embarrassing Resiliency of the Random Walk Model?	604
	Problem Set	605
	Answers to Selected Questions	619
10	Cointegration	623
10.1	Cointegrated Systems	624
	Linear Vector I(0) and I(1) Processes	624
	The Beveridge-Nelson Decomposition	627
	Cointegration Defined	629
10.2	Alternative Representations of Cointegrated Systems	633
	Phillips's Triangular Representation	633
	VAR and Cointegration	636
	The Vector Error-Correction Model (VECM)	638
	Johansen's ML Procedure	640
10.3	Testing the Null of No Cointegration	643
	Spurious Regressions	643
	The Residual-Based Test for Cointegration	644
	Testing the Null of Cointegration	649
10.4	Inference on Cointegrating Vectors	650
	The SOLS Estimator	650
	The Bivariate Example	652
	Continuing with the Bivariate Example	653
	Allowing for Serial Correlation	654
	General Case	657
	Other Estimators and Finite-Sample Properties	658

10.5 Application: The Demand for Money in the United States	659
The Data	660
$(m - p, y, R)$ as a Cointegrated System	660
DOLS	662
Unstable Money Demand?	663
Problem Set	665
Appendix A: Partitioned Matrices and Kronecker Products	670
Addition and Multiplication of Partitioned Matrices	671
Inverting Partitioned Matrices	672
Index	675

COPYRIGHT NOTICE:

Fumio Hayashi: *Econometrics*

is published by Princeton University Press and copyrighted, © 2000, by Princeton University Press. All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher, except for reading and browsing via the World Wide Web. Users are not permitted to mount this file on any network servers.

For COURSE PACK and other PERMISSIONS, refer to entry on previous page. For more information, send e-mail to permissions@pupress.princeton.edu

CHAPTER 1

Finite-Sample Properties of OLS

ABSTRACT

The **Ordinary Least Squares** (OLS) estimator is the most basic estimation procedure in econometrics. This chapter covers the **finite-** or **small-sample properties** of the OLS estimator, that is, the statistical properties of the OLS estimator that are valid for any given sample size. The materials covered in this chapter are entirely standard. The exposition here differs from that of most other textbooks in its emphasis on the role played by the assumption that the regressors are “strictly exogenous.”

In the final section, we apply the finite-sample theory to the estimation of the cost function using cross-section data on individual firms. The question posed in Nerlove’s (1963) study is of great practical importance: are there increasing returns to scale in electricity supply? If yes, microeconomics tells us that the industry should be regulated. Besides providing you with a hands-on experience of using the techniques to test interesting hypotheses, Nerlove’s paper has a careful discussion of why the OLS is an appropriate estimation procedure in this particular application.

1.1 The Classical Linear Regression Model

In this section we present the assumptions that comprise the classical linear regression model. In the model, the variable in question (called the **dependent variable**, the **regressand**, or more generically the **left-hand [-side] variable**) is related to several other variables (called the **regressors**, the **explanatory variables**, or the **right-hand [-side] variables**). Suppose we observe n values for those variables. Let y_i be the i -th observation of the dependent variable in question and let $(x_{i1}, x_{i2}, \dots, x_{iK})$ be the i -th observation of the K regressors. The **sample** or **data** is a collection of those n observations.

The data in economics cannot be generated by experiments (except in experimental economics), so both the dependent and independent variables have to be treated as random variables, variables whose values are subject to chance. A **model**

is a set of restrictions on the joint distribution of the dependent and independent variables. That is, a model is a set of joint distributions satisfying a set of assumptions. The classical regression model is a set of joint distributions satisfying Assumptions 1.1–1.4 stated below.

The Linearity Assumption

The first assumption is that the relationship between the dependent variable and the regressors is linear.

Assumption 1.1 (linearity):

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1, 2, \dots, n), \quad (1.1.1)$$

where β 's are unknown parameters to be estimated, and ε_i is the unobserved error term with certain properties to be specified below.

The part of the right-hand side involving the regressors, $\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$, is called the **regression** or the **regression function**, and the coefficients (β 's) are called the **regression coefficients**. They represent the marginal and separate effects of the regressors. For example, β_2 represents the change in the dependent variable when the second regressor increases by one unit while other regressors are held constant. In the language of calculus, this can be expressed as $\partial y_i / \partial x_{i2} = \beta_2$. The linearity implies that the marginal effect does not depend on the level of regressors. The error term represents the part of the dependent variable left unexplained by the regressors.

Example 1.1 (consumption function): The simple consumption function familiar from introductory economics is

$$CON_i = \beta_1 + \beta_2 YD_i + \varepsilon_i, \quad (1.1.2)$$

where CON is consumption and YD is disposable income. If the data are annual aggregate time-series, CON_i and YD_i are aggregate consumption and disposable income for year i . If the data come from a survey of individual households, CON_i is consumption by the i -th household in the cross-section sample of n households. The consumption function can be written as (1.1.1) by setting $y_i = CON_i$, $x_{i1} = 1$ (a constant), and $x_{i2} = YD_i$. The error term ε_i represents other variables besides disposable income that influence consumption. They include those variables — such as financial assets — that

might be observable but the researcher decided not to include as regressors, as well as those variables — such as the “mood” of the consumer — that are hard to measure. When the equation has only one nonconstant regressor, as here, it is called the **simple regression model**.

The linearity assumption is not as restrictive as it might first seem, because the dependent variable and the regressors can be transformations of the variables in question. Consider

Example 1.2 (wage equation): A simplified version of the wage equation routinely estimated in labor economics is

$$\log(WAGE_i) = \beta_1 + \beta_2 S_i + \beta_3 TENURE_i + \beta_4 EXPR_i + \varepsilon_i, \quad (1.1.3)$$

where $WAGE$ = the wage rate for the individual, S = education in years, $TENURE$ = years on the current job, and $EXPR$ = experience in the labor force (i.e., total number of years to date on all the jobs held currently or previously by the individual). The wage equation fits the generic format (1.1.1) with $y_i = \log(WAGE_i)$. The equation is said to be in the **semi-log** form because only the dependent variable is in logs. The equation is derived from the following nonlinear relationship between the level of the wage rate and the regressors:

$$WAGE_i = \exp(\beta_1) \exp(\beta_2 S_i) \exp(\beta_3 TENURE_i) \exp(\beta_4 EXPR_i) \exp(\varepsilon_i). \quad (1.1.4)$$

By taking logs of both sides of (1.1.4) and noting that $\log[\exp(x)] = x$, one obtains (1.1.3). The coefficients in the semi-log form have the interpretation of *percentage changes*, not changes in levels. For example, a value of 0.05 for β_2 implies that an additional year of education has the effect of raising the wage rate by 5 percent. The difference in the interpretation comes about because the dependent variable is the log wage rate, not the wage rate itself, and the change in logs equals the percentage change in levels.

Certain other forms of nonlinearities can also be accommodated. Suppose, for example, the marginal effect of education tapers off as the level of education gets higher. This can be captured by including in the wage equation the squared term S^2 as an additional regressor in the wage equation. If the coefficient of the squared

term is β_5 , the marginal effect of education is

$$\beta_2 + 2\beta_5 S \quad (= \partial \log(WAGE)/\partial S).$$

If β_5 is negative, the marginal effect of education declines with the level of education.

There are, of course, cases of genuine nonlinearity. For example, the relationship (1.1.4) could not have been made linear if the error term entered additively rather than multiplicatively:

$$WAGE_i = \exp(\beta_1) \exp(\beta_2 S_i) \exp(\beta_3 TENURE_i) \exp(\beta_4 EXPR_i) + \varepsilon_i.$$

Estimation of nonlinear regression equations such as this will be discussed in Chapter 7.

Matrix Notation

Before stating other assumptions of the classical model, we introduce the vector and matrix notation. The notation will prove useful for stating other assumptions precisely and also for deriving the OLS estimator of $\boldsymbol{\beta}$. Define K -dimensional (column) vectors \mathbf{x}_i and $\boldsymbol{\beta}$ as

$$\mathbf{x}_i = \begin{matrix} (K \times 1) \\ \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix} \end{matrix}, \quad \boldsymbol{\beta} = \begin{matrix} (K \times 1) \\ \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \end{matrix}. \quad (1.1.5)$$

By the definition of vector inner products, $\mathbf{x}'_i \boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$. So the equations in Assumption 1.1 can be written as

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, 2, \dots, n). \quad (1.1.1')$$

Also define

$$\mathbf{y} = \begin{matrix} (n \times 1) \\ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \end{matrix}, \quad \boldsymbol{\varepsilon} = \begin{matrix} (n \times 1) \\ \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \end{matrix}, \quad \mathbf{X} = \begin{matrix} (n \times K) \\ \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} \end{matrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix}. \quad (1.1.6)$$

In the vectors and matrices in (1.1.6), there are as many rows as there are observations, with the rows corresponding to the observations. For this reason \mathbf{y} and \mathbf{X} are sometimes called the **data vector** and the **data matrix**. Since the number of

columns of \mathbf{X} equals the number of rows of $\boldsymbol{\beta}$, \mathbf{X} and $\boldsymbol{\beta}$ are conformable and $\mathbf{X}\boldsymbol{\beta}$ is an $n \times 1$ vector. Its i -th element is $\mathbf{x}'_i \boldsymbol{\beta}$. Therefore, Assumption 1.1 can be written compactly as

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{\underbrace{(n \times K)(K \times 1)}_{(n \times 1)}}{\mathbf{X} \boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}.$$

The Strict Exogeneity Assumption

The next assumption of the classical regression model is

Assumption 1.2 (strict exogeneity):

$$E(\varepsilon_i | \mathbf{X}) = 0 \quad (i = 1, 2, \dots, n). \quad (1.1.7)$$

Here, the expectation (mean) is conditional on the regressors for *all* observations. This point may be made more apparent by writing the assumption without using the data matrix as

$$E(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n) = 0 \quad (i = 1, 2, \dots, n).$$

To state the assumption differently, take, for any given observation i , the joint distribution of the $nK + 1$ random variables, $f(\varepsilon_i, \mathbf{x}_1, \dots, \mathbf{x}_n)$, and consider the conditional distribution, $f(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n)$. The conditional mean $E(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n)$ is in general a nonlinear function of $(\mathbf{x}_1, \dots, \mathbf{x}_n)$. The strict exogeneity assumption says that this function is a constant of value zero.¹

Assuming this constant to be zero is not restrictive if the regressors include a constant, because the equation can be rewritten so that the conditional mean of the error term is zero. To see this, suppose that $E(\varepsilon_i | \mathbf{X})$ is μ and $x_{i1} = 1$. The equation can be written as

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i \\ &= (\beta_1 + \mu) + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + (\varepsilon_i - \mu). \end{aligned}$$

If we redefine β_1 to be $\beta_1 + \mu$ and ε_i to be $\varepsilon_i - \mu$, the conditional mean of the new error term is zero. In virtually all applications, the regressors include a constant term.

¹Some authors define the term “strict exogeneity” somewhat differently. For example, in Koopmans and Hood (1953) and Engle, Hendry, and Richards (1983), the regressors are strictly exogenous if \mathbf{x}_i is independent of ε_j for all i, j . This definition is stronger than, but not inconsistent with, our definition of strict exogeneity.

Example 1.3 (continuation of Example 1.1): For the simple regression model of Example 1.1, the strict exogeneity assumption can be written as

$$E(\varepsilon_i \mid YD_1, YD_2, \dots, YD_n) = 0.$$

Since $\mathbf{x}_i = (1, YD_i)'$, you might wish to write the strict exogeneity assumption as

$$E(\varepsilon_i \mid 1, YD_1, 1, YD_2, \dots, 1, YD_n) = 0.$$

But since a constant provides no information, the expectation conditional on

$$(1, YD_1, 1, YD_2, \dots, 1, YD_n)$$

is the same as the expectation conditional on

$$(YD_1, YD_2, \dots, YD_n).$$

Implications of Strict Exogeneity

The strict exogeneity assumption has several implications.

- The *unconditional* mean of the error term is zero, i.e.,

$$E(\varepsilon_i) = 0 \quad (i = 1, 2, \dots, n). \quad (1.1.8)$$

This is because, by the Law of Total Expectations from basic probability theory,² $E[E(\varepsilon_i \mid \mathbf{X})] = E(\varepsilon_i)$.

- If the cross moment $E(xy)$ of two random variables x and y is zero, then we say that x is **orthogonal** to y (or y is orthogonal to x). Under strict exogeneity, the regressors are orthogonal to the error term for *all* observations, i.e.,

$$E(x_{jk}\varepsilon_i) = 0 \quad (i, j = 1, \dots, n; k = 1, \dots, K)$$

or

$$E(\mathbf{x}_j \cdot \varepsilon_i) = \begin{bmatrix} E(x_{j1}\varepsilon_i) \\ E(x_{j2}\varepsilon_i) \\ \vdots \\ E(x_{jK}\varepsilon_i) \end{bmatrix} = \begin{matrix} \mathbf{0} \\ (K \times 1) \end{matrix} \quad (\text{for all } i, j). \quad (1.1.9)$$

²The Law of Total Expectations states that $E[E(y \mid \mathbf{x})] = E(y)$.

The proof is a good illustration of the use of properties of conditional expectations and goes as follows.

PROOF. Since x_{jk} is an element of \mathbf{X} , strict exogeneity implies

$$E(\varepsilon_i | x_{jk}) = E[E(\varepsilon_i | \mathbf{X}) | x_{jk}] = 0 \quad (1.1.10)$$

by the Law of Iterated Expectations from probability theory.³ It follows from this that

$$\begin{aligned} E(x_{jk}\varepsilon_i) &= E[E(x_{jk}\varepsilon_i | x_{jk})] \quad (\text{by the Law of Total Expectations}) \\ &= E[x_{jk} E(\varepsilon_i | x_{jk})] \quad (\text{by the linearity of conditional expectations}^4) \\ &= 0. \quad \blacksquare \end{aligned}$$

The point here is that strict exogeneity requires the regressors be orthogonal not only to the error term from the same observation (i.e., $E(x_{ik}\varepsilon_i) = 0$ for all k), but also to the error term from the other observations (i.e., $E(x_{jk}\varepsilon_i) = 0$ for all k and for $j \neq i$).

- Because the mean of the error term is zero, the orthogonality conditions (1.1.9) are equivalent to zero-correlation conditions. This is because

$$\begin{aligned} \text{Cov}(\varepsilon_i, x_{jk}) &= E(x_{jk}\varepsilon_i) - E(x_{jk}) E(\varepsilon_i) \quad (\text{by definition of covariance}) \\ &= E(x_{jk}\varepsilon_i) \quad (\text{since } E(\varepsilon_i) = 0, \text{ see (1.1.8)}) \\ &= 0 \quad (\text{by the orthogonality conditions (1.1.9)}). \end{aligned}$$

In particular, for $i = j$, $\text{Cov}(x_{ik}, \varepsilon_i) = 0$. Therefore, strict exogeneity implies the requirement (familiar to those who have studied econometrics before) that the regressors be contemporaneously uncorrelated with the error term.

Strict Exogeneity in Time-Series Models

For time-series models where i is time, the implication (1.1.9) of strict exogeneity can be rephrased as: the regressors are orthogonal to the past, current, and future error terms (or equivalently, the error term is orthogonal to the past, current, and future regressors). But for most time-series models, this condition (and *a fortiori* strict exogeneity) is not satisfied, so the finite-sample theory based on strict exogeneity to be developed in this section is rarely applicable in time-series con-

³The Law of Iterated Expectations states that $E[E(y | \mathbf{x}, \mathbf{z}) | \mathbf{x}] = E(y | \mathbf{x})$.

⁴The linearity of conditional expectations states that $E[f(\mathbf{x})y | \mathbf{x}] = f(\mathbf{x})E(y | \mathbf{x})$.

texts. However, as will be shown in the next chapter, the estimator possesses good large-sample properties without strict exogeneity.

The clearest example of a failure of strict exogeneity is a model where the regressor includes the **lagged dependent variable**. Consider the simplest such model:

$$y_i = \beta y_{i-1} + \varepsilon_i \quad (i = 1, 2, \dots, n). \quad (1.1.11)$$

This is called the **first-order autoregressive model** (AR(1)). (We will study this model more fully in Chapter 6.) Suppose, consistent with the spirit of the strict exogeneity assumption, that the regressor for observation i , y_{i-1} , is orthogonal to the error term for i so $E(y_{i-1}\varepsilon_i) = 0$. Then

$$\begin{aligned} E(y_i\varepsilon_i) &= E[(\beta y_{i-1} + \varepsilon_i)\varepsilon_i] \quad (\text{by (1.1.11)}) \\ &= \beta E(y_{i-1}\varepsilon_i) + E(\varepsilon_i^2) \\ &= E(\varepsilon_i^2) \quad (\text{since } E(y_{i-1}\varepsilon_i) = 0 \text{ by hypothesis}). \end{aligned}$$

Therefore, unless the error term is always zero, $E(y_i\varepsilon_i)$ is not zero. But y_i is the regressor for observation $i+1$. Thus, the regressor is not orthogonal to the past error term, which is a violation of strict exogeneity.

Other Assumptions of the Model

The remaining assumptions comprising the classical regression model are the following.

Assumption 1.3 (no multicollinearity): *The rank of the $n \times K$ data matrix, \mathbf{X} , is K with probability 1.*

Assumption 1.4 (spherical error variance):

$$(\text{homoskedasticity}) \quad E(\varepsilon_i^2 | \mathbf{X}) = \sigma^2 > 0 \quad (i = 1, 2, \dots, n),^5 \quad (1.1.12)$$

(no correlation between observations)

$$E(\varepsilon_i\varepsilon_j | \mathbf{X}) = 0 \quad (i, j = 1, 2, \dots, n; i \neq j). \quad (1.1.13)$$

⁵When a symbol (which here is σ^2) is given to a moment (which here is the second moment $E(\varepsilon_i^2 | \mathbf{X})$), by implication the moment is assumed to exist and is finite. We will follow this convention for the rest of this book.

To understand Assumption 1.3, recall from matrix algebra that the rank of a matrix equals the number of linearly independent columns of the matrix. The assumption says that none of the K columns of the data matrix \mathbf{X} can be expressed as a linear combination of the other columns of \mathbf{X} . That is, \mathbf{X} is of **full column rank**. Since the K columns cannot be linearly independent if their dimension is less than K , the assumption implies that $n \geq K$, i.e., there must be at least as many observations as there are regressors. The regressors are said to be **(perfectly) multicollinear** if the assumption is not satisfied. It is easy to see in specific applications when the regressors are multicollinear and what problems arise.

Example 1.4 (continuation of Example 1.2): If no individuals in the sample ever changed jobs, then $TENURE_i = EXPR_i$ for all i , in violation of the no multicollinearity assumption. There is evidently no way to distinguish the tenure effect on the wage rate from the experience effect. If we substitute this equality into the wage equation to eliminate $TENURE_i$, the wage equation becomes

$$\log(WAGE_i) = \beta_1 + \beta_2 S_i + (\beta_3 + \beta_4) EXPR_i + \varepsilon_i,$$

which shows that only the sum $\beta_3 + \beta_4$, but not β_3 and β_4 separately, can be estimated.

The homoskedasticity assumption (1.1.12) says that the conditional second moment, which in general is a nonlinear function of \mathbf{X} , is a constant. Thanks to strict exogeneity, this condition can be stated equivalently in more familiar terms. Consider the conditional variance $\text{Var}(\varepsilon_i | \mathbf{X})$. It equals the same constant because

$$\begin{aligned} \text{Var}(\varepsilon_i | \mathbf{X}) &\equiv \text{E}(\varepsilon_i^2 | \mathbf{X}) - \text{E}(\varepsilon_i | \mathbf{X})^2 \quad (\text{by definition of conditional variance}) \\ &= \text{E}(\varepsilon_i^2 | \mathbf{X}) \quad (\text{since } \text{E}(\varepsilon_i | \mathbf{X}) = 0 \text{ by strict exogeneity}). \end{aligned}$$

Similarly, (1.1.13) is equivalent to the requirement that

$$\text{Cov}(\varepsilon_i, \varepsilon_j | \mathbf{X}) = 0 \quad (i, j = 1, 2, \dots, n; i \neq j).$$

That is, in the joint distribution of $(\varepsilon_i, \varepsilon_j)$ conditional on \mathbf{X} , the covariance is zero. In the context of time-series models, (1.1.13) states that there is no **serial correlation** in the error term.

Since the (i, j) element of the $n \times n$ matrix $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$ is $\varepsilon_i\varepsilon_j$, Assumption 1.4 can be written compactly as

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}) = \sigma^2 \mathbf{I}_n. \quad (1.1.14)$$

The discussion of the previous paragraph shows that the assumption can also be written as

$$\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}) = \sigma^2 \mathbf{I}_n.$$

However, (1.1.14) is the preferred expression, because the more convenient measure of variability is second moments (such as $E(\varepsilon_i^2 | \mathbf{X})$) rather than variances. This point will become clearer when we deal with the large sample theory in the next chapter. Assumption 1.4 is sometimes called the **spherical** error variance assumption because the $n \times n$ matrix of second moments (which are also variances and covariances) is proportional to the identity matrix \mathbf{I}_n . This assumption will be relaxed later in this chapter.

The Classical Regression Model for Random Samples

The sample (\mathbf{y}, \mathbf{X}) is a **random sample** if $\{y_i, \mathbf{x}_i\}$ is i.i.d. (independently and identically distributed) across observations. Since by Assumption 1.1 ε_i is a function of (y_i, \mathbf{x}_i) and since (y_i, \mathbf{x}_i) is independent of (y_j, \mathbf{x}_j) for $j \neq i$, $(\varepsilon_i, \mathbf{x}_i)$ is independent of \mathbf{x}_j for $j \neq i$. So

$$\begin{aligned} E(\varepsilon_i | \mathbf{X}) &= E(\varepsilon_i | \mathbf{x}_i), \\ E(\varepsilon_i^2 | \mathbf{X}) &= E(\varepsilon_i^2 | \mathbf{x}_i), \\ \text{and } E(\varepsilon_i\varepsilon_j | \mathbf{X}) &= E(\varepsilon_i | \mathbf{x}_i) E(\varepsilon_j | \mathbf{x}_j) \quad (\text{for } i \neq j). \end{aligned} \quad (1.1.15)$$

(Proving the last equality in (1.1.15) is a review question.) Therefore, Assumptions 1.2 and 1.4 reduce to

$$\text{Assumption 1.2: } E(\varepsilon_i | \mathbf{x}_i) = 0 \quad (i = 1, 2, \dots, n), \quad (1.1.16)$$

$$\text{Assumption 1.4: } E(\varepsilon_i^2 | \mathbf{x}_i) = \sigma^2 > 0 \quad (i = 1, 2, \dots, n). \quad (1.1.17)$$

The implication of the identical distribution aspect of a random sample is that the joint distribution of $(\varepsilon_i, \mathbf{x}_i)$ does not depend on i . So the *unconditional* second moment $E(\varepsilon_i^2)$ is constant across i (this is referred to as **unconditional homoskedasticity**) and the functional form of the conditional second moment $E(\varepsilon_i^2 | \mathbf{x}_i)$ is the same across i . However, Assumption 1.4—that the *value* of the conditional

second moment is the same across i — does not follow. Therefore, Assumption 1.4 remains restrictive for the case of a random sample; without it, the conditional second moment $E(\varepsilon_i^2 \mid \mathbf{x}_i)$ can differ across i through its possible dependence on \mathbf{x}_i . To emphasize the distinction, the restrictions on the conditional second moments, (1.1.12) and (1.1.17), are referred to as **conditional homoskedasticity**.

“Fixed” Regressors

We have presented the classical linear regression model, treating the regressors as random. This is in contrast to the treatment in most textbooks, where \mathbf{X} is assumed to be “fixed” or deterministic. If \mathbf{X} is fixed, then there is no need to distinguish between the conditional distribution of the error term, $f(\varepsilon_i \mid \mathbf{x}_1, \dots, \mathbf{x}_n)$, and the unconditional distribution, $f(\varepsilon_i)$, so that Assumptions 1.2 and 1.4 can be written as

$$\text{Assumption 1.2: } E(\varepsilon_i) = 0 \quad (i = 1, \dots, n), \quad (1.1.18)$$

$$\begin{aligned} \text{Assumption 1.4: } E(\varepsilon_i^2) &= \sigma^2 \quad (i = 1, \dots, n); \\ E(\varepsilon_i \varepsilon_j) &= 0 \quad (i, j = 1, \dots, n; i \neq j). \end{aligned} \quad (1.1.19)$$

Although it is clearly inappropriate for a nonexperimental science like econometrics, the assumption of fixed regressors remains popular because the regression model with fixed \mathbf{X} can be interpreted as a set of statements conditional on \mathbf{X} , allowing us to dispense with “ $\mid \mathbf{X}$ ” from the statements such as Assumptions 1.2 and 1.4 of the model.

However, the economy in the notation comes at a price. It is very easy to miss the point that the error term is being assumed to be uncorrelated with current, past, and future regressors. Also, the distinction between the unconditional and conditional homoskedasticity gets lost if the regressors are deterministic. Throughout this book, the regressors are treated as random, and, unless otherwise noted, statements conditional on \mathbf{X} are made explicit by inserting “ $\mid \mathbf{X}$.”

QUESTIONS FOR REVIEW

1. (Change in units in the semi-log form) In the wage equation, (1.1.3), of Example 1.2, if *WAGE* is measured in cents rather than in dollars, what difference does it make to the equation? **Hint:** $\log(xy) = \log(x) + \log(y)$.
2. Prove the last equality in (1.1.15). **Hint:** $E(\varepsilon_i \varepsilon_j \mid \mathbf{X}) = E[\varepsilon_j E(\varepsilon_i \mid \mathbf{X}, \varepsilon_j) \mid \mathbf{X}]$. $(\varepsilon_i, \mathbf{x}_i)$ is independent of $(\varepsilon_j, \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n)$ for $i \neq j$.

- [**download online Innovations in Sustainable Consumption: New Economics, Socio-technical Transitions and Social Practices \(Advances in Ecological Economics series\)**](#)
- [**click Top 50 Most Delicious Parfait Recipes \(Recipe Top 50s Book 119\) online**](#)
- [*Island of Terror: Battle of Iwo Jima \(Graphic History, Volume 5\) book*](#)
- [read online Pasta by Hand: A Collection of Italy's Regional Hand-Shaped Pasta for free](#)

- <http://aircon.servicessingaporecompany.com/?lib/Innovations-in-Sustainable-Consumption--New-Economics--Socio-technical-Transitions-and-Social-Practices--Advances->
- <http://nautickim.es/books/The-Motion-Paradox--The-2-500-Year-Old-Puzzle-Behind-All-the-Mysteries-of-Time-and-Space.pdf>
- <http://sidenoter.com/?ebooks/Island-of-Terror--Battle-of-Iwo-Jima--Graphic-History--Volume-5-.pdf>
- <http://aneventshop.com/ebooks/The-First-Bad-Man.pdf>