

Introduction to Econometrics

James H. Stock

HARVARD UNIVERSITY

Mark W. Watson

PRINCETON UNIVERSITY



Boston San Francisco New York
London Toronto Sydney Tokyo Singapore Madrid
Mexico City Munich Paris Cape Town Hong Kong Montreal

Contents

Preface xxvii

PART ONE Introduction and Review 1

CHAPTER 1 Economic Questions and Data 3

1.1 Economic Questions We Examine 4

Question #1: Does Reducing Class Size Improve Elementary School Education? 4

Question #2: Is There Racial Discrimination in the Market for Home Loans? 5

Question #3: How Much Do Cigarette Taxes Reduce Smoking? 5

Question #4: What Will the Rate of Inflation Be Next Year? 6

Quantitative Questions, Quantitative Answers 7

1.2 Causal Effects and Idealized Experiments 8

Estimation of Causal Effects 8

Forecasting and Causality 9

1.3 Data: Sources and Types 10

Experimental versus Observational Data 10

Cross-Sectional Data 11

Time Series Data 11

Panel Data 13

CHAPTER 2 Review of Probability 17

2.1 Random Variables and Probability Distributions 18

Probabilities, the Sample Space, and Random Variables 18

Probability Distribution of a Discrete Random Variable 19

Probability Distribution of a Continuous Random Variable 21

2.2 Expected Values, Mean, and Variance 23

The Expected Value of a Random Variable 23

The Standard Deviation and Variance 24

Mean and Variance of a Linear Function of a Random Variable 25

Other Measures of the Shape of a Distribution 26

2.3 Two Random Variables 29

Joint and Marginal Distributions 29

Conditional Distributions 30

	Independence	34
	Covariance and Correlation	34
	The Mean and Variance of Sums of Random Variables	35
2.4	The Normal, Chi-Squared, Student t, and F Distributions	39
	The Normal Distribution	39
	The Chi-Squared Distribution	43
	The Student t Distribution	44
	The F Distribution	44
2.5	Random Sampling and the Distribution of the Sample Average	45
	Random Sampling	45
	The Sampling Distribution of the Sample Average	46
2.6	Large-Sample Approximations to Sampling Distributions	48
	The Law of Large Numbers and Consistency	49
	The Central Limit Theorem	52
	APPENDIX 2.1 Derivation of Results in Key Concept 2.3	63
CHAPTER 3	Review of Statistics	65
3.1	Estimation of the Population Mean	66
	Estimators and Their Properties	67
	Properties of \bar{Y}	68
	The Importance of Random Sampling	70
3.2	Hypothesis Tests Concerning the Population Mean	71
	Null and Alternative Hypotheses	72
	The p -Value	72
	Calculating the p -Value When σ_Y Is Known	74
	The Sample Variance, Sample Standard Deviation, and Standard Error	75
	Calculating the p -Value When σ_Y Is Unknown	76
	The t -Statistic	77
	Hypothesis Testing with a Prespecified Significance Level	78
	One-Sided Alternatives	80
3.3	Confidence Intervals for the Population Mean	81
3.4	Comparing Means from Different Populations	83
	Hypothesis Tests for the Difference Between Two Means	83
	Confidence Intervals for the Difference Between Two Population Means	84
3.5	Differences-of-Means Estimation of Causal Effects Using Experimental Data	85
	The Causal Effect as a Difference of Conditional Expectations	85
	Estimation of the Causal Effect Using Differences of Means	87

- 3.6 Using the t -Statistic When the Sample Size Is Small 88**
 The t -Statistic and the Student t Distribution 88
 Use of the Student t Distribution in Practice 92
- 3.7 Scatterplot, the Sample Covariance, and the Sample Correlation 92**
 Scatterplots 93
 Sample Covariance and Correlation 94
- APPENDIX 3.1 The U.S. Current Population Survey 105
- APPENDIX 3.2 Two Proofs That \bar{Y} Is the Least Squares Estimator of μ_Y 106
- APPENDIX 3.3 A Proof That the Sample Variance Is Consistent 107

PART TWO Fundamentals of Regression Analysis 109

- CHAPTER 4 Linear Regression with One Regressor 111
- 4.1 The Linear Regression Model 112**
- 4.2 Estimating the Coefficients of the Linear Regression Model 116**
 The Ordinary Least Squares Estimator 118
 OLS Estimates of the Relationship Between Test Scores and the Student–Teacher Ratio 120
 Why Use the OLS Estimator? 121
- 4.3 Measures of Fit 123**
 The R^2 123
 The Standard Error of the Regression 124
 Application to the Test Score Data 125
- 4.4 The Least Squares Assumptions 126**
 Assumption #1: The Conditional Distribution of u_i , Given X_i , Has a Mean of Zero 126
 Assumption #2: (X_i, Y_i) , $i = 1, \dots, n$ Are Independently and Identically Distributed 128
 Assumption #3: Large Outliers Are Unlikely 129
 Use of the Least Squares Assumptions 130
- 4.5 The Sampling Distribution of the OLS Estimators 131**
 The Sampling Distribution of the OLS Estimators 132
- 4.6 Conclusion 135**
- APPENDIX 4.1 The California Test Score Data Set 143
- APPENDIX 4.2 Derivation of the OLS Estimators 143
- APPENDIX 4.3 Sampling Distribution of the OLS Estimator 144

- CHAPTER 5 Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals 148**
- 5.1 Testing Hypotheses About One of the Regression Coefficients 149**
 - Two-Sided Hypotheses Concerning β_1 149
 - One-Sided Hypotheses Concerning β_1 153
 - Testing Hypotheses About the Intercept β_0 155
 - 5.2 Confidence Intervals for a Regression Coefficient 155**
 - 5.3 Regression When X Is a Binary Variable 158**
 - Interpretation of the Regression Coefficients 158
 - 5.4 Heteroskedasticity and Homoskedasticity 160**
 - What Are Heteroskedasticity and Homoskedasticity? 160
 - Mathematical Implications of Homoskedasticity 163
 - What Does This Mean in Practice? 164
 - 5.5 The Theoretical Foundations of Ordinary Least Squares 166**
 - Linear Conditionally Unbiased Estimators and the Gauss-Markov Theorem 167
 - Regression Estimators Other Than OLS 168
 - 5.6 Using the t -Statistic in Regression When the Sample Size Is Small 169**
 - The t -Statistic and the Student t Distribution 170
 - Use of the Student t Distribution in Practice 170
 - 5.7 Conclusion 171**
 - APPENDIX 5.1 Formulas for OLS Standard Errors 180
 - APPENDIX 5.2 The Gauss-Markov Conditions and a Proof of the Gauss-Markov Theorem 182
- CHAPTER 6 Linear Regression with Multiple Regressors 186**
- 6.1 Omitted Variable Bias 186**
 - Definition of Omitted Variable Bias 187
 - A Formula for Omitted Variable Bias 189
 - Addressing Omitted Variable Bias by Dividing the Data into Groups 191
 - 6.2 The Multiple Regression Model 193**
 - The Population Regression Line 193
 - The Population Multiple Regression Model 194
 - 6.3 The OLS Estimator in Multiple Regression 196**
 - The OLS Estimator 197
 - Application to Test Scores and the Student-Teacher Ratio 198

- 6.4 Measures of Fit in Multiple Regression 200**
 The Standard Error of the Regression (*SER*) 200
 The R^2 200
 The “Adjusted R^2 ” 201
 Application to Test Scores 202
- 6.5 The Least Squares Assumptions in Multiple Regression 202**
 Assumption #1: The Conditional Distribution of u_i Given $X_{1i}, X_{2i}, \dots, X_{ki}$ Has a Mean of Zero 203
 Assumption #2: $(X_{1i}, X_{2i}, \dots, X_{ki}, Y_i) i = 1, \dots, n$ Are i.i.d. 203
 Assumption #3: Large Outliers Are Unlikely 203
 Assumption #4: No Perfect Multicollinearity 203
- 6.6 The Distribution of the OLS Estimators in Multiple Regression 205**
- 6.7 Multicollinearity 206**
 Examples of Perfect Multicollinearity 206
 Imperfect Multicollinearity 209
- 6.8 Conclusion 210**
 APPENDIX 6.1 Derivation of Equation (6.1) 218
 APPENDIX 6.2 Distribution of the OLS Estimators When There Are Two Regressors and Homoskedastic Errors 218
- CHAPTER 7 Hypothesis Tests and Confidence Intervals in Multiple Regression 220**
- 7.1 Hypothesis Tests and Confidence Intervals for a Single Coefficient 221**
 Standard Errors for the OLS Estimators 221
 Hypothesis Tests for a Single Coefficient 221
 Confidence Intervals for a Single Coefficient 223
 Application to Test Scores and the Student–Teacher Ratio 223
- 7.2 Tests of Joint Hypotheses 225**
 Testing Hypotheses on Two or More Coefficients 225
 The F -Statistic 227
 Application to Test Scores and the Student–Teacher Ratio 229
 The Homoskedasticity-Only F -Statistic 230
- 7.3 Testing Single Restrictions Involving Multiple Coefficients 232**
- 7.4 Confidence Sets for Multiple Coefficients 234**

7.5 Model Specification for Multiple Regression 235

Omitted Variable Bias in Multiple Regression 236

Model Specification in Theory and in Practice 236

Interpreting the R^2 and the Adjusted R^2 in Practice 237**7.6 Analysis of the Test Score Data Set 239****7.7 Conclusion 244**

APPENDIX 7.1 The Bonferroni Test of a Joint Hypotheses 251

CHAPTER 8 Nonlinear Regression Functions 254**8.1 A General Strategy for Modeling Nonlinear Regression Functions 256**

Test Scores and District Income 256

The Effect on Y of a Change in X in Nonlinear Specifications 260

A General Approach to Modeling Nonlinearities Using Multiple Regression 264

8.2 Nonlinear Functions of a Single Independent Variable 264

Polynomials 265

Logarithms 267

Polynomial and Logarithmic Models of Test Scores and District Income 275

8.3 Interactions Between Independent Variables 277

Interactions Between Two Binary Variables 277

Interactions Between a Continuous and a Binary Variable 280

Interactions Between Two Continuous Variables 286

8.4 Nonlinear Effects on Test Scores of the Student–Teacher Ratio 290

Discussion of Regression Results 291

Summary of Findings 295

8.5 Conclusion 296

APPENDIX 8.1 Regression Functions That Are Nonlinear in the Parameters 307

CHAPTER 9 Assessing Studies Based on Multiple Regression 312**9.1 Internal and External Validity 313**

Threats to Internal Validity 313

Threats to External Validity 314

9.2 Threats to Internal Validity of Multiple Regression Analysis 316

Omitted Variable Bias 316

Misspecification of the Functional Form of the Regression Function 319

Errors-in-Variables 319

Sample Selection 322

Simultaneous Causality 324
Sources of Inconsistency of OLS Standard Errors 325

9.3 Internal and External Validity When the Regression Is Used for Forecasting 327

Using Regression Models for Forecasting 327
Assessing the Validity of Regression Models for Forecasting 328

9.4 Example: Test Scores and Class Size 329

External Validity 329
Internal Validity 336
Discussion and Implications 337

9.5 Conclusion 338

APPENDIX 9.1 The Massachusetts Elementary School Testing Data 344

PART THREE Further Topics in Regression Analysis 347

CHAPTER 10 Regression with Panel Data 349

10.1 Panel Data 350

Example: Traffic Deaths and Alcohol Taxes 351

10.2 Panel Data with Two Time Periods: “Before and After” Comparisons 353

10.3 Fixed Effects Regression 356

The Fixed Effects Regression Model 356
Estimation and Inference 359
Application to Traffic Deaths 360

10.4 Regression with Time Fixed Effects 361

Time Effects Only 361
Both Entity and Time Fixed Effects 362

10.5 The Fixed Effects Regression Assumptions and Standard Errors for Fixed Effects Regression 364

The Fixed Effects Regression Assumptions 364
Standard Errors for Fixed Effects Regression 366

10.6 Drunk Driving Laws and Traffic Deaths 367

10.7 Conclusion 371

APPENDIX 10.1 The State Traffic Fatality Data Set 378

APPENDIX 10.2 Standard Errors for Fixed Effects Regression with Serially Correlated Errors 379

CHAPTER 11	Regression with a Binary Dependent Variable	383
11.1	Binary Dependent Variables and the Linear Probability Model	384
	Binary Dependent Variables	385
	The Linear Probability Model	387
11.2	Probit and Logit Regression	389
	Probit Regression	389
	Logit Regression	394
	Comparing the Linear Probability, Probit, and Logit Models	396
11.3	Estimation and Inference in the Logit and Probit Models	396
	Nonlinear Least Squares Estimation	397
	Maximum Likelihood Estimation	398
	Measures of Fit	399
11.4	Application to the Boston HMDA Data	400
11.5	Summary	407
	APPENDIX 11.1 The Boston HMDA Data Set	415
	APPENDIX 11.2 Maximum Likelihood Estimation	415
	APPENDIX 11.3 Other Limited Dependent Variable Models	418
CHAPTER 12	Instrumental Variables Regression	421
12.1	The IV Estimator with a Single Regressor and a Single Instrument	422
	The IV Model and Assumptions	422
	The Two Stage Least Squares Estimator	423
	Why Does IV Regression Work?	424
	The Sampling Distribution of the TSLS Estimator	428
	Application to the Demand for Cigarettes	430
12.2	The General IV Regression Model	432
	TSLS in the General IV Model	433
	Instrument Relevance and Exogeneity in the General IV Model	434
	The IV Regression Assumptions and Sampling Distribution of the TSLS Estimator	434
	Inference Using the TSLS Estimator	437
	Application to the Demand for Cigarettes	437
12.3	Checking Instrument Validity	439
	Assumption #1: Instrument Relevance	439
	Assumption #2: Instrument Exogeneity	443
12.4	Application to the Demand for Cigarettes	445

12.5	Where Do Valid Instruments Come From?	450
	Three Examples	451
12.6	Conclusion	455
	APPENDIX 12.1 The Cigarette Consumption Panel Data Set	462
	APPENDIX 12.2 Derivation of the Formula for the TSLS Estimator in Equation (12.4)	462
	APPENDIX 12.3 Large-Sample Distribution of the TSLS Estimator	463
	APPENDIX 12.4 Large-Sample Distribution of the TSLS Estimator When the Instrument Is Not Valid	464
	APPENDIX 12.5 Instrumental Variables Analysis with Weak Instruments	466
CHAPTER 13	Experiments and Quasi-Experiments	468
13.1	Idealized Experiments and Causal Effects	470
	Ideal Randomized Controlled Experiments	470
	The Differences Estimator	471
13.2	Potential Problems with Experiments in Practice	472
	Threats to Internal Validity	472
	Threats to External Validity	475
13.3	Regression Estimators of Causal Effects Using Experimental Data	477
	The Differences Estimator with Additional Regressors	477
	The Differences-in-Differences Estimator	480
	Estimation of Causal Effects for Different Groups	484
	Estimation When There Is Partial Compliance	484
	Testing for Randomization	485
13.4	Experimental Estimates of the Effect of Class Size Reductions	486
	Experimental Design	486
	Analysis of the STAR Data	487
	Comparison of the Observational and Experimental Estimates of Class Size Effects	492
13.5	Quasi-Experiments	494
	Examples	495
	Econometric Methods for Analyzing Quasi-Experiments	497
13.6	Potential Problems with Quasi-Experiments	500
	Threats to Internal Validity	500
	Threats to External Validity	502

**13.7 Experimental and Quasi-Experimental Estimates
in Heterogeneous Populations 502**

Population Heterogeneity: Whose Causal Effect? 502

OLS with Heterogeneous Causal Effects 503

IV Regression with Heterogeneous Causal Effects 504

13.8 Conclusion 507

APPENDIX 13.1 The Project STAR Data Set 516

APPENDIX 13.2 Extension of the Differences-in-Differences Estimator to
Multiple Time Periods 517

APPENDIX 13.3 Conditional Mean Independence 518

APPENDIX 13.4 IV Estimation When the Causal Effect Varies Across
Individuals 520

PART FOUR Regression Analysis of Economic Time Series Data 523

CHAPTER 14 Introduction to Time Series Regression and Forecasting 525

14.1 Using Regression Models for Forecasting 527

14.2 Introduction to Time Series Data and Serial Correlation 528

The Rates of Inflation and Unemployment in the United States 528

Lags, First Differences, Logarithms, and Growth Rates 528

Autocorrelation 532

Other Examples of Economic Time Series 533

14.3 Autoregressions 535

The First Order Autoregressive Model 535

The p^{th} Order Autoregressive Model 538

**14.4 Time Series Regression with Additional Predictors and the
Autoregressive Distributed Lag Model 541**

Forecasting Changes in the Inflation Rate Using Past
Unemployment Rates 541

Stationarity 544

Time Series Regression with Multiple Predictors 545

Forecast Uncertainty and Forecast Intervals 548

14.5 Lag Length Selection Using Information Criteria 549

Determining the Order of an Autoregression 551

Lag Length Selection in Time Series Regression with Multiple Predictors 553

14.6 Nonstationarity I: Trends 554

What Is a Trend? 555

Problems Caused by Stochastic Trends 557

Detecting Stochastic Trends: Testing for a Unit AR Root	560
Avoiding the Problems Caused by Stochastic Trends	564
14.7 Nonstationarity II: Breaks	565
What Is a Break?	565
Testing for Breaks	566
Pseudo Out-of-Sample Forecasting	571
Avoiding the Problems Caused by Breaks	576
14.8 Conclusion	577
APPENDIX 14.1 Time Series Data Used in Chapter 14	586
APPENDIX 14.2 Stationarity in the AR(1) Model	586
APPENDIX 14.3 Lag Operator Notation	588
APPENDIX 14.4 ARMA Models	589
APPENDIX 14.5 Consistency of the BIC Lag Length Estimator	589
CHAPTER 15 Estimation of Dynamic Causal Effects	591
15.1 An Initial Taste of the Orange Juice Data	593
15.2 Dynamic Causal Effects	595
Causal Effects and Time Series Data	596
Two Types of Exogeneity	598
15.3 Estimation of Dynamic Causal Effects with Exogenous Regressors	600
The Distributed Lag Model Assumptions	601
Autocorrelated u_t , Standard Errors, and Inference	601
Dynamic Multipliers and Cumulative Dynamic Multipliers	602
15.4 Heteroskedasticity- and Autocorrelation-Consistent Standard Errors	604
Distribution of the OLS Estimator with Autocorrelated Errors	604
HAC Standard Errors	606
15.5 Estimation of Dynamic Causal Effects with Strictly Exogenous Regressors	608
The Distributed Lag Model with AR(1) Errors	609
OLS Estimation of the ADL Model	612
GLS Estimation	613
The Distributed Lag Model with Additional Lags and AR(p) Errors	615
15.6 Orange Juice Prices and Cold Weather	618
15.7 Is Exogeneity Plausible? Some Examples	624
U.S. Income and Australian Exports	625
Oil Prices and Inflation	626

Monetary Policy and Inflation 626

The Phillips Curve 627

15.8 Conclusion 627

APPENDIX 15.1 The Orange Juice Data Set 634

APPENDIX 15.2 The ADL Model and Generalized Least Squares
in Lag Operator Notation 634

CHAPTER 16 Additional Topics in Time Series Regression 637

16.1 Vector Autoregressions 638

The VAR Model 638

A VAR Model of the Rates of Inflation and Unemployment 641

16.2 Multiperiod Forecasts 642

Iterated Multiperiod Forecasts 643

Direct Multiperiod Forecasts 645

Which Method Should You Use? 647

16.3 Orders of Integration and the DF-GLS Unit Root Test 648

Other Models of Trends and Orders of Integration 648

The DF-GLS Test for a Unit Root 650

Why Do Unit Root Tests Have Non-normal Distributions? 653

16.4 Cointegration 655

Cointegration and Error Correction 655

How Can You Tell Whether Two Variables Are Cointegrated? 658

Estimation of Cointegrating Coefficients 660

Extension to Multiple Cointegrated Variables 661

Application to Interest Rates 662

16.5 Volatility Clustering and Autoregressive Conditional Heteroskedasticity 664

Volatility Clustering 665

Autoregressive Conditional Heteroskedasticity 666

Application to Stock Price Volatility 667

16.6 Conclusion 669

APPENDIX 16.1 U.S. Financial Data Used in Chapter 16 674

PART FIVE The Econometric Theory of Regression Analysis 675

CHAPTER 17 The Theory of Linear Regression with One Regressor 677

17.1 The Extended Least Squares Assumptions and the OLS Estimator 678

The Extended Least Squares Assumptions 678

The OLS Estimator 680

17.2	Fundamentals of Asymptotic Distribution Theory	680
	Convergence in Probability and the Law of Large Numbers	681
	The Central Limit Theorem and Convergence in Distribution	683
	Slutsky's Theorem and the Continuous Mapping Theorem	685
	Application to the t -Statistic Based on the Sample Mean	685
17.3	Asymptotic Distribution of the OLS Estimator and t-Statistic	686
	Consistency and Asymptotic Normality of the OLS Estimators	686
	Consistency of Heteroskedasticity-Robust Standard Errors	686
	Asymptotic Normality of the Heteroskedasticity-Robust t -Statistic	688
17.4	Exact Sampling Distributions When the Errors Are Normally Distributed	688
	Distribution of $\hat{\beta}_1$ with Normal Errors	688
	Distribution of the Homoskedasticity-only t -Statistic	690
17.5	Weighted Least Squares	691
	WLS with Known Heteroskedasticity	691
	WLS with Heteroskedasticity of Known Functional Form	692
	Heteroskedasticity-Robust Standard Errors or WLS?	695
	APPENDIX 17.1 The Normal and Related Distributions and Moments of Continuous Random Variables	700
	APPENDIX 17.2 Two Inequalities	702
CHAPTER 18	The Theory of Multiple Regression	704
18.1	The Linear Multiple Regression Model and OLS Estimator in Matrix Form	706
	The Multiple Regression Model in Matrix Notation	706
	The Extended Least Squares Assumptions	707
	The OLS Estimator	708
18.2	Asymptotic Distribution of the OLS Estimator and t-Statistic	710
	The Multivariate Central Limit Theorem	710
	Asymptotic Normality of $\hat{\beta}$	710
	Heteroskedasticity-Robust Standard Errors	711
	Confidence Intervals for Predicted Effects	712
	Asymptotic Distribution of the t -Statistic	713
18.3	Tests of Joint Hypotheses	713
	Joint Hypotheses in Matrix Notation	713
	Asymptotic Distribution of the F -Statistic	714
	Confidence Sets for Multiple Coefficients	714
18.4	Distribution of Regression Statistics with Normal Errors	715
	Matrix Representations of OLS Regression Statistics	715
	Distribution of $\hat{\beta}$ with Normal Errors	716

	Distribution of s_u^2	717
	Homoskedasticity-Only Standard Errors	717
	Distribution of the t -Statistic	718
	Distribution of the F -Statistic	718
18.5	Efficiency of the OLS Estimator with Homoskedastic Errors	719
	The Gauss-Markov Conditions for Multiple Regression	719
	Linear Conditionally Unbiased Estimators	719
	The Gauss-Markov Theorem for Multiple Regression	720
18.6	Generalized Least Squares	721
	The GLS Assumptions	722
	GLS When Ω Is Known	724
	GLS When Ω Contains Unknown Parameters	725
	The Zero Conditional Mean Assumption and GLS	725
18.7	Instrumental Variables and Generalized Method of Moments Estimation	727
	The IV Estimator in Matrix Form	728
	Asymptotic Distribution of the TSLS Estimator	729
	Properties of TSLS When the Errors Are Homoskedastic	730
	Generalized Method of Moments Estimation in Linear Models	733
	APPENDIX 18.1 Summary of Matrix Algebra	743
	APPENDIX 18.2 Multivariate Distributions	747
	APPENDIX 18.3 Derivation of the Asymptotic Distribution of $\hat{\beta}$	748
	APPENDIX 18.4 Derivations of Exact Distributions of OLS Test Statistics with Normal Errors	749
	APPENDIX 18.5 Proof of the Gauss-Markov Theorem for Multiple Regression	751
	APPENDIX 18.6 Proof of Selected Results for IV and GMM Estimation	752
	<i>Appendix</i>	755
	<i>References</i>	763
	<i>Answers to "Review the Concepts" Questions</i>	767
	<i>Glossary</i>	775
	<i>Index</i>	783