Preface

This is a sample lecture notes for undergraduate econometrics. The course aims to help students to establish a solid background in both theoretical and empirical econometrics studies. The lecture notes contain comprehensive information that is sufficient for a two semester course or it can be condensed to fit a one semester course, depending on the course design. The lecture notes cover various topics that start at a moderate level of difficulty, which gradually increases as the course proceeds.

The lecture notes were developed during my time of completing my PhD in the Queen’s University economics department. The lecture notes were created based on the following textbooks:

- “Principles of Econometrics” (3rd Edition) by R. Carter Hill, William Griffiths and Guay C. Lim
- “Introduction to Econometrics” (3rd Edition) by James H. Stock and Mark W. Watson
- “Econometric Theory and Methods” by Russell Davidson and James G. MacKinnon

I used these lecture notes when I was working as a lecturer/TA/private tutor for both undergraduate and graduate students. These lecture notes have received positive feedback and great reviews from my students, which is demonstrated in the enclosed “2012 Winter Term Evaluation Form”.

# Contents

1 Introduction 1

2 The Classical Multiple Linear Regression Model 2
   2.1 The Linear Regression Model 2
   2.2 Assumptions of the Classical Linear Regression Model 2
      2.2.1 Linearity 3
      2.2.2 Full rank 3
      2.2.3 Exogeneity of the independent variables 3
      2.2.4 Homoscedasticity and nonautocorrelation 3
      2.2.5 Exogenously generated data 4
      2.2.6 Normal distribution 4

3 Least Squares 5
   3.1 Least Squares Regression 5
   3.2 Projection 6
   3.3 Partitioned Regression and Partial Regression 6
   3.4 Partial Regression and Partial Correlation Coefficients 7
   3.5 Goodness of Fit and the Analysis of Variance 7
      3.5.1 Adjusted $R^2$-Squared and a Measure of Fit 8
      3.5.2 $R^2$-Squared and the Constant Term in the Model 8
      3.5.3 Comparing Models 8

4 Finite-Sample Properties of the Least Squares Estimator 9
   4.1 Introduction 9
   4.2 Motivating Least Squares 9
   4.3 Unbiased Estimation 10
   4.4 The Gauss Markov Theorem 10
   4.5 The Implications of Stochastic Regressors 10
   4.6 Estimating the Variance of the Least Squares Estimator 11
   4.7 The Normality Assumption and Basic Statistical Inference 11
      4.7.1 Confidence Intervals for Parameters 12
      4.7.2 The Oaxaca Decomposition 12
      4.7.3 Testing the Significance of the Regression 13
   4.8 Finite-Sample Properties of Least Squares 13
   4.9 Data Problem 13
11.4.2 The Goldfeld-Quandt Test ........................................... 62
11.4.3 The Breusch-Pagan/Godfrey LM Test ............................ 62
11.5 Weighted Least Squares When Ω is Known .......................... 62
11.6 Estimation When Ω Contains Unknown Parameters .............. 63
  11.6.1 Two-Step Estimation ............................................. 63
  11.6.2 Maximum Likelihood Estimation ................................. 63
  11.6.3 Model Based Tests for Heteroscedasticity ...................... 64
11.7 Applications ........................................................... 65
  11.7.1 Multiplicative Heteroscedasticity ............................... 65
  11.7.2 Groupwise Heteroscedasticity .................................. 66
11.8 Autoregressive Conditional Heteroscedasticity .................... 67
  11.8.1 The ARCH(1) Model .............................................. 67
  11.8.2 ARCH(q), ARCH-In-Mean and Generalized ARCH Models ... 67
  11.8.3 Maximum Likelihood Estimation of the GARCH Model ........ 68
  11.8.4 Testing for GARCH Effects ..................................... 69

12 Serial Correlation ......................................................... 70
  12.1 Introduction .......................................................... 70
  12.2 The Analysis of Time-Series Data ................................ 70
  12.3 Disturbance Processes .............................................. 70
    12.3.1 Characteristics of Disturbance Processes .................... 70
    12.3.2 AR(1) Disturbances .......................................... 71
  12.4 Some Asymptotic Results for Analyzing Time Series Data ....... 72
    12.4.1 Convergence of Moments – The Ergodic Theorem .......... 72
    12.4.2 Convergence To Normality–A Central Limit Theorem ....... 74
  12.5 Least Squares Estimation .......................................... 75
    12.5.1 Asymptotic Properties of Least Squares .................... 75
    12.5.2 Estimating the Variance of the Least Squares Estimator .. 75
  12.6 GMM Estimation ..................................................... 76
  12.7 Testing for Autocorrelation ....................................... 76
    12.7.1 Lagrange Multiplier Test .................................... 77
    12.7.2 Box and Pierce’s Test and Ljung’s Refinement ............. 77
    12.7.3 The Durbin-Watson Test ...................................... 77
    12.7.4 Testing in the Presence of A Lagged Dependent Variables .. 78
  12.8 Efficient Estimation When Ω is Known ............................ 78
  12.9 Estimation When Ω Is Unknown .................................... 79
    12.9.1 AR(1) Disturbances .......................................... 79
    12.9.2 AR(2) Disturbances .......................................... 79
    12.9.3 Estimation with A Lagged Dependent Variable ................ 80
  12.10 Common Factors ................................................... 80
  12.11 Forecasting in the Presence of Autocorrelation ................ 81

13 Models for Panel Data .................................................. 82
  13.1 Panel Data Models .................................................. 82
  13.2 Fixed Effects ....................................................... 83
    13.2.1 Testing the Significance of the Group Effects .............. 84
    13.2.2 The Within- And Between-Groups Estimators ............... 84
    13.2.3 Fixed Time and Group Effects ................................ 85
15.2.3 Identification Through Covariance Restrictions–The Fully Recursive Model .......................... 125
15.3 Methods of Estimation ............................................................... 126
15.4 Single Equation: Limited Information Estimation Methods ................................. 126
15.4.1 Ordinary Least Squares ....................................................... 126
15.4.2 Estimation by Instrumental Variables .................................... 126
15.4.3 Two-Stage Least Squares ...................................................... 127
15.4.4 GMM Estimation ............................................................... 128
15.4.5 Limited Information Maximum Likelihood and the k Class of Estimators .......... 128
15.4.6 Two-Stage Least Squares in Models that Are Nonlinear in Variables .......... 129
15.5 System Methods of Estimation ................................................... 129
15.5.1 Three-Stage Least Squares ................................................... 130
15.5.2 Full-Information Maximum Likelihood ..................................... 131
15.5.3 GMM Estimation ............................................................... 132
15.5.4 Recursive Systems and Exactly identified Equations ....................... 133
15.6 Specification Tests ................................................................. 134
15.7 Properties of Dynamic Models .................................................. 134
15.7.1 Dynamic Models and Their Multipliers .................................... 134
15.7.2 Stability ................................................................. 136
15.7.3 Adjustment to Equilibrium .................................................. 136

16 Estimation Frameworks in Econometrics ................................................. 138
16.1 Parametric Estimation and Inference ............................................. 138
16.1.1 Classical Likelihood Based Estimation ...................................... 138
16.1.2 Bayesian Estimation ........................................................... 138
16.1.3 Bayesian Analysis of the Classical Regression Model ....................... 139
16.1.4 Interval Estimation ............................................................ 141
16.1.5 Estimation with an Informative Prior Density ............................... 141
16.1.6 Hypothesis Testing ............................................................ 142
16.1.7 The Latent Class Model ....................................................... 143
16.1.8 Estimate by Markov Chain Monte Carlo Simulation ....................... 144
16.2 Semiparametric Estimation .......................................................... 146
16.2.1 Least Absolute Deviations Estimation ..................................... 146
16.2.2 Partially Linear Regression ................................................... 146
16.3 Nonparametric Estimation .......................................................... 147
16.3.1 Kernel Density Estimation .................................................... 147
16.3.2 Nonparametric Regression .................................................... 148
16.4 Properties of Estimators ............................................................ 149
16.4.1 Statistical Properties of Estimators ........................................ 149
16.4.2 Extremum Estimators .......................................................... 149
16.4.3 Assumptions for Asymptotic Properties of Extremum Estimators .......... 149
16.4.4 Asymptotic Properties of Estimators ...................................... 150

17 Maximum Likelihood Estimation ...................................................... 151
17.1 The Likelihood Function and Identification of the Parameters ....................... 151
17.2 Properties of Maximum Likelihood Estimators .................................... 151
17.2.1 Regularity Conditions ......................................................... 152
17.2.2 Properties of Regular Densities .............................................. 152
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.2.3</td>
<td>The Likelihood Equation</td>
<td>153</td>
</tr>
<tr>
<td>17.2.4</td>
<td>The Information Matrix Equality</td>
<td>154</td>
</tr>
<tr>
<td>17.2.5</td>
<td>Asymptotic Properties of the Maximum Likelihood Estimator</td>
<td>154</td>
</tr>
<tr>
<td>17.2.6</td>
<td>Estimating the Asymptotic Variance of the Maximum Likelihood Estimator</td>
<td>156</td>
</tr>
<tr>
<td>17.2.7</td>
<td>Conditional Likelihoods and Econometric Models</td>
<td>157</td>
</tr>
<tr>
<td>17.3</td>
<td>Three Asymptotically Equivalent Test Procedures</td>
<td>157</td>
</tr>
<tr>
<td>17.3.1</td>
<td>The Likelihood Ratio Test</td>
<td>158</td>
</tr>
<tr>
<td>17.3.2</td>
<td>The Wald Test</td>
<td>158</td>
</tr>
<tr>
<td>17.3.3</td>
<td>The Lagrange Multiplier Test</td>
<td>159</td>
</tr>
<tr>
<td>17.4</td>
<td>Application of Maximum Likelihood Estimation</td>
<td>160</td>
</tr>
<tr>
<td>17.4.1</td>
<td>The Normal Linear Regression Model</td>
<td>160</td>
</tr>
<tr>
<td>17.4.2</td>
<td>Maximum Likelihood Estimation of Nonlinear Regression Models</td>
<td>162</td>
</tr>
<tr>
<td>17.4.3</td>
<td>Nonnormal Disturbances–The Stochastic Frontier Model</td>
<td>163</td>
</tr>
<tr>
<td>17.4.4</td>
<td>Conditional Moment Tests of Specification</td>
<td>165</td>
</tr>
<tr>
<td>17.5</td>
<td>Two-Step Maximum Likelihood Estimation</td>
<td>166</td>
</tr>
<tr>
<td>17.6</td>
<td>Maximum Simulated Likelihood Estimation</td>
<td>167</td>
</tr>
<tr>
<td>17.7</td>
<td>Pseudo-Maximum Likelihood Estimation and Robust Asymptotic Covariance</td>
<td>168</td>
</tr>
</tbody>
</table>

| 18     | The Generalized Method of Moments                                    | 171  |
| 18.1   | Consistent Estimation: The Method of Moments                         | 171  |
| 18.1.1 | Random Sampling and Estimating the Parameters of Distributions       | 171  |
| 18.1.2 | Asymptotic Properties of the Method of Moments Estimator             | 172  |
| 18.2   | The Generalized Method of Moments (GMM) Estimator                   | 173  |
| 18.2.1 | Estimation Based on Orthogonality Conditions                          | 173  |
| 18.2.2 | Generalizing the Method of Moments                                   | 175  |
| 18.2.3 | Properties of the GMM Estimator                                      | 176  |
| 18.2.4 | GMM Estimation of Some Specific Econometric Models                  | 179  |
| 18.3   | Testing Hypotheses in the GMM Framework                              | 181  |
| 18.3.1 | Testing the Validity of the Moment Restrictions                      | 181  |
| 18.3.2 | GMM Counterparts to the Wald, LM and LR Tests                        | 182  |

<p>| 19     | Models with Lagged Variables                                          | 183  |
| 19.1   | Dynamic Regression Models                                             | 183  |
| 19.1.1 | Lagged Effects in A Dynamic Model                                    | 183  |
| 19.1.2 | The Lag and Difference Operators                                     | 184  |
| 19.1.3 | Specification Search for the Lag Length                              | 185  |
| 19.2   | Simple Distributed Lag Models                                        | 186  |
| 19.2.1 | Finite Distributed Lag Models                                        | 186  |
| 19.2.2 | An Infinite Lag Model: the Geometric Lag Model                       | 186  |
| 19.3   | Autoregressive Distributed Lag Models                                | 188  |
| 19.3.1 | Estimation of the ARDL Model                                         | 188  |
| 19.3.2 | Computation of the Lag Weights in the ARDL Model                     | 188  |
| 19.3.3 | Stability of a Dynamic Equation                                      | 189  |
| 19.3.4 | Forecasting                                                           | 190  |
| 19.4   | Methodological Issues in the Analysis of Dynamic Models               | 191  |
| 19.4.1 | An Error Correction Model                                            | 191  |</p>
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.4.2</td>
<td>Autocorrelation</td>
<td>191</td>
</tr>
<tr>
<td>19.4.3</td>
<td>Common Factor Restrictions</td>
<td>192</td>
</tr>
<tr>
<td>19.5</td>
<td>Vector Autoregressions</td>
<td>193</td>
</tr>
<tr>
<td>19.5.1</td>
<td>Model Forms</td>
<td>194</td>
</tr>
<tr>
<td>19.5.2</td>
<td>Estimation</td>
<td>194</td>
</tr>
<tr>
<td>19.5.3</td>
<td>Testing Procedures</td>
<td>194</td>
</tr>
<tr>
<td>19.5.4</td>
<td>Exogeneity</td>
<td>195</td>
</tr>
<tr>
<td>19.5.5</td>
<td>Testing for Granger Causality</td>
<td>196</td>
</tr>
<tr>
<td>19.5.6</td>
<td>Impulse Response Functions</td>
<td>196</td>
</tr>
<tr>
<td>19.5.7</td>
<td>Structural VARs</td>
<td>197</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

In the first issue of *Econometrica*, the Econometric Society stated that its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences.

But there are several aspects of the quantitative approach to economics, and no single one of these aspects taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous [sic] with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is the unification that constitutes econometrics.
Chapter 2

The Classical Multiple Linear Regression Model

2.1 The Linear Regression Model

Generic form of the linear regression model is

\[ y = f(x_1, x_2, ..., x_k) + \epsilon = x_1\beta_1 + x_2\beta_2 + ... + x_k\beta_k + \epsilon \]

where

- \( y \) − regressand
- \( x_k \) − regressor or covariates
- \( \epsilon \) − disturbance or error term

The most useful aspects of the multiple regression model is its ability to identify the independent effects of a set of variables on a dependent variable.

2.2 Assumptions of the Classical Linear Regression Model

List of Assumptions

A1. Linearity
A2. Full rank
A3. Exogeneity of the independent variables
A4. Homoscedasticity and nonautocorrelation
A5. Exogenously generated data
A6. Normal distribution
2.2.1 Linearity

\[ y = X\beta + \epsilon \]

The model specifies a linear relationship between \( y \) and \( X \), in matrix form. In practice, the variety of linear functions is unlimited. For example, the **loglinear model** is

\[ \ln y = \beta_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 + \ldots + \beta_k \ln x_k + \epsilon. \]

This equation is also known as the **constant elasticity** form. A **semilog** model is often used to model growth rates

\[ \ln y_t = x'_t \beta + \delta t + \epsilon_t \]

Other variations of the general form

\[ f(y_t) = g(x'_t \beta + \epsilon_t) \]

will allow a tremendous variety of functional forms, all of which fit into our definition of a linear model.

2.2.2 Full rank

\( X \) is an \( n \times k \) matrix with rank \( k \).

Hence, \( X \) has full column rank; the columns of \( X \) are linearly independent and there are at least \( k \) observations. This assumption is known as an **identification condition**.

2.2.3 Exogeneity of the independent variables

The disturbance is assumed to have conditional expected value zero at every observation, which we write as

\[ E(\epsilon_i|X) = 0 \]

In principle, the mean of each \( \epsilon_i \) *conditioned on all observations* \( x_i \) is zero. In words, that no observations on \( x \) convey information about the expected value of the disturbance.

The zero conditional mean implies that the unconditional mean is also zero, since

\[ E(\epsilon_i) = E_X[E(\epsilon_i|X)] = E_X(0) = 0 \]

This also implies that

\[ E(y|X) = X\beta \]

2.2.4 Homoscedasticity and nonautocorrelation

This assumption concerns the variances and covariances of the disturbances

\[ \text{Var}[\epsilon_i|X] = \sigma^2, \quad \text{for all} \quad i = 1, \ldots, n \]

and

\[ \text{Cov}[\epsilon_i, \epsilon_j|X] = 0, \quad \text{for all} \quad i \neq j. \]

Constant variance is labeled **homoscedasticity**, while uncorrelatedness across observations is labeled generically **nonautocorrelation**. Note, this assumption can be summarize as

\[ E[\epsilon'|X] = \sigma^2 I \]

Disturbances that meet this twin assumptions are sometimes called **spherical** disturbances.
2.2.5 Exogenously generated data

The ultimate source of data in $X$ is unrelated (statistically and economically) to the source of $\epsilon$.

2.2.6 Normal distribution

It is convenient to assume that the disturbances are normally distributed, with zero mean and constant variance.

$$\epsilon | X \sim N[0, \sigma^2 I].$$

In view of our description of the source of $\epsilon$, the condition of the central limit theorem will generally apply, at least approximately, and the normality assumption will be reasonable in most settings. Note, normality is not required, sometimes even not necessary, to obtain many of the results we use in multiple regression analysis, although it does prove useful in constructing test statistics.
Chapter 3

Least Squares

3.1 Least Squares Regression

For model
\[ y = X\beta + \epsilon \]

The disturbance associated with the \(i\)th data point is
\[ \epsilon_i = y_i - x'_i\beta \]

For any value of \(b\), we shall estimate \(\epsilon_i\) with the residual
\[ e_i = y_i - x'_ib. \]

The least squares coefficient vector minimizes the sum of squared residuals
\[ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - x'_ib_0)^2 \]

where \(b_0\) denotes the choice for the coefficient vector. In matrix terms, minimizing the sum of squares requires us to choose \(b_0\) to

Minimize_{\(b_0\)} \(S(b_0) = e'_0e_0 = (y - Xb_0)'(y - Xb_0)\)
\[ = y'y' - b'_0X'y - y'Xb_0 + b'_0X'Xb_0 \]
\[ = y'y' - 2y'Xb_0 + b'_0X'Xb_0 \]

The necessary condition for a minimum is
\[ \frac{\partial S(b_0)}{\partial b_0} = -2X'y + 2X'Xb_0 = 0 \]

Let \(b\) be the solution. Then \(b\) satisfies the least squares normal equations,
\[ X'Xb = X'y \]

Therefore, if the inverse of \(X'X\) exists, the solution is
\[ b = (X'X)^{-1}X'y \]
3.2 Projection

The vector of least squares residuals is

\[ e = y - Xb \]

Inserting \( b \) gives

\[ e = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = (I - P)y = My \]

The \( n \times n \) matrix \( P \) and \( M \), which are symmetric \((P = P', M = M')\) and idempotent \((P = P^2, M = M^2)\), are called projection matrices. The word projection is coming from 

\[ \hat{y} = y - e = (I - M)y = X(X'X)^{-1}X'y = Py \]

where the fitted value, \( \hat{y} \), is the projection of the vector \( y \) into the column space of \( X \). 

\( M \) and \( P \) are orthogonal

\[ PM = MP = 0 \]

Also,

\[ PX = X \]

We can see that least squares partitions the vector \( y \) into two orthogonal parts

\[ y = Py + My = \text{projection} + \text{residual} \]

3.3 Partitioned Regression and Partial Regression

Suppose that the regression involves two sets of variables \( X_1 \) and \( X_2 \). Thus

\[ y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon \]

To estimate \( b_2 \), we adopt following theorem

**Frisch-Waugh Theorem**

In the linear least squares regression of vector \( y \) on two sets of variables, \( X_1 \) and \( X_2 \), the subvector \( b_2 \) is the set of coefficients obtained when the residuals from a regression of \( y \) on \( X_1 \) alone and regressed on the set of residuals obtained when each column of \( X_2 \) is regressed on \( X_1 \).

Therefore, following result is obtained

\[ b_2 = (X'_2M_1X_2)^{-1}(X'_2M_1y) \]

This process is commonly called partialing out or netting out the effect of \( X_1 \). For this reason, the coefficients in a multiple regression are often called the partial regression coefficients. Based on Frisch-Waugh Theorem, we find the following.

**Regression with a Constant Term**

The slopes in a multiple regression that contains a constant term are obtained by transforming the data to deviations from their means and then regressing the variable \( y \) in deviation form on the explanatory variables, also in deviation form.
3.4 Partial Regression and Partial Correlation Coefficients

Consider the following situation

\[ y = Xb + e \]
\[ y =Xd + zc + u \]

A new variable \( z \) is added to the original regression, where \( c \) is a scalar. We have

\[ d = (X'X)^{-1}X'(y - zc) = b - (X'X)^{-1}X'zc \]

Therefore,

\[ u = y - Xd - zc \]
\[ = y - X[b - (X'X)^{-1}X'zc] - zc \]
\[ = (y - Xb) + X(X'X)^{-1}X'zc - zc \]
\[ = e - Mzc \]

Now,

\[ u'u = e'e + c^2(z'Mz) - 2cz'Me \]
\[ = e'e + c^2(z'Mz) - 2cz'M(My) \]
\[ = e'e + c^2(z'Mz) - 2cz'M(Xd + zc + u) \]
\[ = e'e + c^2(z'Mz) - 2cz'Mzc \]
\[ = e'e - c^2(z'Mz) \leq e'e \]

Now, we obtain useful result

**Change in the Sum of Squares When a Variable Is Added to a Regression**

If \( e'e \) is the sum of squared residuals when \( y \) is regressed on \( X \) and \( u'u \) is the sum of squared residuals when \( y \) is regressed on \( X \) and \( z \), then

\[ u'u = e'e - c^2(z'Mz) \leq e'e \]

where \( c \) is the coefficient on \( z \) in the long regression and \( Mz = [I - X(X'X)^{-1}X']z \) is the vector of residuals when \( z \) is regressed on \( X \).

3.5 Goodness of Fit and the Analysis of Variance

The **total variation** in \( y \) is the sum of squared deviations:

\[ SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \]

For regression

\[ y = Xb + e \]

we obtain

\[ M^0y = M^0Xb + e \]
where $M^0$ is the $n \times n$ idempotent matrix that transforms observations into deviations from sample means. Note, $M^0e = e$. Then, the total sum of squares is

$$y'M^0y = b'X'M^0Xb + e'e$$

or

$$SST = SSR + SSE$$

We can now obtain a measure of how well the regression line fits the data by using the coefficient of determination:

$$\frac{SSR}{SST} = \frac{b'X'M^0Xb}{y'M^0y} = 1 - \frac{e'e}{y'M^0y}$$

The coefficient of determination is denoted $R^2$. One extreme, $R^2 = 1$, occurs if the values of $x$ and $y$ all lie in the same hyperplane.

### 3.5.1 Adjusted $R$-Squared and a Measure of Fit

One problem with $R^2$ is, it will never decrease when another variable is added to a regression equation. To correct this, the adjusted $R^2$ (denoted as $\bar{R}^2$) is introduced

$$\bar{R}^2 = 1 - \frac{e'e/(n - k)}{y'M^0y/(n - 1)}$$

The connection between $R^2$ and $\bar{R}^2$ is

$$R^2 = 1 - \frac{n - 1}{n - k} (1 - R^2)$$

The adjusted $R^2$ may decline when a variable is added to the set of independent variables.

### 3.5.2 $R$-Squared and the Constant Term in the Model

Another difficulty with $R^2$ concerns the constant term in the model. The proof that $0 \leq R^2 \leq 1$ requires $X$ to contain a column of 1s. If not, then $M^0e \neq e$ and $e'M^0X \neq 0$. Consequently, when we compute

$$R^2 = 1 - \frac{e'e}{y'M^0y}$$

the result is unpredictable. Some computer packages bypass these difficulties by reporting a third $R^2$, the squared sample correlation between the actual values of $y$ and the fitted values from the regression. However, this approach could be deceptive.

### 3.5.3 Comparing Models

The value of $R^2$ is not a absolute basis for comparison. In fact, larger $R^2$ doesn’t imply the model is a better fit in some absolute sense. But it is worth emphasizing that $R^2$ is a measure of linear association between $x$ and $y$ though.
Chapter 4

Finite-Sample Properties of the Least Squares Estimator

4.1 Introduction

In this chapter, we will examine in detail the least squares as an estimator of the model parameters of the classical model.

4.2 Motivating Least Squares

Consider the problem of finding an optimal linear predictor. We seek the minimum mean squared error linear predictor of $y$, which we will denote $x'\gamma$. The expected squared error of this predictor is

$$\text{MSE} = E_y E_x [y - x'\gamma]^2$$

This can be written as

$$\text{MSE} = E_{y,x} \{ y - E[y|x]\}^2 + E_{y,x} \{ E[y|x] - x'\gamma \}^2$$

We seek the $\gamma$ that minimizes this expectation. The necessary condition is

$$\frac{\partial E_y E_x \{[E(y|x) - x'\gamma]^2\}}{\partial \gamma} = E_y E_x \{ \frac{[E(y|x) - x'\gamma]^2}{\partial \gamma} \} = -2E_y E_x \{x[E(y|x) - x'\gamma]\} = 0.$$ 

Finally, we have the equivalent condition

$$E_y E_x [x E(y|x)] = E_y E_x [xx']\gamma$$

Applying Law of Iterated Expectation to the left hand side, we obtain the necessary condition for finding the minimum MSE predictor

$$E_x E_y [xy] = E_x E_y [xx']\gamma$$
4.3 Unbiased Estimation

The least squares estimator is unbiased in every sample

\[ b = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon \]

Now, take expectations, iterating over \( X \),

\[ E(b|X) = \beta + E[(X'X)^{-1}X'\epsilon|X] = \beta \]

4.4 The Variance of the Least Squares Estimator and the Gauss Markov Theorem

The covariance matrix of the least squares slope estimator is

\[ \text{Var}[b|X] = E[(b - \beta)(b - \beta)'|X] \]

\[ = E[(X'X)^{-1}X'\epsilon X(X'X)^{-1}|X] \]

\[ = (X'X)^{-1}X'E[\epsilon\epsilon'|X]X(X'X)^{-1} \]

\[ = (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \]

\[ = \sigma^2(X'X)^{-1} \]

Let \( b_0 = Cy \) be another linear unbiased estimator of \( \beta \), where \( C \) is a \( k \times n \) matrix. If \( b_0 \) is unbiased, then

\[ E[Cy|X] = E[(CX\beta + C\epsilon)|X] = \beta \]

which implies that \( CX = I \). Now let \( D = C - (X'X)^{-1}X' \) so \( Dy = b_0 - b \) and \( DX = CX - I = 0 \). Then,

\[ \text{Var}[b_0|X] = \sigma^2[(D + (X'X)^{-1}X')(D + (X'X)^{-1}X')'] \]

\[ = \sigma^2(X'X)^{-1} + \sigma^2 DD' = \text{Var}[b|X] + \sigma^2 DD' \]

Since a quadratic form in \( DD' \) is \( q'DD'q = z'z \geq 0 \), the conditional covariance matrix of \( b_0 \) equals that of \( b \) plus a nonnegative definite matrix. Hence,

**Gauss-Markov Theorem**

In the classical linear regression model with regressor matrix \( X \), the least squares estimator \( b \) is the minimum variance linear unbiased estimator of \( \beta \). For any vector of constants \( w \), the minimum variance linear unbiased estimator of \( w'\beta \) in the classical regression model is \( w'\hat{b} \), where \( \hat{b} \) is the least squares estimator.

4.5 The Implications of Stochastic Regressors

The conditional variance of \( b \) is

\[ \text{Var}[b|X] = \sigma^2(X'X)^{-1} \]

\[ = E_X[\text{Var}[b|X]] + \text{Var}_X[E[b|X]] \]

The second term is zero since \( E[b|X] = \beta \) for all \( X \), so

\[ \text{Var}[b] = E_X[\sigma^2(X'X)^{-1}] = \sigma^2 E_X[(X'X)^{-1}] \]
Gauss-Markov Theorem (Concluded)
In the classical linear regression model, the least squares estimator \(\hat{b}\) is the minimum variance linear unbiased estimator of \(\beta\) whether \(X\) is stochastic or non-stochastic, so long as the other assumptions of the model continue to hold.

4.6 Estimating the Variance of the Least Squares Estimator
The estimate of variance, by analogy
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2
\]
where \(e_i = y_i - x_i' \hat{b} = \epsilon - x_i' (b - \beta)\). This estimator is distorted because \(\beta\) is not observed directly.

The least squares residuals are
\[
e = My = M[X\beta + \epsilon] = M\epsilon
\]
An estimator of \(\sigma^2\) will be based on the sum of squared residuals
\[
e'\epsilon = \epsilon'M\epsilon
\]
The expected value of this quadratic form is
\[
E[e'e|X] = E[e'M\epsilon|X]
\]
The scalar \(e'M\epsilon\) is a \(1 \times 1\) matrix, so it is equal to its trace.
\[
E[tr(e'M\epsilon)|X] = E[tr(M\epsilon e'|X)] = tr(ME[\epsilon e'|X]) = tr(M\sigma^2 I) = \sigma^2 tr(M)
\]
Note, the trace of \(M\) is
\[
tr[I_n - X(X'X)^{-1}X'] = tr(I_n) - tr[(X'X)^{-1}X'X] = tr(I_n) - tr(I_k) = n - k
\]
Therefore,
\[
E[e'e|X] = (n - k)\sigma^2
\]
An unbiased estimator of \(\sigma^2\) is
\[
s^2 = \frac{e'e}{n - k}
\]
The standard error of the regression is \(s\), the square root of \(s^2\). With \(s^2\), we can then compute
\[
\text{Est.Var}[b|X] = s^2(X'X)^{-1}
\]

4.7 The Normality Assumption and Basic Statistical Inference
This assumption is useful for constructing statistics for testing hypotheses. Since \(b\) is a linear function of the disturbance vector \(\epsilon\)
\[
b|X \sim N[\beta, \sigma^2(X'X)^{-1}]
\]
This specifies a multivariate normal distribution, so each element of \(b|X\) is normally distributed
\[
b_k|X \sim N[\beta_k, \sigma^2(X'X)^{-1}_{kk}]
\]
Testing A Hypothesis about A Coefficient

Let $S_{kk}$ be the $k$th diagonal element of $(X'X)^{-1}$.

$$z_k = \frac{b_k - \beta_k}{\sqrt{s^2 S_{kk}}}$$

has a standard normal distribution. By using $s^2$ instead of $\sigma^2$, we can derive a statistic.

The quantity

$$\frac{(n-k)s^2}{\sigma^2} = \frac{e'e}{\sigma^2} \left( \frac{\epsilon}{\sigma} \right)' M \left( \frac{\epsilon}{\sigma} \right)$$

is an idempotent quadratic form in a standard normal vector $(\epsilon/\sigma)$. Therefore, it has a chi-squared distribution with rank(M) = trace(M) = $n - k$ degrees of freedom.

Also,

$$\frac{b - \beta}{\sigma} = (X'X)^{-1} X' \left( \frac{\epsilon}{\sigma} \right)$$

is independent of $(n-k)s^2/\sigma^2$, according to the following theorem.

**Theorem: Independence of $b$ and $s^2$**

If $\epsilon$ is normally distributed, then the least squares coefficient estimator $b$ is statistically independent of the residual vector $e$ and therefore, all functions of $e$, including $s^2$.

Therefore, the ratio

$$t_k = \frac{(b_k - \beta_k)/\sqrt{s^2 S_{kk}}}{\sqrt{[(n-k)s^2/\sigma^2]/(n-k)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S_{kk}}}$$

has a $t$ distribution with $(n-k)$ degrees of freedom. We can use $t_k$ to test hypotheses or form confidence intervals about the individual elements of $\beta$.

4.7.1 Confidence Intervals for Parameters

We could say that

$$\text{Prob}[b_k - t_{\alpha/2} s_{b_k} \leq \gamma \leq b_k + t_{\alpha/2} s_{b_k}] = 1 - \alpha$$

where $1 - \alpha$ is the desired level of confidence and $t_{\alpha/2}$ is the appropriate critical value from the $t$ distribution with $(n-k)$ degrees of freedom.

4.7.2 Confidence Interval for A Linear Combination of Coefficients: the Oaxaca Decomposition

By extending those results, we can show how to form a confidence interval for a linear function of the parameters. **Oaxaca’s decomposition** provides a frequently used application.

Let $w$ denote a $k \times 1$ vector of known constants. Then the linear combination $c = w'b$ is normally distributed with mean $\gamma = w'\beta$ and variance $\sigma^2_c = w'[\sigma^2(X'X)^{-1}]w$, which we estimate with $s^2_c = w' [s^2(X'X)^{-1}]w$. With these in hand, we can use the earlier results to form a confidence interval for $\gamma$:

$$\text{Prob}[c - t_{\alpha/2} s_{c} \leq \gamma \leq c + t_{\alpha/2} s_{c}] = 1 - \alpha$$
4.7.3 Testing the Significance of the Regression

A question that is usually of interest is whether the regression equation as a whole is significant. This test is a joint test of the hypotheses that all the coefficients except the constant term are zero. The central result needed to carry out the test is the distribution of the statistic

$$F[k - 1, n - k] = \frac{R^2/(k - 1)}{(1 - R^2)/(n - k)}$$

This statistic has an F distribution with $k - 1$ and $n - k$ degrees of freedom. If $F$ is large, then the hypothesis is rejected.

4.8 Finite-Sample Properties of Least Squares

Finite Sample Properties of Least Squares

FS1. $E[b|X] = E[b] = \beta$. Least squares is unbiased

FS2. $\text{Var}[b|X] = \sigma^2(X'X)^{-1}$; $\text{Var}[b] = \sigma^2E[(X'X)^{-1}]$

FS3. Gauss-Markov Theorem: The MVLUE\(^2\) of $w\beta$ is $w'b$

FS4. $E[s^2|X] = E[s^2] = \sigma^2$

FS5. $\text{Cov}[b,e|X] = E[(b - \beta)e'|X] = E[(X'X)^{-1}X'e'EM|X] = 0$ as $X'(\sigma^2I)M = 0$

Results that follow from Assumption A6, normally distributed disturbances

FS6. $b$ and $e$ are statistically independent. It follows that $b$ and $s^2$ are uncorrelated and statistically independent.

FS7. The exact distribution of $b|X$, is $N[\beta, \sigma^2(X'X)^{-1}]$

FS8. $(n - k)s^2/\sigma^2 \sim \chi^2[n - k]$. $s^2$ has mean $\sigma^2$ and variance $2\sigma^4/(n - k)$

FS9. $t[n - k] = (b_k - \beta_k)/[s^2(X'X)_{kk}]^{1/2} \sim t[n - k]$ independently of $k$

FS10. The test statistic for testing the null hypothesis that all slopes in the model are zero, $F[k - 1, n - k] = [R^2/(k - 1)]/[(1 - R^2)/(n - k)]$ has an F distribution with $k - 1$ and $n - k$ degrees of freedom when the null hypothesis is true.

4.9 Data Problem

Three practical problems often arise in the setting of regression analysis, multicollinearity, missing observations and outliers. Although many econometricians tried to offer solutions, but it seems there are no ultimate ones that can solve the problems above nicely. Hence, the solution is heavily depends on the unique data characteristic and model setting.

\(^1\)Note that a large $F$ is induced by a large value of $R^2$

\(^2\)Minimum Variance Linear Unbiased Estimation
Chapter 5

Large-Sample Properties of the Least Squares and Instrumental Variables Estimation

5.1 Introduction

This chapter will examine the asymptotic properties of the parameter estimators in the classical regression model.

5.2 Asymptotic Properties of the Least Squares Estimator

5.2.1 Consistency of the Least Squares Estimator of $\beta$

We make two crucial assumptions. The first is a modification of Assumption A5

A5a. $(x_i, \epsilon_i), i = 1, ..., n$ is a sequence of independent observations.

The second concerns the behavior of the data in large samples;

$$\lim_{n \to \infty} \frac{X'X}{n} = Q, \quad \text{a positive definite matrix.}$$

The least squares estimator may be written

$$b = \beta + \left( \frac{X'X}{n} \right)^{-1} \left( \frac{X'\epsilon}{n} \right)$$

If $Q^{-1}$ exists, then

$$\lim b = \beta + Q^{-1} \lim \left( \frac{X'\epsilon}{n} \right)$$

Let

$$\frac{1}{n} X'\epsilon = \frac{1}{n} \sum_{i=1}^{n} x_i \epsilon_i = \frac{1}{n} \sum_{i=1}^{n} w_i = \bar{w}$$

then

$$\lim b = \beta + Q^{-1} \lim \bar{w}$$
Note,
\[ \text{Var}[\bar{w}|X] = E[\bar{w}\bar{w}'|X] = \frac{1}{n}X'E[\varepsilon\varepsilon'|X]X \frac{1}{n} = \left( \frac{\sigma^2}{n} \right) \left( \frac{X'X}{n} \right) \]
therefore,
\[ \text{Var}[\bar{w}] = \left( \frac{\sigma^2}{n} \right) E \left( \frac{X'X}{n} \right) \]
The variance will collapse to zero if the expectation in parentheses is a constant matrix, so that the leading scalar will dominate the product as \( n \) increases. It then follows that
\[ \lim_{n \to \infty} \text{Var}[\bar{w}] = 0 \cdot Q = 0 \]
Since the mean of \( \bar{w} \) is identically zero and its variance converges to zero, \( \bar{w} \) converges in mean square to zero, so \( \text{plim}\bar{w} = 0 \).
\[ \text{plim} \frac{X'\varepsilon}{n} = 0 \]
so,
\[ \text{plimb} = \beta + Q^{-1} \cdot 0 = \beta \]

5.2.2 Asymptotic Normality of the Least Squares Estimator

Give equation,
\[ \sqrt{n}(b - \beta) = \left( \frac{X'X}{n} \right)^{-1} \left( \frac{1}{\sqrt{n}} \right) X'\varepsilon \]
we know \( \text{plim}(X'X/n)^{-1} = Q^{-1} \)
\[ \left[ \left( \frac{X'X}{n} \right)^{-1} \right] \left( \frac{1}{\sqrt{n}} \right) X'\varepsilon = Q^{-1} \left( \frac{1}{\sqrt{n}} \right) X'\varepsilon \]
We need to establish the limiting distribution of
\[ \left( \frac{1}{\sqrt{n}} \right) X'\varepsilon = \sqrt{n}(\bar{w} - E[\bar{w}]) \]
We can use the multivariate Lindberg-Feller version of the central limit theorem to obtain the limiting distribution of \( \sqrt{n}\bar{w} \), where \( \bar{w} \) is the average of \( n \) independent random vectors \( w_i = x_i\varepsilon_i \), with mean 0 and variances
\[ \text{Var}[x_i\varepsilon_i] = \sigma^2E[x_ix_i'] = \sigma^2Q_i \]
The variance of \( \sqrt{n}\bar{w} \) is
\[ \sigma^2Q_n = \sigma^2 \left( \frac{1}{n} \right) [Q_1 + Q_2 + \ldots + Q_n] \]
where
\[ \lim_{n \to \infty} \sigma^2Q_n = \sigma^2Q \]
then
\[ \left( \frac{1}{\sqrt{n}} \right) X'\varepsilon \xrightarrow{d} N[0, \sigma^2Q] \]
Combining terms
\[ \sqrt{n}(b - \beta) = Q^{-1} \left( \frac{1}{\sqrt{n}} \right) X'\epsilon \xrightarrow{d} N[0, \sigma^2 Q^{-1}] \]
we obtain the asymptotic distribution of \( b \).

**THEOREM**  Asymptotic Distribution of \( b \) with Independent Observations
If \( \{\epsilon_i\} \) are independently distributed with mean zero and finite variance \( \sigma^2 \) and \( x_{ik} \) is such that the Grenander conditions\(^1\) are met, then
\[ b \overset{d}{\sim} N \left[ \beta, \frac{\sigma^2}{n} Q^{-1} \right] \]

### 5.2.3 Consistency of \( s^2 \) and the Estimator of \( \text{Asy.Var}[b] \)

The purpose here is to assess the consistency of \( s^2 \) as an estimator of \( \sigma^2 \).

\[
s^2 = \frac{1}{n-K} \epsilon'M\epsilon \\
= \frac{n}{n-K} \left[ \frac{\epsilon' \epsilon}{n} - \left( \frac{\epsilon'X}{n} \right) \left( \frac{X'X}{n} \right)^{-1} \left( \frac{X'\epsilon}{n} \right) \right]
\]

\( n/(n-K) \) converges to 1 and \( \epsilon'X/n \) converges to 0. That leaves,

\[ \bar{\epsilon}^2 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \]

By the Markov Theorem, what is needed is for \( E[|\epsilon_i^2|^{1+\delta}] \) to be finite, so the minimal assumption thus far is that \( \epsilon_i \) have finite moments up to slightly greater than 2. Then,

\[ \text{plims}^2 s^2 = \sigma^2 \]

by the product rule,

\[ \text{plims}^2 (X'X/n)^{-1} = \sigma^2 Q^{-1} \]

The appropriate estimator of the asymptotic covariance matrix of \( b \) is

\[ \text{Est. Asy.Var}[b] = s^2 (X'X)^{-1} \]

---

\(^1\)Grenander Conditions for Well Behaved Data:

G1. For each column of \( X \), \( x_k \), if \( d^2_{nk} = x_k'x_k \) then \( \lim_{n \to \infty} d^2_{nk} = +\infty \). Hence, \( x_k \) does not degenerate to a sequence of zeros. Sums of squares will continue to grow as the sample size increases. No variable will degenerate to a sequence of zeros.

G2. \( \lim_{n \to \infty} x_{ik}^2/d^2_{nk} = 0 \) for all \( i = 1, ..., n \). This condition implies that no single observation will ever dominate \( x_k'x_k \), and as \( n \to \infty \), individual observations will become less important.

G3. Let \( R_n \) be the sample correlation matrix of the columns of \( X \), excluding the constant term if there is one. Then \( \lim_{n \to \infty} R_n = C \), a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that \( X \) has full rank in a finite sample, so this assumption ensures that the condition will never be violated.
5.2.4 Asymptotic Distribution of A Function of \( b \): the Delta Method

Let \( f(b) \) be a set of \( J \) continuous, linear or nonlinear and continuously differentiable functions of the least squares estimator,

\[
C(b) = \frac{\partial f(b)}{\partial b'}
\]

where \( C \) is the \( J \times K \) matrix whose \( j \)th row is the vector of derivatives of the \( j \)th function with respect to \( b' \). By the Slutsky Theorem

\[
\text{plim} f(b) = f(\beta)
\]

and

\[
\text{plim} C(b) = \frac{\partial f(b)}{\partial b'} = \Gamma
\]

Using our usual linear Taylor series approach, we expand this set of functions in the approximation

\[
f(b) = f(\beta) + \Gamma \times (b - \beta) + \text{higher-order terms}
\]

The higher-order terms become negligible in large samples if \( \text{plim} b = \beta \). Thus the mean of the asymptotic distribution is \( \text{plim} f(b) = f(\beta) \), and the asymptotic covariance matrix is

\[
\{\Gamma \text{Asy.Var} (b - \beta) \Gamma'\}
\]

**THEOREM** Asymptotic Distribution of a Function of \( b \)

If \( f(b) \) is a set of continuous and continuously differentiable functions of \( b \) such that \( \Gamma = \partial f(\beta)/\partial \beta' \), then

\[
f(b) \overset{\text{a}}{\sim} N \left[ f(\beta), \Gamma \left( \frac{\sigma^2}{n} Q^{-1} \right) \Gamma' \right]
\]

In practice, the estimator of the asymptotic covariance matrix would be

\[
\text{Est.Asy.Var}[f(b)] = C[\text{var}(b - \beta)] C'
\]

5.2.5 Asymptotic Efficiency

**DEFINITION** Asymptotic Efficiency

An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

5.3 More General Cases

5.3.1 Heterogeneity in the Distributions of \( x_i \)

In a panel data set, the sample will consist of multiple observations on each of many observational units. The panel data set could be treated as follows. Assume for the moment that the data consist of a fixed number of observations, say \( T \), on a set of \( N \) families, so that the total number of rows in \( X \) is \( n = NT \). The matrix

\[
\bar{Q}_n = \frac{1}{n} \sum_{i=1}^{n} Q_i
\]
in which \( n \) is all the observations in the sample, could be viewed as

\[
\bar{Q}_n = \frac{1}{N} \sum_i \frac{1}{T} \sum_{\text{observations for family } i} Q_{ij} = \frac{1}{N} \sum_{i=1}^N Q_i,
\]

where \( \bar{Q}_i \) = average \( Q_{ij} \) for family \( i \). We might then view the set of observations on the \( i \)th unit as if they were a single observation and apply our convergence arguments to the number of families increasing without bound.

### 5.3.2 Dependent Observations

The second difficult case arises when there are lagged dependent variables among the variables on the right-hand side or, more generally, in time series settings in which the observations are no longer independent or even uncorrelated. Consider following model

\[
y_t = \sum_{i=1}^p \gamma_i y_{t-i} + \epsilon_t
\]

Every observation \( y_t \) is determined by the entire history of the disturbances. Therefore, we have lost the crucial assumption \( E[\epsilon | X] = 0 \), the Gauss-Markov theorem no longer applies.

To establish counterparts to the limiting normal distribution of \((1/\sqrt{n})X'\epsilon\) and convergence of \((1/n)X'X\) to a finite positive definite matrix, it is necessary to make additional assumptions about the regressors. For the disturbances, we replace Assumption A3 following.

**AD3.** \( E[\epsilon_t | X_{t-s}] = 0 \), for all \( s \geq 0 \)

This assumption states that the disturbances are not correlated with any of the history.

We will also replace Assumption A5 with two assumptions. First,

\[
\text{plim} \frac{1}{T-s} \sum_{t=s+1}^T x_t x'_t = Q(s), \quad \text{a finite matrix}, s \geq 0
\]

and \( Q(0) \) is nonsingular if \( T > K \). Second, we assume that the roots of the polynomial

\[
1 - \gamma_1 z - \gamma_2 z^2 - \ldots - \gamma_p z^p = 0
\]

are all outside the unit circle. Formally, we obtain these results with

**AD5.** The series on \( x_t \) is **stationary** and **ergodic**.

This assumption also implies that \( Q(S) \) becomes a matrix of zeros as \( s \) becomes large. These conditions are sufficient to produce \((1/n)X'\epsilon \to 0\) and the consistency of \( b \).

### 5.4 Instrumental Variable and Two-Stage Least Squares Estimation

Suppose that in the classical model \( y_i = x'_i \beta + \epsilon_i \), the \( K \) variables \( x_i \) may be correlated with \( \epsilon_i \). Suppose as well that there exists a set of \( L \) variables \( z_i \), where \( L \) is at least as large as \( K \), such that \( z_i \) is correlated with \( x_i \) but not with \( \epsilon_i \). We cannot estimate \( \beta \) consistently by
using the familiar least squares estimator. But we can construct a consistent estimator of $\beta$ by using the assumed relationships among $z_i$, $x_i$, and $\epsilon_i$.

That is

$$E[x_i\epsilon_i] = \gamma$$

for some nonzero $\gamma$. If the data are “well behaved”, then

$$\text{plim}(1/n)X'\epsilon = \gamma$$

We assume the following:

AI7. $[x_i, z_i, \epsilon_i]$, $i = 1, \ldots, n$, are an i.i.d. sequence of random variables.

AI8a. $E[x_i^2] = Q_{xx, kk} \leq \infty$, a finite constant, $k = 1, \ldots, K$

AI8b. $E[z_i^2] = Q_{zz, ll} \leq \infty$, a finite constant, $l = 1, \ldots, L$

AI8c. $E[z_i x_i] = Q_{zx, lk} \leq \infty$, a finite constant, $l = 1, \ldots, L$, $k = 1, \ldots, K$

AI9. $E[\epsilon_i | z_i] = 0$

We have

$$\text{plim}(1/n)Z'Z = Q_{zz}, \text{ a finite, positive definite (assumed) matrix,}$$

$$\text{plim}(1/n)Z'X = Q_{zx}, \text{ a finite, } L \times K \text{ matrix with rank } K \text{ (assumed),}$$

$$\text{plim}(1/n)Z'\epsilon = 0.$$  

For this more general model, the estimator $b$ is no longer unbiased;

$$E[b | X] = \beta + (X'X)^{-1}X'\eta \neq \beta$$

and it is also inconsistent

$$\text{plim}b = \beta + \text{plim} \left( \frac{X'X}{n} \right)^{-1} \text{plim} \left( \frac{X'\epsilon}{n} \right)^{-1} = \beta + Q_{xx}^{-1}\gamma \neq \beta$$

We can now turn to the instrumental variable estimator. Suppose $Z$ has the same number of variables as $X$. We have assumed that the rank of $Z'X$ is $K$, so now $Z'X$ is a square matrix. It follows that

$$\left[ \text{plim} \left( \frac{Z'X}{n} \right) \right]^{-1} \text{plim} \left( \frac{Z'y}{n} \right) = \beta$$

which leads us to the instrumental variable estimator,

$$b_{IV} = (Z'X)^{-1}Z'y$$

For the asymptotic distribution, first

$$\sqrt{n}(b_{IV} - \beta) = \left( \frac{Z'X}{n} \right)^{-1} \frac{1}{\sqrt{n}} Z'\epsilon,$$

which has the same limiting distribution as $Q_{xx}^{-1}[(1/\sqrt{n})Z'\epsilon]$. It follows that

$$\left( \frac{1}{\sqrt{n}} Z'\epsilon \right) \xrightarrow{d} N[0, \sigma^2 Q_{zz}]$$

and

$$\left( \frac{Z'X}{n} \right)^{-1} \left( \frac{1}{\sqrt{n}} Z'\epsilon \right) \xrightarrow{d} N[0, \sigma^2 Q_{xx}^{-1}Q_{zz}Q_{xx}^{-1}]$$

Finally, these compose the following theorem
THEOREM  Asymptotic Distribution of the Instrumental Variables Estimator

If Assumptions A1, A2, A13, A4, AS5, AS5a, AI7, AI89a-c and AI9 all hold for \([y_i, x_i, z_i, \epsilon_i]\), where \(z\) is a valid set of \(L = K\) instrumental variables, then the asymptotic distribution of the instrumental variables estimator \(b_{IV} = (Z'X)^{-1}Z'y\) is

\[
b_{IV} \sim N \left[ \beta, \frac{s^2}{n} Q_{xx}'Q_{xz}Q_{zx}^{-1} \right]
\]

where \(Q_{zx} = \plim(Z'X/n)\) and \(Q_{zx} = \plim(Z'Z/n)\).

To estimate the asymptotic covariance matrix, we will require an estimator of \(s^2\). The natural estimator is

\[
s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'b_{IV})^2
\]

There is a remaining detail. If \(Z\) contains more variables than \(X\), then much of the preceding is unusable, because \(Z'X\) will be \(L \times K\) with rank \(K < L\) and will thus not have an inverse. A better choice is the projection of the columns of \(X\) in the column space of \(Z\)

\[
\hat{X} = Z(Z'Z)^{-1}Z'X
\]

then we have

\[
b_{IV} = (\hat{X}'X)^{-1}\hat{X}'y
\]
\[
= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y
\]

The proofs of consistency and asymptotic normality for this estimator are exactly the same as before.

In practice, \(b_{IV}\) can be computed in two steps, first by computing \(\hat{X}\), then by the least squares regression. For this reason, this is called the **two-stage least squares (2SLS)** estimator.

One should be careful of this approach, however, in the computation of the asymptotic covariance matrix; \(s^2\) should not be based on \(\hat{X}\). The estimator

\[
s^2_{IV} = \frac{(y - \hat{X}b_{IV})(y - \hat{X}b_{IV})}{n}
\]

is inconsistent for \(s^2\), with or without a correction for degrees of freedom.

### 5.5 Hausman’s Specification Test

The difference between the asymptotic covariance matrices of the two estimators is

\[
\text{Asy.Var}[b_{IV}] - \text{Asy.Var}[b_{LS}] = \frac{s^2}{n} \plim \left[ (X'Z(Z'Z)^{-1}Z'X)^{-1} - (X'X)^{-1} \right]
\]

Since \(X'Z(Z'Z)^{-1}Z'X\) is smaller, in the matrix sense, than \(X'X\), its inverse is larger. That is if LS is consistent, then it is a preferred estimator.

The logic of Hausman’s approach is as follows. Define \(d = b_{IV} - b_{LS}\). Under the null hypothesis, \(\plim d = 0\), whereas under the alternative, \(\plim d \neq 0\). Test this hypothesis with a Wald statistic,

\[
H = d'\{\text{Est.Asy.Var}[d]\}^{-1}d
\]
The asymptotic covariance matrix
\[
\text{Asy.Var}[b_{IV} - b_{LS}] = \text{Asy.Var}[b_{IV}] + \text{Asy.Var}[b_{LS}] - \text{Asy.Cov}[b_{IV}, b_{LS}] - \text{Asy.Cov}[b_{LS}, b_{IV}]
\]

Hausman gives a fundamental result

the covariance between an efficient estimator, \(b_E\), of a parameter vector, \(\beta\), and its difference from an inefficient estimator, \(b_I\), of the same parameter vector, \(b_E - b_I\), is zero.

For our case, \(b_E\) is \(b_{LS}\) and \(b_I\) is \(b_{IV}\). By Hausman’s result

\[
\text{Cov}[b_E, b_E - b_I] = \text{Var}[b_E] - \text{Cov}[b_E, b_I] = 0
\]

so,

\[
\text{Asy.Var}[b_{IV} - b_{LS}] = \text{Asy.Var}[b_{IV}] - \text{Asy.Var}[b_{LS}]
\]

then the statistic will be

\[
H = \frac{d'[\hat{X}'\hat{X}]^{-1} - (X'X)^{-1}d}{s^2}
\]

However, unless \(X\) and \(Z\) have no variables in common, the rank of the matrix in this statistic is less than \(K\), and the ordinary inverse will not even exist.

Since \(Z'(Z'Z)^{-1}Z'\) is an idempotent matrix, \(\hat{X}'\hat{X} = \hat{X}'X\). Then,

\[
d = (\hat{X}'\hat{X})^{-1}\hat{X}'y - (X'X)^{-1}X'y
\]

\[
= (\hat{X}'\hat{X})^{-1}[\hat{X}'y - (\hat{X}'\hat{X})(X'X)^{-1}X'y]
\]

\[
= (\hat{X}'\hat{X})^{-1}\hat{X}'(y - X(X'X)^{-1}X'y)
\]

\[
= (\hat{X}'\hat{X})^{-1}\hat{X}'e
\]

where \(e\) is the vector of least squares residuals. Recall that \(K_0\) of the columns in \(\hat{X}\) are the original variables in \(X\). Suppose that these variables are the first \(K_0\). Thus, the first \(K_0\) rows of \(\hat{X}'e\) are the same as the first \(K_0\) rows of \(X'e\), which are, of course 0. So, we can write \(d\) as

\[
d = (\hat{X}'\hat{X})^{-1}0_{X's} = (\hat{X}'\hat{X})^{-1}\begin{pmatrix} 0 \\ X'se \end{pmatrix}
\]

Finally, denote the entire matrix in \(H\) by \(W\). Then, denoting the whole matrix product by \(P\), we obtain

\[
H = [0' \ q'](\hat{X}'\hat{X})^{-1}W(\hat{X}'\hat{X})^{-1}\begin{pmatrix} 0 \\ q* \end{pmatrix} = q*P_{ss'q*}
\]

where \(P_{ss'}\) is the lower right \(K' \times K'\) submatrix of \(P\).

In preceding, Wald test requires a generalized inverse. An alternative approach devised by Wu(1993) is simpler. An \(F\) statistic with \(K'\) and \(n - K_0'\) degrees of freedom can be used to est the joint significance of the elements of \(\gamma\) in the augmented rgression

\[
y = X\beta + \hat{X}'\gamma + \epsilon^*
\]

where \(\hat{X}'\) are the fitted values in regressions of the variables in \(X^*\) on \(Z\).

Hausman test is general. We have a pair of estimators, \(\hat{\theta}_E\) and \(\hat{\theta}_I\), such that under \(H_0:\ \hat{\theta}_E\) and \(\hat{\theta}_I\) are both consistent and \(\hat{\theta}_E\) is efficient relative to \(\hat{\theta}_I\), while under \(H_1:\ \hat{\theta}_I\) remains consistent while \(\hat{\theta}_E\) is inconsistent, then we can form a test of the hypothesis by referring the "Hausman statistic"

\[
H = (\hat{\theta}_I - \hat{\theta}_E)'\{\text{Est.Asy.Var}[\hat{\theta}_I] - \text{Est.Asy.Var}[\hat{\theta}_E]\}^{-1}(\hat{\theta}_I - \hat{\theta}_E) \xrightarrow{d} \chi^2[J]
\]

21
5.6 Measurement Error

5.6.1 Least Squares Attenuation

Assume that the model
\[ y^* = \beta x^* + \epsilon \]
conforms to all the assumptions of the classical normal regression model. Suppose, however, that the observed data are only imperfectly measured versions of \( y^* \) and \( x^* \). Observed \( y \) and \( x \) contain errors of measurement. We assume that,
\[ y = y^* + v \quad \text{with} \quad v \sim N[0, \sigma_v^2], \]
\[ x = x^* + u \quad \text{with} \quad u \sim N[0, \sigma_u^2] \]
Assume, as well, that \( u \) and \( v \) are independent of each other and of \( y^* \) and \( x^* \). By substituting, obtain
\[ y = \beta x + [\epsilon - \beta u] = \beta x + w \]
Note, measurement error on the dependent variable can be absorbed into the disturbance of the regression and ignored. The regressor above is correlated with the disturbance:
\[ \text{Cov}[x, w] = \text{Cov}[x^* + u, \epsilon - \beta u] = -\beta \sigma_u^2 \]
This result violates one of the central assumptions of the classical model. The LS estimator \( \hat{b} \) is inconsistent. To find the probability limits,
\[ \text{plim} \frac{1}{n} \sum_{i=1}^{n} (x_i^* + u_i)(\beta x_i^* + \epsilon_i) = \frac{\beta}{1 + \sigma_u^2/Q^*} \]
where \( Q^* = \text{plim}(1/n) \sum_i x_i^* \). Clearly, the greater the variability in the measurement error, the worse the bias. The effect of biasing the coefficient toward zero is called attenuation.

In a multiple regression model, matters only get worse.
\[ \text{plim} b = \left[ Q^* + \Sigma_{uu} \right]^{-1} Q^* \beta = \beta - \left[ Q^* + \Sigma_{uu} \right]^{-1} \Sigma_{uu} \beta \]
This probability limit is a mixture of all the parameters in the model.

5.6.2 Instrumental Variables Estimation

For the general case, \( y = X^* \beta + \epsilon \), \( X = X^* + U \), suppose that there exists a matrix of variables \( Z \) that is not correlated with the disturbances or the measurement error but is correlated with regressors, \( X \). \( b_{IV} = (Z'X)^{-1}Z'y \) is consistent and asymptotically normally distributed with asymptotic covariance matrix that is estimated with
\[ \text{Est.Asy.Var}[b_{IV}] = \hat{\sigma}^2[Z'X]^{-1}[Z'Z][X'X]^{-1} \]
Chapter 6

Inference and Prediction

6.1 Introduction

In this chapter, we will examine some applications of hypothesis tests using the classical model.

6.2 Restrictions and Nested Models

Consider for example, a simple model of investment, \( I_t \),

\[
\ln I_t = \beta_1 + \beta_2 i_t + \beta_3 \Delta p_t + \beta_4 \ln Y_t + \beta_5 t + \epsilon_t
\]

which states that investors are sensitive to nominal interest rate, \( i_t \), the rate of inflation, \( \Delta p_t \), real output, \( \ln Y_t \), and other factors which trend upward through time, embodied in the time trend, \( t \). Consider the stronger hypothesis, “investors care only about real interest rates”. This resulting equation

\[
\ln I_t = \beta_1 + \beta_2 (i_t - \Delta p_t) + \beta_4 \ln Y_t + \beta_5 t + \epsilon_t
\]

is now restricted; in the context of the first model, the implication is that \( \beta_2 + \beta_3 = 0 \). This subset of values is contained within the unrestricted set. In this way, the models are said to be nested.

Now, consider an alternative pair of models: Model 0: “Investors care only about inflation;” Model 1: “Investors care only about nominal interest rate.” In this case, the two specifications are both subsets of the unrestricted model, but neither model is obtained as a restriction on the other. These two models are nonnested.

6.3 Two Approaches to Testing Hypotheses

We can ask whether the estimates come reasonably close to satisfying the restrictions implied by the hypothesis. We can also ascertain whether this loss of fit results merely from sampling error or whether it is so large as to cast doubt on the validity of the restrictions.

6.3.1 The \( F \) Statistic and the Least Squares Discrepancy

We now consider testing a set of \( J \) linear restrictions stated in the null hypothesis

\[
H_0 : R\beta - q = 0
\]
against the alternative hypothesis

\[ H_0 : R\beta - q \neq 0 \]

Given the least squares estimator \( b \), our interest centers on the discrepancy vector \( Rb - q = 0 \). It is unlikely that \( m \) will be exactly 0. The statistical question is whether the deviation of \( m \) from 0 can be attributed to sampling error or whether it is significant.

We can base a test of \( H_0 \) on the Wald criterion:

\[
W = m'\{\text{Var}[m|X]\}^{-1}m = (Rb - q)'[\sigma^2R(X'X)^{-1}R']^{-1}(Rb - q)
\]

\[
= \frac{(Rb - q)'[\sigma^2R(X'X)^{-1}R']^{-1}(Rb - q)}{\sigma^2}
\]

\[
\sim \chi^2(J)
\]

However, this chi-squared statistic is not usable because of the unknown \( \sigma^2 \). By using \( s^2 \) instead of \( \sigma^2 \) and dividing the result by \( J \), we obtained a usable \( F \) statistic with \( J \) and \( n - K \) degrees of freedom. Then,

\[
F(J, n - K) = \frac{W \sigma^2}{J s^2} = \frac{(Rb - q)'[\sigma^2R(X'X)^{-1}R']^{-1}(Rb - q)/J}{[(n - K)s^2/\sigma^2]/(n - K)} = \frac{(Rb - q)'[\sigma^2R(X'X)^{-1}R']^{-1}(Rb - q)}{J}
\]

For testing one linear restriction of the form

\[ H_0 : r_1\beta_1 + r_2\beta_2 + \ldots + r_k\beta_k = r'\beta = q \]

Consider an alternative approach. The sample estimate of \( r'\beta \) is

\[ r_1b_1 + r_2b_2 + \ldots + r_Kb_K = r'b = \hat{q} \]

If \( \hat{q} \) differs significantly from \( q \), then we conclude that the sample data are not consistent with the hypothesis. It is natural to base the test on

\[ t = \frac{\hat{q} - q}{\text{se} (\hat{q})} \]

There is a useful relationship between the statistics. We can write the square of \( t \) statistic as

\[
t^2 = \frac{(\hat{q} - q)^2}{\text{Var}(\hat{q} - q|X)} = \frac{(r'b - q)'[s^2(X'X)^{-1}r']^{-1}(r'b - q)}{1}
\]

It follows, therefore, that for testing a single restriction, the \( t \) statistic is the square root of the \( F \) statistic that would be used to test that would be used to test that hypothesis.
6.3.2 The Restricted Least Squares Estimator

A different approach to hypothesis testing focuses on the fit of the regression. The restricted least squares estimator is obtained as the solution to

\[
\min_{b_0} S(b_0) = (y - Xb_0)'(y - Xb_0) \text{ subject to } Rb_0 = q
\]

A lagrangian function for this problem can be written

\[
L^*(b_0, \lambda) = (y - Xb_0)'(y - Xb_0) + 2\lambda'(Rb_0 - q)
\]

The solutions \(b_\star\) and \(\lambda_\star\) will satisfy the necessary conditions

\[
\frac{\partial L^*}{\partial b_\star} = -2X'(y - Xb_\star) + 2R'R_\star = 0
\]
\[
\frac{\partial L^*}{\partial \lambda_\star} = 2(Rb_\star - q) = 0
\]

The explicit solutions are

\[
b_\star = b - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(Rb - q) = b - Cm
\]
\[
\lambda_\star = [R(X'X)^{-1}R']^{-1}(Rb - q)
\]

Also,

\[
\text{Var}[b_\star | X] = \sigma^2(X'X)^{-1} - \sigma^2(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}
\]

\[
= \text{Var}[b | X] - \text{a nonnegative definite matrix}
\]

The constrained solution \(b_\star\) is equal to the unconstrained solution \(b\) plus a term that accounts for the failure of the unrestricted solution to satisfy the constraints.

6.3.3 The Loss of the Fit from Restricted Least Squares

Consider the change in the fit of a multiple regression when a variable \(z\) is added to a model that already contains \(K - 1\) variables, \(x\).

\[
R^2_{xz} = R^2_X + (1 - R^2_X)r_{yz}^2
\]

where \(R^2_{xz}\) is the new \(R^2\) after \(z\) is added, \(R^2_X\) is the original \(R^2\) and \(r_{yz}\) is the partial correlation between \(y\) and \(z\), controlling for \(x\).

The fit for the restricted least squares coefficients cannot be better than that of the unrestricted solution. Let \(e_\star\) equal \(y - Xb_\star\). Then, using a familiar device,

\[
e_\star = y - Xb - X(b_\star - b) = e - X(b_\star - b)
\]

The new sum of squared deviations is

\[
'e_\star e_\star = e'e + (b_\star - b)'X'X(b_\star - b) \geq e'e
\]

The loss of fit is

\[
e_\star e_\star - e'e = (Rb - q)'[R(X'X)^{-1}R']^{-1}(Rb - q)
\]
6.4 Nonnormal Disturbances and Large Sample Tests

In this section, we will examine results that will generalize the familiar procedures. These large-sample results suggest that although the usual $t$ and $F$ statistics are still usable, in the more general case without the special assumption of normality, they are viewed as approximations whose quality improves as the sample size increases.

Assuming the data are well behaved, the asymptotic distribution of the least squares coefficient estimator $b$, is given by

$$ b \xrightarrow{a} N \left[ \beta, \frac{\sigma^2}{n} Q^{-1} \right] $$

where $Q = \text{plim} \left( \frac{X'X}{n} \right)$

The second result we will need concerns the estimator of $\sigma^2$

$$ \text{plims}^2 = \sigma^2, \quad \text{where} \quad s^2 = \frac{e'e}{n-K} $$

The sample statistic for testing the hypothesis that one of the coefficients, $\beta_k$ equals a particular value $\beta^0_k$ is

$$ t_k = \frac{\sqrt{n}(b_k - \beta^0_k)}{\sqrt{\sigma^2(X'X)/n_k}} $$

Under the null hypothesis, with normally distributed disturbances, $t_k$ is exactly distributed as $t$ with $n-K$ degrees of freedom. Denominator of $t_k$ converges to $\sqrt{\sigma^2 Q^{-1}}$. The large-sample distribution of $t_k$ is the same as that of

$$ \tau_k = \frac{\sqrt{n}(b_k - \beta^0_k)}{\sqrt{\sigma^2 Q^{-1}}} $$

But $\tau_k = (b_k - E[b_k])/(\text{Asy.Var}[b_k])^{1/2}$ from the asymptotic normal distribution (under the hypothesis $\beta_k = \beta^0_k$), so it follows that $\tau_k$ has a standard normal asymptotic distribution, and this result is the large-sample distribution of our $t$ statistic. Thus, as a large sample approximation, we will use the standard normal distribution to approximate the true distribution of the test statistic $t_k$ and use the critical values from the standard normal distribution for testing hypotheses.

For $F$ statistics

$$ F = \frac{(Rb - q)'[\sigma^2 R(X'X)^{-1}R']^{-1}(Rb - q)}{J(s^2/\sigma^2)} $$

Since plims$^2 = \sigma^2$, and plim$(X'X/n) = Q$, the denominator of $F$ converges to $J$, if $F$ has a limiting distribution, the it is the same as the limiting distribution of

$$ W^* = \frac{1}{J}(Rb - q)'[\sigma^2/n Q^{-1} R']^{-1}(Rb - q) $$

$$ = \frac{1}{J}(Rb - q)'\{\text{Asy.Var}[Rb-q]\}^{-1}(Rb - q) $$

**THEOREM** Limiting Distribution of the Wald Statistic

If $\sqrt{n}(b - \beta) \xrightarrow{d} N[0, \sigma^2 Q^{-1}]$ and if $H_0 : R \beta - q = 0$ is true, then

$$ W = (Rb - q)'\{R \sigma^2 (X'X)^{-1} R'\}^{-1}(Rb - q) = JF \xrightarrow{d} \chi^2[J] $$
Proof: Since $R$ is a matrix of constants and $R\beta = q$,

$$\sqrt{n}R(b - \beta) = \sqrt{n}(Rb - q) \xrightarrow{d} N[0, R(\sigma^2Q^{-1})R']$$

For convenience, write this equation as

$$z \xrightarrow{d} N[0, P]$$

We define the inverse square root of a positive definite matrix $P$ as another matrix, say $T$ such that $T^2 = P^{-1}$, and denote $T$ as $P^{-1/2}$. Let $T$ be the inverse square root of $P$. Then, by the same reasoning

if $z \xrightarrow{d} N[0, P]$, then $P^{-1/2}z \xrightarrow{d} N[0, P^{-1/2}PP^{-1/2}] = N[0, I]$

The limiting distribution of

$$(P^{-1/2}z)'(P^{-1/2}z) = z'P^{-1}z \xrightarrow{d} \chi^2(J)$$

Reassembling the parts from before, we have shown that the limiting distribution of

$$n(Rb - q)'[R(\sigma^2Q^{-1})R']^{-1}(Rb - q)$$

is chi-squared, with $J$ degrees of freedom.

The appropriate critical values for the $F$ test of the restrictions $R\beta - q = 0$ converge from above to $1/J$ times those for a chi-squared test based on the Wald statistic.

6.5 Testing Nonlinear Restrictions

The general problem is that of testing a hypothesis that involves a nonlinear function of the regression coefficients:

$$H_0 : c(\beta) = q$$

The counterpart to the test statistic we used earlier would be

$$z = \frac{c(\hat{\beta}) - q}{\text{estimated standard error}}$$

A linear Taylor series approximation to $c(\hat{\beta})$ around the true parameter vector $\beta$ is

$$c(\hat{\beta}) \approx c(\beta) + \left( \frac{\partial c(\beta)}{\partial \beta} \right)(\hat{\beta} - \beta)$$

If plim $\hat{\beta} = \beta$, then we are justified in using $c(\hat{\beta})$ as an estimate of $c(\beta)$. The variance of the nonlinear function is approximately equal to the variance of the right-hand side, which is, then

$$\text{Var}[c(\hat{\beta})] \approx \left( \frac{\partial c(\beta)}{\partial \beta} \right)' \text{Var}[\hat{\beta}] \left( \frac{\partial c(\beta)}{\partial \beta} \right)$$

We use our sample estimates in computing the derivatives, and using $g(\hat{\beta})$ to estimate $g(\beta) = \partial c(\beta)/\partial \beta$. To estimate the variance of the estimator, we can use $s^2(X'X)^{-1}$. 

27
6.6 Prediction

After the estimation of parameters, a common use of regression is for prediction. Suppose that we wish to predict the value of \( y^0 \) associated with a regressor vector \( x^0 \). This value would be

\[
y^0 = x^0' \beta + \epsilon^0
\]

It follows from the Gauss-Markov theorem that \( \hat{y}^0 = x^0' b \) is the minimum variance linear unbiased estimator of \( E[y^0|x^0] \). The forecast error is

\[
e^0 = y^0 - \hat{y}^0 = (\beta - b)' x^0 + \epsilon^0
\]

The prediction variance to be applied to this estimate is

\[
\text{Var}[e^0|X, x^0] = \sigma^2 + \text{Var}[(\beta - b)' x^0|X, x^0] = \sigma^2 + x^0' [\sigma^2 (X'X)^{-1}] x^0
\]

If the regression contains a constant term, then an equivalent expression is

\[
\text{Var}[e^0] = \sigma^2 \left[ 1 + \frac{1}{n^0} + \sum_{j=1}^{K-1} \sum_{k=1}^{K-1} (x_j^0 - \bar{x}_j)(x_k^0 - \bar{x}_k)(Z'M^0Z)^{jk} \right]
\]

where \( Z \) is the \( K-1 \) columns of \( X \) not including the constant. The first term \( \sigma^2 \) is constant, which implies that no matter how much data we have, we can never predict perfectly.

Various measures have been proposed for assessing the predictive accuracy of forecasting models. Two measures that are based on the residuals from the forecasts are the root mean squared error

\[
RMSE = \sqrt{\frac{1}{n^0} \sum_i (y_i - \hat{y}_i)^2}
\]

and the mean absolute error

\[
MAE = \frac{1}{n^0} \sum_i |y_i - \hat{y}_i|
\]

where \( n^0 \) is the number of periods being forecasted. These statistics have an obvious scaling problem. Several measures that are scale free are based on the Theil U Statistic:

\[
U = \sqrt{\frac{(1/n^0) \sum (y_i - \hat{y}_i)^2}{(1/n^0) \sum y_i^2}}
\]

Large values indicate a poor forecasting performance. An alternative is to compute the measure in terms of the changes in \( y \):

\[
U_\Delta = \sqrt{\frac{(1/n^0) \sum (\Delta y_i - \Delta \hat{y}_i)^2}{(1/n^0) \sum \Delta y_i^2}}
\]

---

1Prediction, not forecasting. The term “forecasting” is usually for usage of the time series models.
Chapter 7

Functional Form and Structural Change

7.1 Introduction

In this chapter, we are concerned with the functional form of the regression model.

7.2 Using Binary Variables

7.2.1 Binary Variables in Regression

In recent applications, researchers in many fields have studied the effects of treatment on some kind of response. These examples can all be formulated in regression models involving a single dummy variable:

\[ y_i = x_i' \beta + \delta d_i + \epsilon_i \]

One of the important issues in policy analysis concerns measurement of such treatment effects when the dummy variable results from an individual participation decision.

A dummy variable that takes the value one only for one observation has the effect of deleting that observation from computation of the least squares slopes and variance estimator.

7.2.2 Several Categories

When there are several categories, a set of binary variables is necessary. Correcting for seasonal factors in macroeconomic data is a common application. We could write a consumption function for quarterly data as

\[ C_t = \beta_1 + \beta_2 x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \epsilon_t \]

where \( x_t \) is disposable income. Note that only three of the four quarterly dummy variables are included in the model. If the fourth were included, then the four dummy variables would sum to one at every observation\(^1\), which would reproduce the constant term—a case of perfect multicollinearity. This is known as the dummy variable trap.

\(^1\)Depending on the application, it might be preferable to have four separate dummy variables and drop the overall constant.
The preceding is a means of deseasonalizing the data. Consider the alternative formulation:

\[ C_t = \beta x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \epsilon_t \]

### 7.2.3 Threshold Effects and Categorical Variables

There are cases, however, in which the dummy variables represents levels of some underlying factor that might have been measured directly if this were possible. For example, education is a case in which we typically observe certain thresholds rather than, say, years of education. For example, that our interest is in a regression of the form

\[ \text{income} = \beta_1 + \beta_2 \text{age} + \text{effect of education} + \epsilon \]

The data on education might consist of the highest level of education attained, such as high school (HS), undergraduate (B), master’s (M), or Ph.D (P). A flexible model would use three or four binary variables, one for each level of education. Thus, we would write

\[ \text{income} = \beta_1 + \beta_2 \text{age} + \delta_B B + \delta_M M + \delta_P P + \epsilon \]

### 7.2.4 Spline Regression

If one is examining income data for a large cross section of individuals of varying ages in a population, then certain patterns with regard to some age thresholds will be clearly evident. In particular, throughout the range of values of age, income will be rising, but the slope might change at some distinct milestones, for example, at age 18, when the typical individual graduates from high school, and at age 22, when he or she graduates from college. Restricted regression and what is known as a **spline** function can be used to achieve the desired effect.

The function we wish to estimate is

\[ E[\text{income}|\text{age}] = \begin{cases} 
\alpha^0 + \beta^0 \text{age} & \text{if age} < 18 \\
\alpha^1 + \beta^1 \text{age} & \text{if age} \geq 18 \text{ and age} < 22 \\
\alpha^2 + \beta^2 \text{age} & \text{if age} \geq 22
\end{cases} \]

The threshold values, 18, 22, are called **knots**. Let

\[ d_1 = 1 \quad \text{if age} \geq t_1^* \]
\[ d_2 = 1 \quad \text{if age} \geq t_2^* \]

where \( t_1^* = 18 \) and \( t_2^* = 22 \). To combine all three equations, we use

\[ \text{income} = \beta_1 + \beta_2 \text{age} + \gamma_1 d_1 + \delta_1 d_1 \text{age} + \gamma_2 d_2 + \delta_2 d_2 \text{age} + \epsilon \]

To make the function **piecewise continuous**, we require that the segments joint at the knots

\[ \beta_1 + \beta_2 t_1^* = (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_1^* \]

and

\[ (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_2^* = (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2) t_2^* \]

Collecting terms, we obtain

\[ \text{income} = \beta_1 + \beta_2 \text{age} + \delta_1 d_1 (\text{age} - t_1^*) + \delta_2 d_2 (\text{age} - t_2^*) + \epsilon \]

We can test the hypothesis that the slope of the function is constant with the joint test of the two restrictions \( \delta_1 = 0 \) and \( \delta_2 = 0 \).
7.3 Nonlinearity in the Variables

7.3.1 Functional Forms

A commonly used form of regression model is the loglinear model,
\[
\ln y = \ln \alpha + \sum_k \beta_k \ln X_k + \epsilon = \beta_1 + \sum_k \beta_k x_k + \epsilon
\]
In this model, the coefficients are elasticities:
\[
\left( \frac{\partial y}{\partial x_k} \right) \left( \frac{x_k}{y} \right) = \frac{\partial \ln y}{\partial \ln x_k} = \beta_k
\]
In the loglinear equation, measured changes are in proportional or percentage terms; \( \beta_k \) measures the percentage change in \( y \) associated with a one percent change in \( x_k \).

A hybrid of the linear and loglinear models is the semilog equation
\[
\ln y = \beta_1 + \beta_2 x + \epsilon
\]
The coefficients in the semilog model are partial- or semi-elasticities; in the equation above, \( \beta_2 \) is \( \partial \ln y / \partial x \).

Another useful formulation of the regression model is one with interaction terms. For example, a model relating braking distance \( D \) to speed \( S \) and road wetness \( W \) might be
\[
D = \beta_1 + \beta_2 S + \beta_3 W + \beta_4 SW + \epsilon
\]
In this model
\[
\frac{\partial E[D|S,W]}{\partial S} = \beta_2 + \beta_4 W
\]
which implies that the marginal effect of higher speed on braking distance is increased when the road is wetter (assuming that \( \beta_4 \) is positive).

7.3.2 intrinsic Linearity and Identification

The loglinear model illustrates an intermediate case of a nonlinear regression model. The equation is intrinsically linear by our definition; by taking logs of \( Y_i = \alpha X_i^{\beta_2} e^{\epsilon_i} \), we obtain
\[
\ln Y_i = \ln \alpha + \beta_2 \ln X_i + \epsilon_i
\]
or
\[
y_i = \beta_1 + \beta_2 x_i + \epsilon_i
\]
Although this equation is linear in most respects, something has changed in that it is no longer linear in \( \alpha \). Written in terms of \( \beta_1 \), we obtain a fully linear model. But that may not be the form of interest. Nothing is lost, of course, since \( \beta_1 \) is just \( \ln \alpha \). If \( \beta_1 \) can be estimated, then an obvious estimate of \( \alpha \) is suggested.

**Definition** Intrinsic Linearity

In the classical linear regression model, if the \( K \) parameters \( \beta_1, \beta_2, \ldots, \beta_k \) can be written as \( K \) one-to-one, possibly nonlinear functions of a set of \( K \) underlying parameters \( \theta_1, \theta_2, \ldots, \theta_K \), then the model is intrinsically linear in \( \theta \).

The emphasis in intrinsic linearity is on “one to one”. If the conditions are met, then the model an be estimated in terms of the functions \( \beta_1, \beta_2, \ldots, \beta_k \), and the underlying parameters derived after these are estimated. The one-to-one correspondence is an identification condition. If the condition is met, then the underlying parameters of the regression (\( \theta \)) are said to be exactly identified in terms of the parameters of the linear model \( \beta \).
7.4 Modeling and Testing for a Structural Break

One of the more common applications of the $F$ test is in tests of structural change.

7.4.1 Different Parameter Vectors

The gasoline consumption data span two very different periods. The embargo of 1973 marked a transition in this market. It is possible that the entire relationship described by our regression model changed in 1974. To test this as a hypothesis, we could proceed as follows: Denote the first 14 years of the data in $y$ and $X$ as $y_1$ and $X_1$ and the remaining years as $y_2$ and $X_2$.

Denoting the data matrices as $y$ and $X$, we find that the unrestricted least squares estimator is

$$b = (X'X)^{-1}X'y = \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'y_1 \\ X_2'y_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

which is least squares applied to the two equations separately. Therefore, the total sum of squared residuals from this regression will be the sum of the two residual sums of squares from the two separate regressions:

$$e'e = e'_1e_1 + e'_2e_2$$

The restricted coefficient vector can be obtained in two ways. Formally, the restriction $\beta_1 = \beta_2$ is $R\beta = q$, where $R = [I : -I]$ and $q = 0$. The general result given earlier can be applied directly. An easier way to proceed is to build the restriction directly into the model. If the two coefficient vectors are the same, the

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

7.4.2 Insufficient Observations

In some circumstances, the data series are not long enough to estimate one or the other of the separate regressions for a test of structural change. Fisher (1970) has shown that in such a circumstance, a valid way to proceed is as follows:

1. Estimate the regression, using the full data set, and compute the restricted sum of squared residuals, $e'_s e_s$.

2. Using the longer subperiod to estimate the regression, and compute the unrestricted sum of squares, $e'_1 e_1$. This latter computation is done assuming that with only $n_2 < K$ observations, we could obtain a perfect fit and thus contribute zero to the sum of squares.

3. The $F$ statistic is then computed, using

$$F[n_2, n_1 - K] = \frac{(e'_s e_s - e'_1 e_1)/n_2}{e'_1 e_1/(n_1 - K)}$$

Note that the numerator degrees of freedom is $n_2$, not $K$. 
### 7.4.3 Tests of Structural Break with Unequal Variances

An important assumption made in using the Chow test is that the disturbance variance is the same in both regressions. In the restricted model, if this is not true, the first $n_1$ elements of $\epsilon$ have variance $\sigma_1^2$, whereas the next $n_2$ have variance $\sigma_2^2$, and so on. The restricted model is, therefore, heteroscedastic, and our results for the classical regression model no longer apply.

If the sample size is reasonably large, then we have a test that is valid whether or not the disturbance variances are the same. Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are two consistent and asymptotically normally distributed estimators of a parameter based on independent samples, with asymptotic covariance matrices $V_1$ and $V_2$.

$\hat{\theta}_1 - \hat{\theta}_2$ has mean 0 and asymptotic covariance matrix $V_1 + V_2$.

Under the null hypothesis, the Wald statistic,

$$W = (\hat{\theta}_1 - \hat{\theta}_2)'(\hat{V}_1 + \hat{V}_2)^{-1}(\hat{\theta}_1 - \hat{\theta}_2)$$

has a limiting chi-squared distribution with $K$ degrees of freedom. A test that the difference between the parameters is zero can be based on this statistic.

### 7.5 Tests of Model Stability

In this section, we will examine two tests that are based on the idea that a regime change might take place slowly, and at an unknown point in time, or that the regime underlying the observed data might simply not be stable at all.

#### 7.5.1 Hansen’s Test

Hansen’s (1992) test of model stability is based on a cumulative sum of the least squares residuals. From the least squares normal equations, we have

$$\sum_{t=1}^{T} x_t e_t = 0 \quad \text{and} \quad \sum_{t=1}^{T} (e_t^2 - \frac{\epsilon^2}{n}) = 0$$

Let the vector $f_t$ be the $(K+1) \times 1$ $t$th observation in this pair of sums. Then, $\sum_{t=1}^{T} f_t = 0$. Let the sequence of partial sums be $s_t = \sum_{r=1}^{t} f_r$, so $s_T = 0$. Finally, let $F = T \sum_{t=1}^{T} f_t f_t'$ and $S = \sum_{t=1}^{T} s_t s_t'$. Hansen’s test statistic can be computed simply as $H = \text{tr}(F^{-1}S)$. Large values of $H$ give evidence against the hypothesis of model stability. The logic of Hansen’s test is that if the model is table through the $T$ periods, then the cumulative sums in $S$ will not differ greatly from those in $F$.

#### 7.5.2 Recursive Residuals and the CUSUMS Test

A similar logic underlies an alternative test of model stability proposed by Brown, Durbin and Evans (1975) based on recursive residuals. The technique is appropriate for time-series data and might be used if one is uncertain about when a structural change might have taken place. The null hypothesis is that the coefficient vector $\beta$ is the same in every period; the alternative is simply that is not.
Suppose that the sample contains a total of $T$ observations. The $t$th recursive residual is the ex post prediction error for $y_t$ when the regression is estimated using only the first $t-1$ observations. It is also labeled a **one step ahead prediction error**;

$$e_t = y_t - x'_t b_{t-1}$$

The forecast variance of this residual is

$$\sigma^2_{ft} = \sigma^2 [1 + x'_t (X'_{t-1} X_{t-1})^{-1} x_t]$$

Let the $r$th scaled residual be

$$w_r = \frac{e_r}{\sqrt{1 + x'_t (X'_{t-1} X_{t-1})^{-1} x_t}}$$

Under the hypothesis that the coefficients remain constant during the full sample period, $w_r \sim N[0, \sigma^2]$ and is independent of $w_s$ for all $s \neq r$.

The CUSUM Test is based on the cumulated sum of the residuals:

$$W_t = \sum_{r=K+1}^{r=t} \frac{w_r}{\sigma}$$

where $\hat{\sigma}^2 = (T - K - 1)^{-1} \sum_{r=K+1}^{T} (w_r - \bar{w})^2$ and $\bar{w} = (T - K)^{-1} \sum_{r=K+1}^{T} w_r$. Under the null hypothesis, $W_t$ has a mean of zero and a variance approximately equal to the number of residuals being summed. The test is performed by plotting $W_t$ against $t$. Confidence bounds for the sum are obtained by plotting the two lines that connect the points $[K, \pm a(T - K)^{1/2}]$ and $[T, \pm 3a(T - K)^{1/2}]$. Values of $a$ that correspond to various significance levels can be found in their paper.

### 7.5.3 Unknown Timing of the Structural Break*

The testing procedures described in this section all assume that the point of the structural break is known. When this corresponds to a discrete historical even, this is a reasonable assumption. But in some applications, the timing of the break may be unknown.

We suppose that the model $E[m(y_t, x_t | \beta)] = 0$ is to be estimated by GMM using $T$ observations. The hypothesis to be investigated is as follows: Let $[\pi T] = T_1$ denote the integer part of $\pi T$ where $0 < \pi < 0$. Under the null hypothesis, the model $E[m(y_t, x_t | \beta)] = 0$ is stable for the entire sample period. Under the alternative hypothesis, the model $E[m(y_t, x_t | \beta_1)] = 0$ applies to observations $1,...,[\pi T]$ and model $E[m(y_t, x_t | \beta_2)] = 0$ applies to the remaining $T - [\pi T]$ observations.

Suppose $\pi$ were known. Then, the optimal GMM estimator for the first subsample would be obtained by minimizing with respect to the parameters $\beta_1$ the criterion function

$$q_1(\pi) = \tilde{m}_1'(\pi | \beta_1)[\text{Est.

$$Var}\sqrt{[\pi T]} \tilde{m}_1'(\pi | \beta_1)]^{-1} \tilde{m}_1(\pi | \beta_1)$$

$$= \tilde{m}_1'(\pi | \beta_1)[W_1(\pi)]^{-1} \tilde{m}_1(\pi | \beta_1)$$

where

$$m_1(\pi | \beta_1) = \frac{1}{[\pi T]} \sum_{t=1}^{[\pi T]} m_t(y_t, x_t | \beta_1)$$
The asymptotic covariance matrix will generally be computed using a first round estimator in

\[ \hat{W}_1(\pi) = \frac{1}{[\pi T]} \sum_{t=1}^{[\pi T]} m_t(\pi|\hat{\beta}_1^0)m'_t(\pi|\hat{\beta}_1^0) \]

In this time-series setting, it would be natural to accommodate serial correlation in this estimator. Following Hall and Sen (1999), the counterpart to the Newey-West (1987a) estimator would be

\[ \hat{W}_1(\pi) = \hat{W}_{1,0}(\pi) + \sum_{j=1}^{B(T)} [\hat{W}_{1,j}(\pi) + \hat{W}'_{1,j}(\pi)] \]

where \( \hat{W}_{1,j}(\pi) \) is given

\[ \hat{W}_{1,j}(\pi) = \frac{1}{[\pi T]} \sum_{t=j+1}^{[\pi T]} m_t(\pi|\hat{\beta}_1^0)m'_t(\pi|\hat{\beta}_1^0) \]

\( B(T) \) is the bandwidth, chosen to be \( O(T^{1/4}) \) and \( w_{j,T} \) is the kernel. Newey and West’s value for this is the Bartlett kernel, \([1 - j/(1 + B(T))]}\). The asymptotic covariance matrix for the GMM estimator would then be computed using

\[ \text{Est.Asy.Var}[\hat{\beta}_1] = \frac{1}{[\pi T]} \left[ \tilde{G}^{-1}_1 \hat{W}^{-1}_1(\pi) \tilde{G}_1(\pi) \right]^{-1} = \hat{V}_1 \]

where

\[ \tilde{G}_1(\pi) = \frac{1}{[\pi T]} \sum_{t=1}^{[\pi T]} \frac{\partial m_t(\pi|\hat{\beta}_1)}{\partial \hat{\beta}_1^0} \]

Estimators for the second sample are found by changing the summations to \([\pi T] + 1, ..., T\) and for the full sample by summing form 1 to \( T \).

Still assuming that \( \pi \) is known, the three standard test statistics for testing the null hypothesis of model constancy against the alternative of structural break at \([\pi T]\) would be as follows: The Wald statistic is

\[ W_T(\pi) = [\hat{\beta}_1 \pi - \hat{\beta}_2 \pi'] \{ \hat{V}_1(\pi) + \hat{V}_2(\pi) \}^{-1} [\hat{\beta}_1 \pi - \hat{\beta}_2 \pi] \]

The likelihood ratio-like statistic would be

\[ LR_T(\pi) = [q_1(\pi|\hat{\beta}_1) + q_2(\pi|\hat{\beta}_2)] - [q_1(\pi|\hat{\beta}) + q_2(\pi|\hat{\beta})] \]

where \( \hat{\beta} \) is based on the full sample.

The Lagrange multiplier statistic is the most convenient of the three. All matrices with subscript “\( T \)” are based on the full sample GMM estimator

\[ LM_T(\pi) = \frac{T}{\pi(1 - \pi)} \tilde{m}_1(\pi|\hat{\beta}_T)'\tilde{V}_T^{-1}\tilde{G}_T'[\tilde{G}_T'\tilde{V}_T^{-1}\tilde{G}_T]^{-1}\tilde{G}_T'\tilde{V}_T^{-1}\tilde{m}_1(\pi|\hat{\beta}_T) \]

The LM statistic is simpler, as it requires the model only to be estimated once, using the full sample. In each case, the statistic has a limiting chi-squared distribution with \( K \) degrees of freedom where \( K \) is the number of parameters in the model.

Since \( \pi \) is unknown, the preceding does not solve the problem posed at the outset. In principle the preceding suggests a maximum likelihood estimator of \( \pi \) if ML is used as the estimation method. However, a wide accepted solution haven’t been offered yet.
Chapter 8

Specification Analysis and Model Selection

8.1 Introduction

In this chapter we turn to some broader techniques that relate to choosing a specific model when there is more than one competing candidate.

8.2 Specification Analysis and Model Building

8.2.1 Bias Caused by Omission of Relevant Variables

Suppose that a correctly specified regression model would be

\[ y = X_1\beta_1 + X_2\beta_2 + \epsilon \]

where the two parts of \( X \) have \( K_1 \) and \( K_2 \) columns, respectively. if we regress \( y \) on \( X_1 \) without including \( X_2 \), then the estimator is

\[
b_1 = (X_1'X_1)^{-1}X_1'y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon
\]

Taking the expectation, we see that unless \( X_1'X_2 = 0 \) or \( \beta_2 = 0 \), \( b_1 \) is biased. The well known result is the omitted variable formula:

\[
E[b_1|X] = \beta_1 + P_{1,2}\beta_2
\]

where

\[
P_{1,2} = (X_1'X_1)^{-1}X_1'X_2
\]

Each column of the \( K_1 \times K_2 \) matrix \( P_{1,2} \) is the column of slopes in the least squares regression of the corresponding column of \( X_2 \) on the columns of \( X_1 \).

8.2.2 Pretest Estimation

The variance of \( b_1 \) is

\[
\text{Var}[b_1|X] = \sigma^2(X_1'X_1)^{-1}
\]

If we had computed the correct regression, including \( X_2 \), then

\[
\text{Var}[b_{1,2}|X] = \sigma^2(X_1'M_2X_1)^{-1}
\]
or
\[
\text{Var}[b_{1,2}|X] = \sigma^2[X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1}
\]

compare the covariance matrices of \(b_1\) and \(b_{1,2}\) more easily by comparing their inverses

\[
\text{Var}[b_1|X]^{-1} - \text{Var}[b_{1,2}|X]^{-1} = (1/\sigma^2)X_1'X_2(X_2'X_2)^{-1}X_2'X_1
\]

which is nonnegative definite. We conclude that although \(b_1\) is biased, its variance is never larger than that of \(b_{1,2}\).

Suppose, for instance, that \(X_1\) and \(X_2\) are each a single column and that the variables are measured as deviations from their respective means. Then

\[
\text{Var}[b_1|X] = \frac{\sigma^2}{s_{11}}, \quad \text{where } s_{11} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2
\]

whereas

\[
\text{Var}[b_{1,2}|X] = \sigma^2[x_1'x_1 - x_1'x_2(x_2'x_2)^{-1}x_2'x_1]^{-1} = \frac{\sigma^2}{s_{11}(1 - r_{12}^2)}
\]

where

\[
r_{12}^2 = \frac{(x_1'x_2)^2}{x_1'x_1x_2'x_2}
\]

is the squared sample correlation between \(x_1\) and \(x_2\). The more highly correlated \(x_1\) and \(x_2\) are, the larger is the variance of \(b_{1,2}\) compared with that of \(b_1\).

When there is a choice between two estimators \(b_1\) and \(b_2\). What researchers usually do actually creates a third estimator. It is common to include the problem variable provisionally. If its \(t\) ratio is sufficiently large, it is retained; otherwise it is discarded. This third estimator is called a pretest estimator.

### 8.2.3 Inclusion of Irrelevant Variables

If the regression model is correctly given by

\[
y = X_1\beta_1 + \epsilon
\]

The inclusion of the irrelevant variables \(X_2\) in the regression is equivalent to failing to impose \(\beta_2 = 0\) in the estimation. The least squares estimator of \(\beta\) is obviously unbiased, and \(s^2\) is also unbiased

\[
E \left[ \frac{\epsilon'\epsilon}{n - K_1 - K_2} \right] = \sigma^2
\]

In this instance, the cost is the reduced precision of the estimates. The covariance matrix in the short regression is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.

### 8.2.4 Model Building–A General to Simple Strategy

Building on the work of Hendry [e.g., (1995)] and aided by advances in estimation hardware and software, researchers are now more comfortable beginning their specification searches with large elaborate models involving many variables and perhaps long and complex lag structures. The attractive strategy is then to adopt a general-to-simple, downward reduction of the model to the preferred specification.
8.3 Choosing Between Nonnested Models

For the present, we are interested in comparing two competing linear models:

\[ H_0 : y = X\beta + \epsilon_0 \]

and

\[ H_1 : y = Z\gamma + \epsilon_1 \]

The classical procedures we have considered thus far provide no means of forming a preference for one model or the other.

8.3.1 An Encompassing Model

The encompassing approach is one in which the ability of one model to explain features of another is tested. Model 0 "encompasses" Model 1 if the features of Model 1 can be explained by Model 0 but the reverse is not true. Let \( \bar{X} \) be the set of variables in \( X \) that are not in \( Z \), define \( \bar{Z} \) likewise with respect to \( X \), and let \( W \) be the variables that the models have in common. Then \( H_0 \) and \( H_1 \) could be combined in a "supermodel":

\[ y = \bar{X}\bar{\beta} + \bar{Z}\bar{\gamma} + W\delta + \epsilon \]

In principle, \( H_1 \) is rejected if it is found that \( \bar{\gamma} = 0 \) by a conventional \( F \) test, whereas \( H_0 \) is rejected if it is found that \( \bar{\beta} = 0 \). Two problems with this approach. First, \( \delta \) remains a mixture of parts of \( \beta \) and \( \gamma \), and it is not established by the \( F \) test that either of these parts is zero. Hence, this test does not really distinguish between \( H_0 \) and \( H_1 \); it distinguishes between \( H_1 \) and a hybrid model. Second, this compound model may have an extremely large number of regressors. In a time-series setting, the problem of collinearity may be severe.

Consider an alternative approach. If \( H_0 \) is correct, then \( y \) will, apart from the random disturbance \( \epsilon \), be fully explained by \( X \). Suppose we then attempt to estimate \( \gamma \) by regression of \( y \) on \( Z \).

It is straightforward to show that the test can be carried out by using a standard \( F \) test to test the hypothesis that \( \gamma_1 = 0 \) in the augmented regression

\[ y = X\beta + Z_1\gamma_1 + \epsilon_1 \]

where \( Z_1 \) is the variables in \( Z \) that are not in \( X \).

8.3.2 Comprehensive Approach–The \( J \) Test

The underpinnings of the comprehensive approach are tied to the density function as the characterization of the data generating process. Let \( f_0(y_i|\text{data}, \beta_0) \) be the assumed density under Model 0 and define the alternative likewise as \( f_1(y_i|\text{data}, \beta_1) \). Then, a comprehensive model which subsumes both of these is

\[
f_c(y_i|\text{data}, \beta_0, \beta_1) = \frac{[f_0(y_i|\text{data}, \beta_0)]^{1-\lambda}[f_1(y_i|\text{data}, \beta_1)]^\lambda}{\int_{\text{range of } y_i} [f_0(y_i|\text{data}, \beta_0)]^{1-\lambda}[f_1(y_i|\text{data}, \beta_1)]^\lambda dy_i}
\]

Estimation of the comprehensive model followed by a test of \( \lambda = 0 \) or 1 is used to assess the validity of Model 0 or 1, respectively.
The $J$ test proposed by Davidson and MacKinnon (1982) can be shown to be an application of this principle to the linear regression model. Their suggested alternative to the preceding compound model is

$$y = (1 - \lambda)X\beta + \lambda(Z\gamma) + \epsilon$$

In this model, a test of $\lambda = 0$ would be a test against $H_1$. Davidson and MacKinnon’s $J$ test consists of estimating $\gamma$ by a least squares regression of $y$ on $Z$ followed by a least squares regression of $y$ on $X$ and $Z\hat{\gamma}$, the fitted values in the first regression. Davidson and MacKinnon show that as $n \to \infty$, if $H_1$ is true, then the probability that $\hat{\lambda}$ will differ significantly from zero approaches 1.

### 8.3.3 The Cox Test

The Cox statistics for testing the hypothesis that $X$ is the correct set of regressors and that $Z$ is not is

$$c_{01} = \frac{n}{2} \ln \left[ \frac{s_Z^2}{s_X^2 + (1/n)b'X'M_ZXb} \right] = \frac{n}{2} \ln \left[ \frac{s_Z^2}{s_{ZX}^2} \right]$$

where

- $M_Z = I - Z(Z'Z)^{-1}Z'$,
- $M_X = I - X(X'X)^{-1}X'$,
- $b = (X'X)^{-1}X' y$,
- $s_Z^2 = e_Z'e_Z/n = \text{mean-squared residual in the regression of } y \text{ on } Z$,
- $s_X^2 = e_X'e_X/n = \text{mean-squared residual in the regression of } y \text{ on } X$,
- $s_{ZX}^2 = s_X^2 + b'X'M_ZXb/n$.

The hypothesis is tested by comparing

$$q = \frac{c_{01}}{\text{Est.Var}[c_{01}]}^{1/2} = \frac{c_{01}}{\sqrt{s_Z^2/b'X'M_ZM_XM_ZXb}}$$

to the critical value from the standard normal table. A large value of $q$ is evidence against the null hypothesis ($H_0$).

#### Estimation Procedure:

1. Regress $y$ on $X$ to obtain $b$ and $\hat{y}_X = Xb$, $e_X = y - Xb$, $s_X^2 = e_X'e_X/n$.
2. Regress $y$ on $Z$ to obtain $d$ and $\hat{y}_Z = Zd$, $e_Z = y - Zd$, $s_Z^2 = e_Z'e_Z/n$.
3. Regress $\hat{y}_X$ on $Z$ to obtain $Xd$ and $e_{Z,X} = \hat{y}_X - Zd_X = M_ZXb$, $e_{Z,X}'e_{Z,X} = b'X'M_ZXb$.
4. Regress $e_{Z,X}$ on $X$ and compute residuals $e_{X,ZX}, e_{X,ZX}'e_{X,ZX} = b'X'M_ZM_XM_ZXb$.
5. Compute $s_{ZX}^2 = s_X^2 + e_{Z,X}'e_{Z,X}/n$.
6. Compute $c_{01} = \frac{n}{2} \log \frac{s_Z^2}{s_{ZX}^2}$, $v_{01} = \frac{s_X^2 (e_{X,ZX}'e_{X,ZX})}{s_{ZX}^2}, q = \frac{c_{01}}{v_{01}}$.
Therefore, the Cox statistic can be computed simply by computing a series of least squares regressions.

Pesaran and Hall (1988) have extended the Cox test to testing which of two nonnested restricted regressions is preferred. The modeling framework is

\[
H_0 : \quad y = X_0 \beta_0 + \epsilon_0, \quad \text{Var}[\epsilon_0|X_0] = \sigma_0^2 I, \quad \text{subject to } R_0 \beta_0 = q_0 \\
H_1 : \quad y = X_1 \beta_1 + \epsilon_1, \quad \text{Var}[\epsilon_1|X_1] = \sigma_1^2 I, \quad \text{subject to } R_1 \beta_1 = q_1.
\]

Let

\[
G_i = (X'_i X_i)^{-1} - (X'_i X_i)^{-1} R'_i [R_i (X'_i X_i)^{-1} R'_i]^{-1} R_i (X'_i X_i)^{-1}, \quad i = 0, 1
\]

and

\[
T_i = X_i G_i X'_i, \quad m_i = \text{rank}(R_i), \quad k_i = \text{rank}(X_i), \quad h_i = k_i - m_i \quad \text{and} \quad d_i = n - h_i \text{ where } n \text{ is the sample size.}
\]

The following steps produce the needed statistics:

1. Compute \( e_i \) = the residuals from the restricted regression, \( i = 0, 1 \).
2. Compute \( e_{10} \) by computing the residuals from the restricted regression of \( y - \epsilon_0 \) on \( X_1 \). Compute \( e_{01} \) likewise by reversing the subscripts.
3. Compute \( e_{100} \) as the residuals from the restricted regression of \( y - e_{10} \) on \( X_0 \) and \( e_{110} \) likewise by reversing the subscripts.

Let \( v_i, v_{ij} \) and \( v_{ijk} \) denote the sums of squared residuals in Steps 1, 2, and 3 and let \( s^2_i = e'_i e_i / d_i \).
4. Compute \( \text{trace } (B^2_0) = h - 1 - \text{trace}[(T_0 T_1)^2] - \{ h_1 - \text{trace}[(T_0 T_1)^2] \}^2 / (n - h_0) \) and \( \text{trace } (B^2_1) \) likewise by reversing subscripts.
5. Compute \( s^2_{10} = (v_{10} + s^2_0 \text{trace}(I - T_0 - T_1 + T_0 T_1)) \) and \( s^2_{01} \) likewise.

A Wald test based on Godfrey and Pesaran (1993) is based on the difference between an estimator of \( \sigma^2 \) and the probability limit of this estimator assuming that \( H_0 \) is true

\[
W_0 = \sqrt{n(v_1 - v_0 - v_{10})}/\sqrt{4v_0 v_{100}}.
\]

Under the null hypothesis of Model 0, the limiting distribution of \( W_0 \) is standard normal. An alternative statistic based on Cox’s likelihood approach is

\[
N_0 = (n/2) \ln(s^2_1/s^2_{10})/\sqrt{4v_{100} s^2_0/(s^2_{10})^2}
\]

### 8.4 Model Selection Criteria

As we have seen, \( R^2 \) cannot fall when variables are added to a model, so there is a built-in tendency to overfit the model. With this thought in mind, the **adjusted \( R^2 \)**

\[
\bar{R}^2 = 1 - \frac{n - 1}{n - K} (1 - \bar{R}^2) = 1 - \frac{n - 1}{n - K} \left( \frac{e' e}{\sum_{i=1}^n (y_i - \bar{y})^2} \right)
\]

has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model.

Two alternative fit measures that have seen suggested are the **Akaike information criterion**, \( AIC(K) = s^2_0 (1 - R^2) e^{2K/n} \).
and the Schwartz or Bayesian information criterion,

\[ BIC(K) = s_y^2(1 - R^2)n^{K/n} \]

(There is no degrees of freedom correction in \( s_y^2 \)). Both measures improve (decline) as \( R^2 \) increases, but, everything else constant, degrade as the model size increases. Logs are usually more convenient; the measures reported by most software are

\[ AIC(K) = \log \left( \frac{e'e}{n} \right) + \frac{2K}{n} \]
\[ BIC(K) = \log \left( \frac{e'e}{n} \right) + \frac{K \log n}{n} \]

Both prediction criteria have their virtues, and neither has an obvious advantage over the other.
Chapter 9

Nonlinear Regression Models

9.1 Nonlinear Regression Models

The general form of the nonlinear regression model is

\[ y_i = h(x_i, \beta) + \epsilon_i \]

The linear model is obviously a special case. Moreover, some models which appear to be nonlinear, such as

\[ y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} e^\epsilon \]

become linear after a transformation, in this case after taking logarithms.

9.1.1 Assumptions of the Nonlinear Regression Model

We assume that there is an underlying probability distribution, or data generating process (DGP) for the observable \( y_i \) and a true parameter vector, \( \beta \), which is a characteristic of that DGP. The following are the assumptions of the nonlinear regression model:

1. **Functional Form:** The conditional mean function for \( y_i \) given \( x_i \) is

   \[ E[y_i|x_i] = h(x_i, \beta), \quad i = 1, \ldots, n \]

   where \( h(x_i, \beta) \) is a twice continuously differentiable function.

2. **Identifiability of the model parameters:** The parameter vector in the model is identified (estimable) if there is no nonzero parameter \( \beta^0 \neq \beta \) such that \( h(x_i, \beta^0) = h(x_i, \beta) \) for all \( x_i \).

3. **Zero Mean of the Disturbance:** It follows from Assumption 1 that we may write

   \[ y_i = h(x_i, \beta) + \epsilon_i \]

   where \( E[\epsilon_i|h(x_i, \beta)] = 0 \)

4. **Homoscedasticity and Nonautocorrelation:** As in the linear model, we assume conditional homoscedasticity,

   \[ E[\epsilon_i^2|h(x_j, \beta), j = 1, \ldots, n] = \sigma^2, \quad \text{a finite constant,} \]

   and nonautocorrelation

   \[ E[\epsilon_i \epsilon_j|h(x_i, \beta), h(x_j, \beta), j = 1, \ldots, n] = 0 \quad \text{for all } j \neq i \]
Data Generating Process: The data generating process for $x_i$ is assumed to be a well behaved population such that first and second moments of the data an be assumed to converge to fixed, finite population counterparts.

Underlying Probability Model: There is a well defined probability distribution generating $\epsilon_i$.

9.1.2 The Linearized Regression

The nonlinear regression model is $y = h(x, \beta) + \epsilon$. A linear Taylor series approximation of $h(x, \beta)$ at a particular value for the parameter vector, $\beta^0$:

$$ h(x, \beta) \approx h(x, \beta^0) + \sum_{k=1}^{K} \frac{\partial h(x, \beta^0)}{\partial \beta_k^0} (\beta_k - \beta_k^0) $$

This form of the equation is called the linearized regression model. By collecting terms, we obtain

$$ h(x, \beta) \approx h(x, \beta^0) - \sum_{k=1}^{K} \beta_k^0 \left( \frac{\partial h(x, \beta^0)}{\partial \beta_k^0} \right) + \sum_{k=1}^{K} \beta_k \left( \frac{\partial h(x, \beta^0)}{\partial \beta_k^0} \right) $$

Let $x_k^0$ equal the $k$th partial derivative, $\partial h(x, \beta^0)/\partial \beta_k^0$. Then,

$$ h(x, \beta) \approx \left[ h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0 \right] + \sum_{k=1}^{K} x_k^0 \beta_k $$

which may be written

$$ h(x, \beta) \approx h^0 - x^0 \beta^0 + x^0 \beta $$

which implies that

$$ y \approx h^0 - x^0 \beta^0 + x^0 \beta + \epsilon $$

By placing the known terms on the left-hand side of the equation, we obtain a linear equation:

$$ y^0 = y - h^0 + x^0 \beta^0 = x^0 \beta + \epsilon^0 $$

With a value of $\beta^0$ in hand, we could compute $y^0$ and $x^0$ and then estimate the parameters of above function by linear least squares.

9.1.3 Large Sample Properties of the Nonlinear Least Squares Estimator

The derivatives of the regression function, which are called the pseudoregressors in the linearized model when they are computed at the true parameter values. For the nonlinear regression model, we assume that

$$ \operatorname{plim} \frac{1}{n} X^0 X^0 = \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{h(x_i, \beta^0)}{\partial \beta_0} \right) \left( \frac{h(x_i, \beta^0)}{\partial \beta_0'} \right) = Q^0 $$

where $Q^0$ is a positive definite matrix. Also,

$$ \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} x_i^0 \epsilon_i = 0 $$
Finally, asymptotic normality can be established under general conditions if
\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i^0 \epsilon_i \xrightarrow{d} N[0, \sigma^2 Q^0] \]
The nonlinear least squares criterion function is
\[ S(b) = \frac{1}{2} \sum_{i=1}^{n} \left( y_i - h(x_i, b) \right)^2 = \frac{1}{2} \sum_{i=1}^{n} e_i^2 \]
The values of the parameters that minimize the sum of squared deviations are the nonlinear least squares estimators.

**THEOREM** Consistency of the Nonlinear Least Squares Estimator
If the following assumptions hold:
1. The parameter space containing \( \beta \) is compact
2. For any vector \( \beta^0 \) in that parameter space, \( \text{plim}(1/n)S(\beta^0) = q(\beta^0) \), a continuous and differentiable function.
3. \( q(\beta^0) \) has a unique minimum at the true parameter vector, \( \beta \),

Then, the nonlinear least squares estimator is consistent.

**THEOREM** Asymptotic Normality of the Nonlinear Least Squares Estimator
If the pseudoregressors are well behaved, then
\[ b \overset{a}{\sim} N\left[ \beta, \frac{\sigma^2}{n} \sqrt{Q^0 - 1} \right] \]
where
\[ Q^0 = \text{plim} \frac{1}{n} X^0 X^0 \]
The sample estimate of the asymptotic covariance matrix is
\[ \text{Est. Asy. Var}[b] = \hat{\sigma}^2 (X^0 X^0)^{-1} \]

9.1.4 Computing the Nonlinear Least Squares Estimator
Minimizing the sum of squares is a standard problem in nonlinear optimization that can be solved by a number of methods. The method of Gauss-Newton is often used. In the linearized regression model, if a value of \( \beta^0 \) is available, then the linear regression model can be estimated by linear least squares. Once a parameter vector is obtained, it can play the role of a new \( \beta^0 \), and the computation can be done again. The iteration can continue until the difference between successive parameter vectors is small enough to assume convergence.

This iterative solution to the minimization problem is
\[ b_{t+1} = b_t + \left( \sum_{i=1}^{n} x_i^0 x_i^0 \right)^{-1} \left( \sum_{i=1}^{n} x_i^0 (y_i - h_i^0 + x_i^0 b_t) \right) \]
\[ = b_t + (X^0 X^0)^{-1} (X^0 X^0 - 1) e^0 \]
\[ = b_t + \Delta_t \]
where all terms on the right-hand side are evaluated at \( b \) and \( e^0 \) is the vector of nonlinear least squares residuals.

A consistent estimator of \( \sigma^2 \) is based on the residuals:

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - h(x_i, b)]^2
\]

A degrees of freedom correction, \( 1/(n - K) \), where \( K \) is the number of elements in \( \beta \), is NOT strictly necessary here, because all results are asymptotic in any event. Davidson and MacKinnon (1993) argue that on average, the function above will underestimate \( \sigma^2 \), and one should use the degrees of freedom correction.

Once the nonlinear least squares estimates are in hand, inference and hypothesis tests can proceed in the same fashion as prescribed in Chapter 7. A minor problem arises in evaluating the fit of the regression in that the familiar measure,

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

is no longer guaranteed to be in the range of 0 to 1.

9.2 Hypothesis Testing and Parametric Restrictions

9.2.1 Significance Tests for Restrictions: \( F \) and Wald Statistics

The hypothesis to be tested is

\[
H_0 : r(\beta) = q
\]

where \( r(\beta) \) is a column vector of \( J \) continuous functions of the elements of \( \beta \). It is necessary, however, that they be overidentifying restrictions.

Let \( b \) be the unrestricted, nonlinear least squares estimator, and let \( b^* \) be the estimator obtained when the constraints of the hypothesis are imposed.

The nonlinear analog to the familiar \( F \) statistic based on the fit of the regression would be

\[
F[J, n - K] = \frac{[S(b^*) - S(b)]/J}{S(b)/(n - K)}
\]

This equation has the appearance of our earlier \( F \) ratio. In the nonlinear setting, however, neither the numerator nor the denominator has exactly the necessary chi-squared distribution, so the \( F \) distribution is only approximate.

The Wald test is based on the distance between \( r(b) \) and \( q \). If the unrestricted estimates fail to satisfy the restrictions, then doubt is cast on the validity of the restrictions. The statistic is

\[
W = [r(b) - q]'\{\text{Est. Asy. Var}[r(b) - q]\}^{-1}[r(b) - q]
\]

\[
= [r(b) - q]'\{R(b)\hat{V}R'(b)\}^{-1}[r(b) - q]
\]

where

\[
\hat{V} = \text{Est. Asy. Var}[b]
\]

and \( R(b) \) is evaluated at \( b \), the estimate of \( \beta \). Under the null hypothesis, this statistic has a limiting chi-squared distribution with \( J \) degrees of freedom.
9.2.2 Tests Based on the LM Statistic

The Lagrange Multiplier Test is based on the decrease in the sum of squared residuals that would result if the restrictions in the restricted model were released. Let \( e_* \) be the vector of residuals \( y_i - h(x_i, b_*) \) computed using the restricted estimates. Recall that we defined \( X^0 \) as an \( n \times K \) matrix of derivatives computed at a particular parameter vector. Let \( X^*_0 \) be this matrix computed at the restricted estimates. Then the Lagrange multiplier statistic for the nonlinear regression model is

\[
LM = \frac{e'_* X^*_0 [X^*_0 X^0]^{-1} X^0 e_*}{e'_* e_*/n}
\]

Under \( H_0 \), this statistic has a limiting chi-squared distribution with \( J \) degrees of freedom.

9.2.3 A Specification Test for Nonlinear Regressions: The \( PE \) Test

MacKinnon, White, and Davidson (1983) have extended the \( J \) test to nonlinear regressions. The specific hypothesis to be tested is

\[
H_0: y = h^0(x, \beta) + \epsilon_0
\]

versus

\[
H_1: g(y) = h^1(z, \gamma) + \epsilon_1
\]

where \( x \) and \( z \) are regressor vectors and \( \beta \) and \( \gamma \) are the parameters. We form the compound model

\[
y = (1 - \alpha)h^0(x, \beta) + \alpha h^1(z, \gamma) + \epsilon
\]

where \( \alpha \) is the parameter of interest.

Presumably, both \( \beta \) and \( \gamma \) could be estimated in isolation by nonlinear least squares. Suppose that a nonlinear least squares estimate of \( \gamma \) has been obtained. One approach is to insert this estimate in equation above and then estimate \( \gamma \) and \( \alpha \) by nonlinear least squares. The \( J \) test amounts to testing the hypothesis that \( \alpha \) equals zero.

Davidson and MacKinnon (1981) propose what may be a simpler alternative. Given an estimate of \( \beta \) say \( \hat{\beta} \), approximate the first \( h^0(x, \beta) \) with a linear Taylor series at this point,

\[
h^0(x, \beta) \approx h^0(x, \hat{\beta}) + \left[ \frac{\partial h^0(\cdot)}{\partial \beta} \right](\beta - \hat{\beta}) = \hat{h}^0 + \hat{H}^0 \beta - \hat{H}^0 \hat{\beta}
\]

Using this device,

\[
y - \hat{h}^0 + \hat{H}^0 \beta = \hat{H}^0 \beta + \alpha [h^1(z, \hat{\gamma}) - h^0(x, \hat{\beta})] + \epsilon
\]

in which \( \beta \) and \( \alpha \) can be estimated by linear least squares. If it is found that \( \hat{\alpha} \) is significantly different from zero then \( H_0 \) is rejected.

Now we can generalized the test to allow a nonlinear function, \( g(y) \), in \( H_1 \). Davidson and MacKinnon require \( g(y) \) to be monotonic, continuous, and continuously differentiable and not to introduce any new parameters. The compound model that forms the basis of the test is

\[
(1 - \alpha)[y - h^0(x, \beta)] + \alpha [g(y) - h^1(z, \gamma)] = \epsilon
\]
Use the same linear Taylor series expansion for \( h^0(x, \beta) \) on the left-hand side and replace both \( y \) and \( h^0(x, \beta) \) with \( \hat{h}^0 \) on the right. The resulting model is

\[
y - \hat{h}^0 + \hat{H}^0 \beta = \hat{H}^0 \beta + \alpha[\hat{h}_1 - g(\hat{h}^0)] + \epsilon
\]

As before, with an initial estimate of \( \beta \), this model can be estimated by least squares. This modified form of the \( J \) test is labeled the \( P_E \) test\(^1\).

### 9.3 Alternative Estimators for Nonlinear Regression Models

In this section, we will consider two extensions of these results. First, as in the linear case, there can be situations in which the assumption that \( \text{Cov}[x_i, \epsilon_i] = 0 \) is not reasonable. Second, there will be models in which it is convenient to estimate the parameters in two steps, estimating one subset at the first step and then using these estimates in a second step at which the remaining parameters are estimated.

#### 9.3.1 Nonlinear Instrumental Variables Estimation

In the nonlinear model

\[
y_i = h(x_i, \beta) + \epsilon_i
\]

the covariates \( x_i \) may be correlated with the disturbances. We would expect this effect to be transmitted to the pseudoregressors, \( x_i^0 = \partial h(x_i, \beta) / \partial \beta \). Suppose that there is a set of variables \( \{z_1, \ldots, z_L\} \) such that

\[
\text{plim}(1/n) Z' \epsilon = 0
\]

and

\[
\text{plim}(1/n) Z' X^0 = Q_{ZX}^0 \neq 0
\]

where \( X^0 \) is the matrix of pseudoregressors in the linearized regression, evaluated at the true parameter values. The linearized regression model is given

\[
y = h(X, \beta) + \epsilon \approx \hat{h}^0 + X^0(\beta - \beta^0) + \epsilon
\]

or

\[
y^0 \approx X^0 \beta + \epsilon
\]

where

\[
y^0 = y - \hat{h}^0 + X^0 \beta^0
\]

We have assumed that

\[
\text{plim}(1/n) Z' y^0 = \text{plim}(1/n) Z' X^0 \beta
\]

Suppose, as did before, that there are the same number of instrumental variables as there are parameters, that is, columns in \( X^0 \). Then the estimator used before is suggested:

\[
b_{IV} = (Z' X^0)^{-1} Z' y^0
\]

The logic is sound, but there is a problem with this estimator. The unknown parameter vector \( \beta \) appears on both sides of \( \text{plim}(1/n) Z' y^0 = \text{plim}(1/n) Z' X^0 \beta \). We might consider

\(^1\)As the author discuss, it is probably not as powerful as any of the Wald or Lagrange multiplier tests that we have considered. In their experience, however, it has sufficient power for applied research and is clearly simple to carry out.
the approach that with some initial estimator in hand, iterate back and forth between the
instrumental variables regression and recomputing the pseudoregressors until the process
converges to the fixed point that we seek.

We will consider the problem in general terms. The estimation criterion for nonlinear
instrumental variables is a quadratic form,

\[
\text{Min}_\beta S(\beta) = \frac{1}{2} \left\{ [y - h(X,\beta)]' Z ) (Z'Z)^{-1} \left\{ Z'[y - h(X,\beta)] \right\} \right. \\
= \frac{1}{2} \epsilon(\beta)' Z(Z'Z)^{-1} Z' \epsilon(\beta)
\]

The first-order conditions for minimization of this weighted sum of squares are

\[
\frac{\partial S(\beta)}{\partial \beta} = -X_0' Z(Z'Z)^{-1} Z' \epsilon(\beta) = 0
\]

This result is the same one we had for the linear model with \(X^0\) in the role of \(X\).

**THEOREM**  **Asymptotic Distribution of the Nonlinear Instrumental Variables Estimator**

With well behaved instrumental variables and pseudoregressors,

\[
b_{IV} \overset{d}{\sim} N[\beta, \sigma^2 (Q_{XX} Z (Q_{ZZ})^{-1} Q_{ZX}^{-1})^{-1}]
\]

We estimate the asymptotic covariance matrix with

\[
\text{Est. Asy. Var}[b_{IV}] = \hat{\sigma}^2 [\hat{X}^0' Z (Z'Z)^{-1} Z' \hat{X}^0]^{-1}
\]

where \(\hat{X}^0\) is \(X^0\) computed using \(b_{IV}\).

**9.3.2 Two-Step Nonlinear Least Squares Estimation**

In this section, we consider a special case of this general class of models in which the non-
linear regression model depends on a second set of parameters that is estimated separately.

The model is

\[
y = h(x,\beta,w,\gamma) + \epsilon
\]

We consider cases in which the auxiliary parameter \(\gamma\) is estimated separately in a model
that depends on an additional set of variables \(w\). The estimation procedure is as follows.

1. Estimate \(\gamma\) by least squares, nonlinear least squares, or maximum likelihood.

2. Estimate \(\beta\) by nonlinear least squares regression of \(y\) on \(h(x,\beta,w,c)\). Let \(\sigma^2 V_b\) by
the asymptotic covariance matrix of this estimator of \(\beta\), assuming \(\gamma\) is known and let
\(s^2 \hat{V}_b\) be any appropriate estimator of \(\sigma^2 V_b = \sigma^2 (X^0'X^0)^{-1}\), where \(X^0\) is the matrix of
pseudoregressors evaluated at the true parameter values \(x_i^0 = \partial h(x_i,\beta,w_i,\gamma)/\partial \beta\).

The asymptotic covariance matrix for the two-step estimator is provided by the following
theorem.
THEOREM Asymptotic Distribution of the Two-Step Nonlinear Least Squares Estimator [Murphy and Topel (1985)]

Under the standard conditions assumed for the nonlinear least squares estimator, the second-step estimator of $\beta$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$V_b^* = \sigma^2 V_b + V_b [CV_c V' - CV_c R' - RV_c C'] V_b$$

where

$$C = n \text{plim} \frac{1}{n} \sum_{i=1}^{n} x_i^0 \epsilon_i^2 \left( \frac{\partial h(x_i, \beta, w_i, \gamma)}{\partial \gamma} \right)$$

and

$$R = n \text{plim} \frac{1}{n} \sum_{i=1}^{n} x_i^0 \epsilon_i \left( \frac{\partial g(w_i, \gamma)}{\partial \gamma} \right)$$

The function $\partial g(\cdot)/\partial \gamma$ in the definition of $R$ is the gradient of the $i$th term in the log-likelihood function if $\gamma$ is estimated by maximum likelihood. If $\gamma$ appears as the parameter vector in a regression model,

$$z_i = f(w_i, \gamma) + u_i$$

then $\partial g(\cdot)/\partial \gamma$ will be a derivative of the sum of squared deviations function,

$$\frac{\partial g(\cdot)}{\partial \gamma} = u_i \frac{\partial f(w_i, \gamma)}{\partial \gamma}$$

If this is a linear regression, then the derivative vector is just $w_i$. 

49
Chapter 10

Nonspherical Disturbances–The Generalized Regression Model

10.1 Introduction

The generalized linear regression model is

\[ y = X\beta + \epsilon, \]

\[ E[\epsilon|X] = 0, \]

\[ E[\epsilon\epsilon'|X] = \sigma^2 \Omega = \Sigma \]

where \( \Omega \) is a positive definite matrix. For present purposes, we will retain the linear specification and refer to our model simply as the generalized regression model. Two cases we will consider in detail are heteroscedasticity and autocorrelation. Disturbances are heteroscedastic when they have different variances. Autocorrelation is usually found in time-series data. Economic time series often display a “memory” in that variation around the regression function is not independent from one period to the next. Panel Data sets, consisting of cross sections observed at several points in time, may exhibit both characteristics.

10.2 Least Squares and Instrumental Variables Estimation

10.2.1 Finite-Sample Properties of Ordinary Least Squares

**THEOREM** Finite Sample Properties of \( b \) in the Generalized Regression Model

If the regressors and disturbances are uncorrelated, then the unbiasedness of least squares is unaffected by violations of assumption \( E[\epsilon\epsilon'|X] = \sigma^2 I \). The least squares estimator is unbiased in the generalized regression model. With nonstochastic regressors, or conditional on \( X \), the sampling variance of the least
The least squares estimator is

\[
\text{Var}[b | X] = E[(b - \beta)(b - \beta)' | X] \\
= E[(X'X)^{-1}X'e'e'X(X'X)^{-1} | X] \\
= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\
= \frac{\sigma^2}{n} \left( \frac{1}{n}X'X \right)^{-1} \left( \frac{1}{n}X'\Omega X \right) \left( \frac{1}{n}X'X \right)^{-1}
\]

If the regressors are stochastic, then the unconditional variance is \(E_X[\text{Var}[b | X]]\). 

\(b\) is a linear function of \(\epsilon\). Therefore, if \(\epsilon\) is normally distributed, then

\(b | X \sim N[\beta, \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}]\)

### 10.2.2 Asymptotic Properties of Least Squares

If \(\text{Var}[b | X]\) converges to zero, then \(b\) is mean square consistent. With well-behaved regressors, \((X'X/n)^{-1}\) will converge to a constant matrix.

\[
\frac{\sigma^2}{n} \left( \frac{X'\Omega X}{n} \right) = \left( \frac{\sigma^2}{n} \right) \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{ij} x_i x_j' \right)
\]

we see that though the leading constant will, by itself, converge to zero, the matrix is a sum of \(n^2\) terms, divided by \(n\). Thus, the product is a scalar that is \(O(1/n)\) times a matrix that is, at least at this juncture, \(O(n)\), which is \(O(1)\).

**THEOREM** Consistency of OLS in the Generalized Regression Model

If \(Q = \text{plim}(X'X/n)\) and \(\text{plim}(X'\Omega X/n)\) are both finite positive definite matrices, then \(b\) is consistent for \(\beta\). Under the assumed conditions,

\[
\text{plimb} = \beta
\]

Ordinary least squares is consistent in the generalized regression model if:

1. The smallest characteristic root of \(X'X\) increases without bound as \(n \to \infty\), which implies that \(\text{plim}(X'X)^{-1} = 0\). If the regressors satisfy the Grenander conditions G1 through G3 of Section 5.2, then they will meet this requirement.

2. The largest characteristic root of \(\Omega\) is finite for all \(n\). For the heteroscedastic model, the variances are the characteristic roots, which requires them to be finite. For models with autocorrelation, the requirements are that the elements of \(\Omega\) be finite and that the off-diagonal elements not be too large relative to the diagonal elements. We will examine this condition at several points below.

The least squares estimator is asymptotically normally distributed if the limiting distribution of

\[
\sqrt{n}(b - \beta) = \left( \frac{X'X}{n} \right)^{-1} \frac{1}{\sqrt{n}} X'\epsilon
\]

51
is normal. If \( \text{plim}(X'X/n) = Q \), then the limiting distribution of the right-hand side is the same as that of

\[
v_{n,LS} = Q^{-1} \frac{1}{\sqrt{n}} X' \epsilon = Q^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \epsilon_i
\]

where \( x'_i \) is a row of \( X \). The exact variance of the sum is

\[
E_x \left[ \text{Var} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \epsilon_i \right] \right] = \frac{\sigma^2}{n} \sum_{i=1}^{n} \omega_i Q_i
\]

which, for our purposes, we would require to converge to a positive definite matrix.

In the heteroscedastic case, if the variances of \( \epsilon_i \) are finite and are not dominated by any single term, so that the conditions of the Lindberg-Feller central limit theorem apply to \( v_{n,LS} \), then the least squares estimator is asymptotically normally distributed with covariance matrix

\[
\text{Asy. Var}[b] = \frac{\sigma^2}{n} Q^{-1} \text{plim} \left( \frac{1}{n} X' \Omega X \right) Q^{-1}
\]

**THEOREM Asymptotic Distribution of \( b \) in the GR Model**

If the regressors are sufficiently well behaved and the off-diagonal terms in \( \Omega \) diminish sufficiently rapidly, then the least squares estimator is asymptotically normally distributed with mean \( \beta \) and covariance matrix given as

\[
\text{Asy. Var}[b] = \frac{\sigma^2}{n} Q^{-1} \text{plim} \left( \frac{1}{n} X' \Omega X \right) Q^{-1}
\]

### 10.2.3 Asymptotic Properties of the Instrumental Variables Estimator

In the classical model, we constructed an estimator around a set of variables \( Z \) that were uncorrelated with \( \epsilon \),

\[
b_{IV} = (X'Z)(Z'Z)^{-1}X'y
\]

\[
= \beta + [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z' \epsilon
\]

Suppose that \( X \) and \( Z \) are well behaved

\[
\text{plim}(1/n)Z'Z = Q_{ZZ}, \text{ a positive definite matrix.}
\]

\[
\text{plim}(1/n)Z'X = Q_{ZZ} = Q_{XZ}', \text{ a nonzero matrix.}
\]

\[
\text{plim}(1/n)X'X = Q_{XX}, \text{ a positive definite matrix.}
\]

To avoid a string of matrix computations that may not fit on a single line, for convenience let

\[
Q_{XX.Z} = [Q_{xz}Q_{zz}^{-1}Q_{xz}]^{-1}Q_{xz}Q_{zz}^{-1}
\]

\[
= \text{plim} \left[ \left( \frac{1}{n} X'Z \right) \left( \frac{1}{n} Z'Z \right)^{-1} \left( \frac{1}{n} Z'X \right) \right]^{-1} \left( \frac{1}{n} X'Z \right) \left( \frac{1}{n} Z'Z \right)^{-1}
\]

then,

\[
\text{plimb}_{IV} = \beta + Q_{XX.Z} \text{plim} \left( \frac{1}{n} Z' \epsilon \right) = \beta
\]

52
The large sample behavior of $b_{IV}$ depends on the behavior of

$$v_{n,IV} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \epsilon_i$$

We will once again rely on the results of Anderson (1971) or Amemiya (1985) that under very general conditions,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \epsilon_i \overset{d}{\to} N \left[ 0, \sigma^2 \text{plim} \left( \frac{1}{n} Z' \Omega Z \right) \right]$$

**THEOREM** Asymptotic Distribution of the IV Estimator in the Generalized Regression Model

If the regressors and the instrumental variables are well behaved in the fashions discussed above, then

$$b_{IV} \sim N[\beta, V_{IV}]$$

where

$$V_{IV} = \frac{\sigma^2}{n} (Q_{XX,Z}) \text{plim} \left( \frac{1}{n} Z' \Omega Z \right) (Q'_{XX,Z})$$

### 10.3 Robust Estimation of Asymptotic Covariance Matrices

What is required is an estimator of the $K(K + 1)/2$ unknown elements in the matrix.

$$\text{plim} Q_* = \text{plim} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x'_j$$

The least squares estimator $b$ is a consistent estimator of $\beta$, which implies that the least squares residuals $\epsilon_i$ are “pointwise” consistent estimators of their population counterparts $\epsilon_i$. The general approach, then, will be to use $X$ and $e$ to devise an estimator of $Q_*$. Consider the heteroscedasticity case first. We seek an estimator of

$$Q_* = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 x_i x'_i$$

White (1980) has shown that under very general conditions, the estimator

$$S_0 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 x_i x'_i$$

has

$$\text{plim} S_0 = \text{plim} Q_*$$

With some fairly mild assumptions about $x_i$, then, we could invoke a law of large numbers to state that of $Q_*$ has a probability limit, then

$$\text{plim} Q_* = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 x_i x'_i = \text{plim} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 x_i x'_i$$
The final detail is to justify the replacement of $\epsilon_i$ with $e_i$ in $S_0$. The end result is that the **White heteroscedasticity consistent estimator**

\[
\text{Est. Asy. Var}[b] = \frac{1}{n} \left( \frac{1}{n} X'X \right)^{-1} \left( \sum_{i=1}^{n} e_i^2 x_i x_i' \right) \left( \frac{1}{n} X'X \right)^{-1}
\]

\[
= n(X'X)^{-1} S_0 (X'X)^{-1}
\]

can be used to estimate the asymptotic covariance matrix of $b$.

The extension of White’s result to the more general case of autocorrelation is much more difficult. The natural counterpart for estimating

\[
Q_* = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j'
\]

would be

\[
\hat{Q}_* = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} e_i e_j x_i x_j'
\]

Unlike the heteroscedasticity case, the matrix above is $1/n$ times a sum of $n^2$ terms, so it is difficult to conclude yet that it will converge to anything at all. A sufficient condition is that terms with subscript pairs $|i - j|$ grow smaller as the distance between them grows larger.

The practical problem is that $\hat{Q}_*$ need not be positive definite. Newey and West (1987a) have devised an estimator that overcomes this difficulty:

\[
\hat{Q}_* = S_0 + \frac{1}{n} \sum_{l=1}^{L} \sum_{t=l+1}^{n} w_l e_t e_{t-l} (x_t x_{t-l} + x_{t-l} x_t'),
\]

\[
w_l = 1 - \frac{l}{(L+1)}
\]

The **Newey-West autocorrelation consistent covariance estimator** is surprisingly simple and relatively easy to implement. Current practice specifies $L \approx T^{1/4}$

### 10.4 Generalized Method of Moments Estimation

At this point, we no longer require that $E[\epsilon_i|x_i] = 0$. Our model is

\[
y_i = x_i' \beta + \epsilon_i
\]

where $E[\epsilon_i|z_i] = 0$ for $K$ variables in $x_i$ and for some set of $L$ instrumental variables, $z_i$, where $L > K$. In particular, we will consider three cases:

- Classical regression: Var$[\epsilon_i|X,Z] = \sigma^2$.
- Heteroscedasticity: Var$[\epsilon_i|X,Z] = \sigma_i^2$.
- Generalized model: Cov$[\epsilon_t, \epsilon_s|X,Z] = \sigma^2 \omega_{ts}$.
where \( Z \) and \( X \) are the \( n \times L \) and \( n \times K \) observed data matrices.

The assumption \( E[\epsilon_i | z_i] = 0 \) implies the following orthogonality condition:

\[
\text{Cov}[z_i, \epsilon_i] = 0, \quad \text{or} \quad E[z_i(y_i - x_i' \beta)] = 0
\]

By summing the terms, we find that this further implies the population moment equation,

\[
E \left[ \frac{1}{n} \sum_{i=1}^{n} z_i(y_i - x_i' \beta) \right] = E[\bar{m}(\beta)] = 0
\]

Its empirical moment equation is

\[
\left[ \frac{1}{n} \sum_{i=1}^{n} z_i(y_i - x_i' \hat{\beta}) \right] = \left[ \frac{1}{n} \sum_{i=1}^{n} m_i(\hat{\beta}) \right] = \bar{m}(\hat{\beta}) = 0
\]

In the absence of other information about the data generating process, we can use the empirical moment equation as the basis of our estimation strategy.

The empirical moment condition is \( L \) equations in \( K \) unknowns. There are three possibilities to consider:

1. Underidentified: \( L < K \). If there are fewer moment equations than there are parameters, then it will not be possible to find a solution to the equation system above.

2. Exactly identified. If \( L = K \), then you can easily show that the single solution to our equation system is the familiar instrumental variables estimator,

\[
\hat{\beta} = (Z'X)^{-1}Z'y
\]

3. Overidentified. If \( L > K \), then there is no unique solution to the equation system \( \bar{m}(\hat{\beta}) = 0 \). In this instance, that would mean choosing the estimator based on the criterion function

\[
\text{Min}_{\beta} q = \bar{m}(\hat{\beta})\bar{m}(\hat{\beta})
\]

Assumption GMM1. Convergence of the moments. The population moment converges in probability to its population counterpart. That is, \( \bar{m}(\beta) \to 0 \).

Assumption GMM2. Identification. The parameters are identified in terms of the moment equations.

Assumption GMM3. Limiting Normal Distribution for the Sample Moments. The population moment obeys a central limit theorem or some similar variant.

For the particular model we are studying here,

\[
\bar{m}(\beta) = \frac{1}{n}(Z' y - Z' X \beta)
\]

\[
G(\beta) = \frac{1}{n}Z' X
\]

\[
\Gamma(\beta) = Q Z X
\]
The remaining detail, which is the crucial one for the model we are examining, is for us to determine,

$$V = \text{Asy.Var}\left[\sqrt{n}\hat{m}(\beta)\right]$$

Given the form of \(\hat{m}(\beta)\),

$$V = \frac{1}{n}\text{Var}\left[\sum_{i=1}^{n} z_i e_i\right] = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma^2 \omega_{ij} z_i z_j' = \sigma^2 \frac{Z'\Omega Z}{n}$$

The estimators of \(V\) for our three cases will be

- **Classical Regression:**
  $$\hat{V} = \frac{(e'e/n)}{n} \sum_{i=1}^{n} z_i z_i' = \frac{(e'e/n)}{n} Z'Z$$

- **Heteroscedastic:**
  $$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 z_i z_i'$$

- **General:**
  $$\hat{V} = \frac{1}{n} \left[ \sum_{i=1}^{n} e_i^2 z_i z_i' + \sum_{l=1}^{L} \sum_{t=l+1}^{n} \left(1 - \frac{l}{(L+1)}\right) e_t e_{t-l}(z_t z_{t-l} + z_{t-l} z_t') \right]$$

Collecting all the terms so far, then, we have

$$\text{Est. Asy.Var}[\hat{\beta}] = \frac{1}{n} \frac{[\hat{G}(\hat{\beta})']^{-1} \hat{G}(\hat{\beta})' \hat{V} \hat{G}(\hat{\beta}) [\hat{G}(\hat{\beta})']^{-1}}{n([X'Z](Z'X))^{-1}(X'Z)\hat{V}(Z'X)(X'Z)(Z'X)^{-1}}$$

The class of minimum distance estimators is defined by the solutions to the criterion function

$$\text{Min}_{\beta q} = \hat{m}(\beta)' W \hat{m}(\beta)$$

where \(W\) is any positive definite weighting matrix.

**THEOREM Minimum Distance Estimators**

If \(\text{plim}\hat{m}(\beta) = 0\) and if \(W\) is a positive definite matrix, then \(\text{plim}\hat{\beta} = \text{Argmin}\{q = \hat{m}(\beta)' W \hat{m}(\beta)\} = \beta\). The minimum distance estimator is consistent. It is also asymptotically normally distributed and has asymptotic covariance matrix.

$$\text{Asy.Var}[\hat{\beta}_{MD}] = \frac{1}{n} [\hat{G}'W\hat{G}]^{-1} \hat{G}'WVW\hat{G} [\hat{G}'W\hat{G}]^{-1}$$

**THEOREM Generalized Method of Moments Estimators**

The minimum distance estimator obtained by using \(W = V^{-1}\) is the generalized method of moments, or GMM estimator. The GMM estimator is consistent, asymptotically normally distributed, and has asymptotic covariance matrix equal to

$$\text{Asy. Var}[\hat{\beta}_{GMM}] = \frac{1}{n} [\hat{G}'V^{-1}\hat{G}]^{-1}$$
For the generalized regression model, these are

\[ \hat{\beta}_{GMM} = \left( \left( X^T Z \right) \hat{V}^{-1} \left( Z^T X \right) \right)^{-1} \left( X^T Z \right) \hat{V}^{-1} \left( Z^T y \right) \]

and

\[ \text{Asy. Var}[\hat{\beta}_{GMM}] = \left[ \left( X^T Z \right) \hat{V} \left( Z^T X \right) \right]^{-1} \]

The process of GMM estimation will have to proceed in two steps:

**Step 1.** Use \( W = I \) to obtain a consistent estimator of \( \beta \). Then, estimate \( V \) with

\[ \hat{V} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 z_i \]

**Step 2.** Use \( W = \hat{V}^{-1} \) to compute the GMM estimator.

### 10.5 Efficient Estimation by Generalized Least Squares

#### 10.5.1 Generalized Least Squares (GLS)

Since \( \Omega \) is a positive definite symmetric matrix, it can be factored into

\[ \Omega = C \Lambda C' \]

where the columns of \( C \) are the characteristic vectors of \( \Omega \) and the characteristic roots of \( \Omega \) are arrayed in the diagonal matrix \( \Lambda \). Let \( \Lambda^{1/2} \) be the diagonal matrix with \( \hat{i} \)th diagonal element \( \sqrt{\lambda_i} \), and let \( T = C \Lambda^{1/2} \). Then \( \Omega = T T' \). Also, let \( P' = C \Lambda^{-1/2} \), so \( \Omega^{-1} = P' P \).

Premultiply the model by \( P \) to obtain,

\[ Py = PX \beta + P \epsilon \]

or

\[ y* = X* \beta + \epsilon* \]

The variance of \( \epsilon* \) is

\[ E[\epsilon* \epsilon'_*] = P \sigma^2 \Omega \]

In the classical model, ordinary least squares is efficient; hence,

\[ \hat{\beta} = (X'_* X'_*)^{-1} X'_* y* \]

\[ = (X' P' P^{-1} X') P' P y \]

\[ = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \]

is the efficient estimator of \( \beta \). This estimator is the generalized least squares (GLS).

**THEOREM Properties of the Generalized Least Squares Estimator**

If \( E[\epsilon* | X*] = 0 \), then

\[ E[\hat{\beta} | X*] = E[(X'_* X'_*)^{-1} X'_* y* | X*] = \beta + E[(X'_* X'_*)^{-1} X'_* \epsilon* | X*] = \beta \]

The GLS estimator \( \hat{\beta} \) is unbiased. This result is equivalent to \( E[P \epsilon | PX] = 0 \), but since \( P \) is a matrix of known constants, we return to the familiar requirement \( E[\epsilon | X] = 0 \). The requirement that the regressors and disturbances be uncorrelated is unchanged.
The GLS estimator is consistent if \( \text{plim}(1/n)X'X = Q_* \), where \( Q_* \) is a finite positive definite matrix. Making the substitution, we see that this implies

\[
\text{plim}[(1/n)X'\Omega^{-1}X]^{-1} = Q_*^{-1}
\]

We require the transformed data \( X_* = PX \), not the original data \( X \), to be well behaved. Under the assumption, the following hold:

The GLS estimator is asymptotically normally distributed, with mean \( \beta \) and sampling variance

\[
\text{Var}[\hat{\beta}|X_*] = \sigma^2(X'_*X_*)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}
\]

The GLS estimator \( \hat{\beta} \) is the minimum variance linear unbiased estimator in the generalize regression model.

For testing the \( J \) linear restrictions, \( R\beta = q \), the appropriate statistic is

\[
F[J, n-K] = \frac{(R\hat{\beta} - q)'[R\hat{\sigma}^2(X'_*X_*)^{-1}R']^{-1}(R\hat{\beta} - q)}{\hat{\sigma}^2}/J
\]

where the residual vector is

\[\hat{\epsilon} = y_* - X_*\hat{\beta}\]

and

\[\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-K} = \frac{(y - X\hat{\beta})'(\Omega^{-1}(y - X\hat{\beta}))}{n-K}\]

The constrained GLS residuals, \( \hat{\epsilon}_c = y_* - X_*\hat{\beta}_c \), are based on

\[\hat{\beta}_c = \hat{\beta} - [X'\Omega^{-1}X]^{-1}R'[R(\Omega^{-1}X)^{-1}R']^{-1}(R\hat{\beta} - q)\]

A more appealing fit measure might be based on the residuals from the original model once the GLS estimator is in hand, such as

\[R^2_G = 1 - \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\]

### 10.5.2 Feasible Generalized Least Squares

To use the results of section 10.5.1, \( \Omega \) must be known. The typical problem involves a small set of parameters \( \theta \) such that \( \Omega = \Omega(\theta) \). Suppose, then, that \( \hat{\theta} \) is a consistent estimator of \( \theta \). To make GLS estimation feasible, we shall use \( \hat{\Omega} = \Omega(\hat{\theta}) \) instead of the true \( \Omega \). It would seem that if \( \text{plim}\hat{\theta} = \theta \), then using \( \hat{\Omega} \) is asymptotically equivalent to using the true \( \Omega \). Let the feasible generalized least squares (FGLS) estimator be denoted

\[\hat{\beta} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y\]
10.6 Maximum Likelihood Estimation

This section considers efficient estimation when the disturbances are normally distributed. If the disturbances are multivariate normally distributed, then the log-likelihood function for the sample is

\[
\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'\Omega^{-1}(y - X\beta) - \frac{1}{2} \ln |\Omega|
\]

Since \(\Omega\) is a matrix of known constants, the maximum likelihood estimator of \(\beta\) is the vector that minimizes the generalized sum of squares,

\[
S_*(\beta) = (y - X\beta)'\Omega^{-1}(y - X\beta)
\]

The necessary conditions for maximizing \(L\) are

\[
\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} X'\Omega^{-1}(y - X\beta) = \frac{1}{\sigma^2} X'_*(y_* - X_*\beta) = 0
\]

\[
\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)'\Omega^{-1}(y - X\beta) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y_* - X_*\beta)'(y_* - X_*\beta)
\]

The solutions are the OLS estimators using the transformed data:

\[
\hat{\beta}_{ML} = (X'_*X_*)^{-1}X'_*y_* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y,
\]

\[
\hat{\sigma}^2_{ML} = \frac{1}{n} (y_* - X_*\hat{\beta})'(y_* - X_*\hat{\beta}) = \frac{1}{n} (y - X\hat{\beta})'\Omega^{-1}(y - X\hat{\beta}),
\]

which implies that with normally distributed disturbances, generalized least squares is also maximum likelihood. When \(\Omega\) is unknown and must be estimated, then it is necessary to maximize the log likelihood with respect to the full set of parameters \([\beta, \sigma^2, \Omega]\) simultaneously.
Chapter 11

Heteroscedasticity

11.1 Introduction

Regression disturbances whose variances are not constant across observations are heteroscedastic.

11.2 Ordinary Least Squares Estimation

11.2.1 Inefficiency of Least Squares

It follows from our earlier results that $b$ is inefficient relative to the GLS estimator. By how much will depend on the setting, but there is some generality to the pattern. As might be expected, the greater is the dispersion in $\omega_i$ across observations, the greater the efficiency of GLS over OLS.

11.2.2 The Estimated Covariance Matrix of $b$

The precise form of the heteroscedasticity is usually unknown, however. In that case, generalized least squares is not usable, and we may need to salvage what we can from the results of ordinary least squares.

As usual,

$$s^2 = \frac{\epsilon' \epsilon}{n-K} - \frac{\epsilon' X (X'X)^{-1} X' \epsilon}{n-K}$$

Taking the two parts separately yields

$$E \left[ \frac{\epsilon' \epsilon}{n-K} \right] = \frac{\text{tr}E[\epsilon' \epsilon | X]}{n-K} = \frac{n\sigma^2}{n-K}$$

As $n \to \infty$, the term will converge to $\sigma^2$. In addition,

$$E \left[ \frac{\epsilon' X (X'X)^{-1} X' \epsilon}{n-K} \right] = \frac{\text{tr}{\left[ \frac{(X'X)^{-1} X'^\prime \Omega X}{n} \right]}}{n-K} = \frac{\sigma^2}{n-K} \text{tr}\left[ \left( \frac{X'X}{n} \right)^{-1} Q_n^* \right]$$

where $Q_n^*$ is

$$Q_n^* = \frac{X'\Omega X}{n} = \frac{1}{n} \sum_{i=1}^n \omega_i x_i x_i'$$
The term will converge to zero if \( b \) is consistent because both matrices in the product are finite. Therefore,

\[
\text{If } b \text{ is consistent, then } \lim_{n \to \infty} E[s^2] = \sigma^2
\]

### 11.3 GMM Estimation of the Heteroscedastic Regression Model

The GMM estimator in the heteroscedastic regression model is produced by the empirical moment equations

\[
\frac{1}{n} \sum_{i=1}^{n} x_i(y_i - x_i' \hat{\beta}_{GMM}) = \frac{1}{n} X' \hat{\epsilon}(\hat{\beta}_{GMM}) = \bar{m}(\hat{\beta}_{GMM}) = 0
\]

The estimator is obtained by minimizing

\[
q = \bar{m}'(\hat{\beta}_{GMM})W\bar{m}(\hat{\beta}_{GMM})
\]

where \( W \) is a positive definite weighting matrix. The optimal weighting matrix would be

\[
W = \{ \text{Asy.Var}[\sqrt{n}\bar{m}(\beta)] \}^{-1}
\]

which is the inverse of

\[
\text{Asy. Var}[\sqrt{n}\bar{m}(\beta)] = \text{Asy. Var} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \epsilon_i \right] = \text{plim}_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \sigma^2 \omega_i x_i x'_i = \sigma^2 Q^*\]

The optimal weighting matrix would be \([\sigma^2 Q^*]^{-1}\).

### 11.4 Testing for Heteroscedasticity

Heteroscedasticity poses potentially severe problems for inferences based on least squares. One can rarely be certain that the disturbances are heteroscedastic however, and unfortunately, what form the heteroscedasticity takes if they are. As such, it is useful to be able to test for homoscedasticity and if necessary, modify our estimation procedures accordingly.

#### 11.4.1 White’s General Test

For formulate the most of the available tests, it is necessary to specify, at least in rough terms the nature of the heteroscedasticity. It would be desirable to be able to test general hypothesis of the form

\[
H_0 : \sigma_i^2 = \sigma^2 \quad \text{for all } i, \\
H_1 : \text{Not } H_0
\]

The correct covariance matrix for the least squares estimator is

\[
\text{Var}[b|X] = \sigma^2[X'X]^{-1}[X'\Omega X][X'X]^{-1}
\]

The conventional estimator is \( V = s^2[X'X]^{-1} \). If there is no heteroscedasticity, then \( V \) will give a consistent estimator of \( \text{Var}[b|X] \), whereas if there is, then it will not. White has devised a statistical test based on this observation. A simple operational version of his test
is carried out by obtaining \( nR^2 \) in the regression of \( e_i^2 \) on a constant and all unique variables contained in \( x \) and all the squares and cross products of the variables in \( x \). The statistic is asymptotically distributed as chi-squared with \( P - 1 \) degrees of freedom, where \( P \) is the number of regressors in the equation, including the constant.

The **White test** is extremely general. The test may reveal heteroscedasticity, but it may instead simply identify some other specification error (such as the omission of \( x^2 \) from a simple regression). In addition, the White test is **nonconstructive**. If we reject the null hypothesis, then the result of the test gives no indication of what to do next.

### 11.4.2 The Goldfeld-Quandt Test

For the Goldfeld-Quandt test, we assume that the observations can be divided into two groups in such a way that under the hypothesis of homoscedasticity, the disturbance variances would be the same in the two groups, whereas under the alternative, the disturbance variances would differ systematically. The most favorable case for this would be the **groupwise heteroscedastic** model such as \( \sigma_i^2 = \sigma^2 x_i^2 \) for some variable \( x \). By ranking the observations based on this \( x \), we can separate the observations into those with high and low variances. The test is applied by dividing the sample into two groups with \( n_1 \) and \( n_2 \) observations. To obtain statistically independent variance estimators, the regression is then estimated separately with the two sets of observations. The test statistic is

\[
F[n_1 - K, n_2 - K] = \frac{e_1'e_1/(n_1 - K)}{e_2'e_2/(n_2 - K)}
\]

where we assume that the disturbance variance is larger in the first sample. Under the null hypothesis of homoscedasticity, this statistic has an \( F \) distribution with \( n_1 - K \) and \( n_2 - K \) degrees of freedom. The sample value can be referred to the standard \( F \) table to carry out the test, with a large value leading to the rejection of the null hypothesis.

### 11.4.3 The Breusch-Pagan/Godfrey LM Test

Breusch and Pagan have devised a **Lagrange multiplier test** of the hypothesis that \( \sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha'z_i) \), where \( z_i \) is a vector of independent variables. The model is homoscedastic if \( \alpha = 0 \). The test can be carried out with a simple regression:

\[
LM = \frac{1}{2} \text{ explained sum of squares in the regression of } e_i^2/(e'e/n) \text{ on } z_i
\]

For computational purposes, let \( Z \) be the \( n \times P \) matrix of observations on \((1, z_i)\), and let \( g \) be the vector of observations of \( g_i = e_i^2/(e'e/n) - 1 \). Then

\[
LM = \frac{1}{2} [g'Z(Z'Z)^{-1}Z'g]
\]

### 11.5 Weighted Least Squares When \( \Omega \) is Known

The GLS estimator is

\[
\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y
\]

Consider the most general case, \( \text{Var}[e_i|x_i] = \sigma_i^2 = \sigma^2 \omega_i \). Then \( \Omega^{-1} \) is a diagonal matrix whose \( i \)th diagonal element is \( 1/\omega_i \). The GLS estimator is obtained by regressing \( Py \) and
Applying ordinary least squares to the transformed model, we obtained the \textbf{weighted least squares (WLS)} estimator.

\[ \hat{\beta} = \left[ \sum_{i=1}^{n} w_i x_i x_i' \right]^{-1} \left[ \sum_{i=1}^{n} w_i x_i y_i \right] \]

where \( w_i = 1/\omega_i \). Note, the weighted least squares estimator is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances.

### 11.6 Estimation When \( \Omega \) Contains Unknown Parameters

The new problem is that we must first find consistent estimators of the unknown parameters in \( \Omega(\alpha) \). Two methods are typically used, two-step GLS and maximum likelihood.

#### 11.6.1 Two-Step Estimation

For the heteroscedastic model, the GLS estimator is

\[ \hat{\beta} = \left[ \sum_{i=1}^{n} \left( \frac{1}{\sigma_i^2} \right) x_i x_i' \right]^{-1} \left[ \sum_{i=1}^{n} \left( \frac{1}{\sigma_i^2} \right) x_i y_i \right] \]

The \textbf{two-step estimators} are computed by first obtaining estimates \( \hat{\sigma}_i^2 \), usually using some function of the ordinary least squares residuals. Then, \( \hat{\beta} \) uses equation above.

The two-step estimator may be iterated by recomputing the residuals after computing the FGLS estimates and then reentering the computation.

#### 11.6.2 Maximum Likelihood Estimation

The log-likelihood function for a sample of normally distributed observations is

\[ \ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \left[ \ln \sigma_i^2 + \frac{1}{\sigma_i^2} (y_i - x_i' \beta)^2 \right] \]

For simplicity, let

\[ \sigma_i^2 = \sigma^2 f_i(\alpha) \]

where \( \alpha \) is the vector of unknown parameters in \( \Omega(\alpha) \) and \( f_i(\alpha) \) is indexed by \( i \) to indicate that it is a function of \( z_i \). Note that \( \Omega(\alpha) = \text{diag}[f_i(\alpha)] \). The log-likelihood function is

\[ \ln L = -\frac{n}{2} \ln(2\pi) + \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \left[ \ln f_i(\alpha) + \frac{1}{\sigma_i^2} \left( \frac{1}{f_i(\alpha)} \right) (y_i - x_i' \beta)^2 \right] \]

For convenience in what follows, substitute \( \epsilon_i \) for \( (y_i - x_i' \beta) \), denote \( f_i(\alpha) \) as simply \( f_i \), and denote the vector of derivatives \( \partial f_i(\alpha)/\partial \alpha \) as \( g_i \). Then, the derivatives of the log-likelihood
function are
\[ \frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} x_i \frac{\epsilon_i}{\sigma^2 f_i} \]
\[ \frac{\partial \ln L}{\partial \sigma^2} = \sum_{i=1}^{n} \left( \frac{1}{2\sigma^2} \right) \left( \frac{\epsilon_i^2}{\sigma^2 f_i} - 1 \right) \]
\[ \frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \left( \frac{1}{2} \right) \left( \frac{\epsilon_i^2}{\sigma^2 f_i} - 1 \right) \left( \frac{1}{f_i} \right) g_i \]

The **maximum likelihood estimators** are those values of \( \beta, \sigma^2, \) and \( \alpha \) that simultaneously equate these derivatives to zero. The likelihood equations are generally highly nonlinear and will usually require an iterative solution.

Let \( G \) be the \( n \times M \) matrix with \( i^{th} \) row equal to \( \partial f_i / \partial \alpha' = g_i' \) and let \( i \) denote an \( n \times 1 \) column vector of 1s. The asymptotic covariance matrix for the maximum likelihood estimator in this model is
\[
\left( -E \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right] \right)^{-1} = 
\begin{bmatrix}
(1/\sigma^2) X' \Omega^{-1} X & 0 & 0 \\
0 & n/(2\sigma^4) & (1/(2\sigma^4)) \Omega^{-1} G' \\
0 & (1/(2\sigma^4)) G \Omega^{-1} i & (1/2) G' \Omega^{-2} G
\end{bmatrix}^{-1}
\]

where \( \gamma' = [\beta', \sigma^2, \alpha'] \).

### 11.6.3 Model Based Tests for Heteroscedasticity

**Wald Test** The Wald statistic is computed by extracting from the full parameter vector and its estimated asymptotic covariance matrix the subvector \( \hat{\alpha} \) and its asymptotic covariance matrix. Then,
\[ W = \hat{\alpha} \{ \text{Est. Asy. Var}[\hat{\alpha}] \}^{-1} \hat{\alpha} \]

**Likelihood Ratio Test** The results of the homoscedastic least squares regression are generally used to obtain the initial values for the iterations. The restricted log-likelihood value is a by-product of the initial setup; \( \log - L_R = -(n/2)[1 + \ln 2\pi + \ln(e'e/n)] \). The unrestricted log-likelihood, \( \log - L_U \), is obtained as the objective function for the estimation. Then, the statistic for the test is
\[ LR = -2(\ln - L_R - \ln - L_U) \]

**Lagrange Multiplier Test** To set up the LM test, we refer back to the model above. At the restricted estimates \( \alpha = 0, \beta = b, \sigma^2 = e'e/n \) (not \( n - K \)), \( f_i = 1 \) and \( \Omega(0) = I \). Thus, the first derivatives vector evaluated at the least squares estimates is
\[
\begin{align*}
\left. \frac{\partial \ln L}{\partial \beta} \right|_{(\beta = b, \sigma^2 = e'e/n, \hat{\alpha} = 0)} &= 0 \\
\left. \frac{\partial \ln L}{\partial \sigma^2} \right|_{(\beta = b, \sigma^2 = e'e/n, \hat{\alpha} = 0)} &= 0 \\
\left. \frac{\partial \ln L}{\partial \alpha} \right|_{(\beta = b, \sigma^2 = e'e/n, \hat{\alpha} = 0)} &= \sum_{i=1}^{n} \left( \frac{1}{2} \left( \frac{\epsilon_i^2}{e'e/n} - 1 \right) \right) g_i = \sum_{i=1}^{n} \frac{1}{2} v_i g_i
\end{align*}
\]

64
The negative expected inverse of the Hessian is

\[
\left( -E \left[ \frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} \right]_{\alpha=0} \right)^{-1} = \begin{bmatrix}
(1/\sigma^2)X'X & 0 \\
0' & n/(2\sigma^4)
\end{bmatrix} \begin{bmatrix}
0 \\
0' [1/(2\sigma^4)]g' (1/2)G'G
\end{bmatrix}^{-1} = \{ -E(H) \}^{-1}
\]

where \( g = \sum_{i=1}^{n} g_i \) and \( G'G = \sum_{i=1}^{n} g_i g_i' \). The LM statistic will be

\[
LM = \left[ \frac{\partial \ln L}{\partial \gamma} \right] (\gamma = b, e' e/n, 0) \left[ \frac{\partial \ln L}{\partial \gamma} \right]^{-1} \left[ -E[H] \right]^{-1} \left[ -E[H] \right] \left[ \frac{\partial \ln L}{\partial \gamma} \right] (\gamma = b, e' e/n, 0)
\]

With a bit of algebra for the partitioned inverse, you can show that this reduces to

\[
LM = \frac{1}{2} \left\{ \sum_{i=1}^{n} v_i g_i \right\} \left[ \sum_{i=1}^{n} (g_i - \bar{g}_i)(g_i - \bar{g}_i)' \right]^{-1} \left\{ \sum_{i=1}^{n} v_i g_i \right\}
\]

### 11.7 Applications

#### 11.7.1 Multiplicative Heteroscedasticity

Harvey’s (1976) model of multiplicative heteroscedasticity is a very flexible, general model that includes most of the useful formulations as special cases. The general formulation is

\[
\sigma_i^2 = \sigma^2 \exp(z_i' \alpha)
\]

More generally, a model with heteroscedasticity of the form

\[
\sigma_i^2 = \sigma^2 \prod_{m=1}^{M} z_{im}^{\alpha_m}
\]

results if the logs of the variables are placed in \( z_i \).

Let \( z_i \) include a constant term so that \( z_i' = [1, q_i'] \), where \( q_i \) is the original set of variables, and let \( \gamma' = [\ln \sigma^2, \alpha'] \). Then, the model is simply \( \sigma_i^2 = \exp(\gamma' z_i) \). Once the full parameter vector is estimated, \( \exp(\gamma_1) \) provides the estimator of \( \sigma^2 \).

The log-likelihood is

\[
\ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \ln \sigma_i^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_i^2}
\]

\[
= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} z_i' \gamma - \frac{1}{2} \sum_{i=1}^{n} \frac{\epsilon_i^2}{\exp(z_i' \gamma)}
\]

The likelihood equations are

\[
\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} x_i \frac{\epsilon_i}{\exp(z_i' \gamma)} = X'\Omega\epsilon = 0
\]

\[
\frac{\partial \ln L}{\partial \gamma} = \frac{1}{2} \sum_{i=1}^{n} z_i \left( \frac{\epsilon_i^2}{\exp(z_i' \gamma)} - 1 \right) = 0
\]
The terms in the Hessian are
\[
\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \frac{1}{\exp(z_i'^\gamma)} x_i x_i' = -X'\Omega^{-1}X,
\]
\[
\frac{\partial^2 \ln L}{\partial \beta \partial \gamma'} = -\sum_{i=1}^{n} \frac{\epsilon_i'}{\exp(z_i'^\gamma)} x_i z_i',
\]
\[
\frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} = -\frac{1}{2} \sum_{i=1}^{n} \frac{\epsilon_i'^2}{\exp(z_i'^\gamma)} z_i z_i'.
\]

The expected value of \( \partial^2 \ln L/\partial \beta \partial \gamma' \) is 0 since \( E[\epsilon_i | x_i, z_i] = 0 \). The expected value of the fraction in \( \partial^2 \ln L/\partial \gamma \partial \gamma' \) is \( E[\epsilon_i'^2 \sigma_i^2 | x_i, z_i] = 1 \). Let \( \delta = [\beta, \gamma] \). Then,
\[
-E \left( \frac{\partial^2 \ln L}{\partial \delta \partial \delta'} \right) = \left[ \begin{array}{cc} X'\Omega^{-1}X & 0 \\ 0' & \frac{1}{2}Z'Z \end{array} \right] = -H
\]

The scoring method is
\[
\delta_{t+1} = \delta_t - H_t^{-1}g_t,
\]
where \( \delta_t \) (i.e., \( \beta_t, \gamma_t \), and \( \Omega_t \)) is the estimate at iteration \( t \), \( g_t \) is the two-part vector of first derivatives [\( \partial \ln L/\partial \beta_t \), \( \partial \ln L/\partial \gamma_t \)] and \( H_t \) is partitioned likewise.

The updated value of \( \gamma \) is computed by adding the vector of slopes in the least squares regression of \([\epsilon_i'^2 / \exp(z_i'^\gamma) - 1] \) on \( z_i \) to the old one. Thereafter, the iteration is simply:

1. Estimate the disturbance variance \( \sigma_i^2 \) with \( \exp(\gamma_i'z_i) \)
2. Compute \( \beta_{t+1} \) by FGLS
3. Update \( \gamma_t \) using the regression described in the preceding paragraph
4. Compute \( d_{t+1} = [\beta_{t+1}, \gamma_{t+1}] - [\beta_t, \gamma_t] \). If \( d_{t+1} \) is large, then return to step 1.

### 11.7.2 Groupwise Heteroscedasticity

A groupwise heteroscedastic regression has structural equations
\[
y_{i} = x_{i}'\beta + \epsilon_{i}, \quad i = 1, ..., n
\]
\[
E[\epsilon_i | x_i] = 0, \quad i = 1, ..., n
\]

The \( n \) observations are grouped into \( G \) groups, each with \( n_g \) observations. The slope vector is the same in all groups, but within group \( g \):
\[
\text{Var}[\epsilon_{ig} | x_{ig}] = \sigma_{ig}^2, \quad i = 1, ..., n_g
\]

If the variances are known, then the GLS estimator is
\[
\hat{\beta} = \left[ \sum_{g=1}^{G} \left( \frac{1}{\sigma_g^2} X_g'X_g \right) \right]^{-1} \left[ \sum_{g=1}^{G} \left( \frac{1}{\sigma_g^2} X_g'y_g \right) \right]
\]

Since \( X_g'y_g = X_g'X_gb_g \), where \( b_g \) is the OLS estimator in the \( g \)th subset of observations,
\[
\hat{\beta} = \left[ \sum_{g=1}^{G} \left( \frac{1}{\sigma_g^2} \right) X_g'X_g \right]^{-1} \left[ \sum_{g=1}^{G} \left( \frac{1}{\sigma_g^2} \right) X_g'y_g \right] = \left[ \sum_{g=1}^{G} V_g \right]^{-1} \left[ \sum_{g=1}^{G} V_gb_g \right] = \sum_{g=1}^{G} W_gb_g
\]

This result is a matrix weighted average of the \( G \) least squares estimators.
11.8 Autoregressive Conditional Heteroscedasticity

11.8.1 The ARCH(1) Model

The simplest form of this model is the ARCH(1) model,

\[ y_t = \beta' x_t + \epsilon_t \]

\[ \epsilon_t = u_t \sqrt{\alpha_0 + \alpha_1 \epsilon_{t-1}^2}, \]

where \( u_t \) is distributed as standard normal. It follows that \( E[\epsilon_t | x_t, \epsilon_{t-1}] = 0 \), so that \( E[\epsilon_t | x_t] = \beta' x_t \). Therefore, this model is a classical regression model. But

\[ \text{Var}[\epsilon_t | \epsilon_{t-1}] = E[\epsilon_t^2 | \epsilon_{t-1}] = E[u_t^2 | \alpha_0 + \alpha_1 \epsilon_{t-1}^2] = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 \]

so \( \epsilon_t \) is conditionally heteroscedastic, not with respect to \( x_t \) as we considered in the preceding sections, but with respect to \( \epsilon_{t-1} \). The unconditional variance of \( \epsilon_t \) is

\[ \text{Var}[\epsilon_t] = \text{Var}[E(\epsilon_t | \epsilon_{t-1})] + E[\text{Var}(\epsilon_t | \epsilon_{t-1})] = \alpha_0 + \alpha_1 E[\epsilon_{t-1}^2] = \alpha_0 + \alpha_1 \text{Var}[\epsilon_{t-1}] \]

If the process generating the disturbances is weakly stationary, then the unconditional variance is not changing over time so

\[ \text{Var}[\epsilon_t] = \text{Var}[\epsilon_{t-1}] = \frac{\alpha_0}{1 - \alpha_1} \]

For this ratio to be finite and positive, \( |\alpha_1| \) must be less than 1. Then, unconditionally, \( \epsilon_t \) is distributed with mean zero and variance \( \sigma^2 = \alpha_0/(1 - \alpha_1) \). Therefore, the model obeys the classical assumptions, and ordinary least squares is the most efficient linear unbiased estimator of \( \beta \).

11.8.2 ARCH(\(q\)), ARCH-In-Mean and Generalized ARCH Models

The natural extension of the ARCH(1) model presented before is a more general model with longer lags. The ARCH(\(q\)) process,

\[ \sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + ... + \alpha_q \epsilon_{t-q}^2 \]

is a \( q \)th order moving average [MA(\(q\))] process. The ARCH-in-Mean or ARCH-M, model suggested by Engle, Lilien, and Robins(1987) is another natural extension. The model states that

\[ y_t = \beta' x_t + \delta \sigma_t^2 + \epsilon_t \]

\[ \text{Var}[\epsilon_t | \Psi_t] = ARCH(q) \]

Among the interesting implications of this modification of the standard model is that under certain assumptions, \( \delta \) is the coefficient of relative risk aversion.

The model of generalized autoregressive conditional heteroscedasticity (GARCH) is defined as follows. The underlying regression is the usual one in

\[ y_t = \beta' x_t + \epsilon_t \]

\[ \epsilon_t = u_t \sqrt{\alpha_0 + \alpha_1 \epsilon_{t-1}^2}, \]

67
Conditioned on an information set at time \( t \), denoted \( \Psi_t \), the distribution of the disturbance is assumed to be

\[ \epsilon_t | \Psi_t \sim N(0, \sigma_t^2) \]

where the conditional variance is

\[ \sigma_t^2 = \alpha_0 + \delta_1 \sigma_{t-1}^2 + \delta_2 \sigma_{t-2}^2 + \ldots + \delta_p \sigma_{t-p}^2 + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \ldots + \alpha_q \epsilon_{t-q}^2 \]

Define

\[ z_t = [1, \sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2, \epsilon_{t-1}^2, \epsilon_{t-2}^2, \ldots, \epsilon_{t-q}^2]' \]

and

\[ \gamma = [\alpha_0, \delta_1, \delta_2, \ldots, \delta_p, \alpha_1, \ldots, \alpha_q]' = [\alpha_0', \delta', \alpha']'. \]

Then

\[ \sigma_t^2 = \gamma' z_t \]

Notice that the conditional variance is defined by an autoregressive-moving average [ARMA(\( p, q \))] process in the innovations \( \epsilon_t^2 \). The model is a GARCH(\( p, q \)) model, where \( p \) refers, as before, to the order of the autoregressive part. The stationary conditions are important in this context to ensure that the moments of the normal distribution are finite.

### 11.8.3 Maximum Likelihood Estimation of the GARCH Model

Bollerslev describes a method of estimation based on the BHHH algorithm. Following the suggestions of Harvey (1976), it turns out that there is a simpler way to estimate the GARCH model that is also very illuminating.

For normally distributed disturbances, the log-likelihood for a sample of \( T \) observations is

\[ \ln L = \sum_{t=1}^{T} -\frac{1}{2} \left[ \ln(2\pi) + \ln \sigma_t^2 + \frac{\epsilon_t^2}{\sigma_t^2} \right] = \sum_{t=1}^{T} \ln f_t(\theta) = \sum_{t=1}^{T} l_t(\theta), \]

where \( \epsilon_t = y_t - x_t' \beta \) and \( \theta = (\beta', \alpha_0, \alpha', \delta')' = (\beta', \gamma')' \). Let \( l_t \) denote \( \ln f_t(\theta) \). The first derivatives with respect to the variance parameters are

\[ \frac{\partial l_t}{\partial \gamma} = -\frac{1}{2} \left[ 1 - \frac{\epsilon_t^2}{\sigma_t^2} \right] \frac{\partial \sigma_t^2}{\partial \gamma} = \frac{1}{2} \left( \frac{1}{\sigma_t^2} \right) \frac{\partial \sigma_t^2}{\partial \gamma} \left( \frac{\epsilon_t^2}{\sigma_t^2} - 1 \right) = \frac{1}{2} \left( \frac{1}{\sigma_t^2} \right) g_t v_t = b_t v_t \]

Suppose, for now, that there are no regression parameters. Newton’s method for estimating the variance parameters would be

\[ \gamma^{i+1} = \gamma^i - H^{-1} g \]

where \( H \) indicates the Hessian and \( g \) is the first derivative vector. Following Harvey’s suggestion, we will use the method of scoring instead. After taking expectations, the iteration reduces to a linear regression of \( u_t = \epsilon_t \) on regressors \( w_t = g_t/\sigma_t^2 \). That is,

\[ \gamma^{i+1} = \hat{\gamma}^i + \left[ W_*' W_* \right]^{-1} W_*' u_t = \hat{\gamma}^i + \left[ W_*' W_* \right]^{-1} \left( \frac{\partial \ln L}{\partial \gamma} \right) \]

where row \( t \) of \( W_* \) is \( w_{*t} \).
Consider then the slope parameters, $\beta$. The same type of modified scoring method as used earlier produces the iteration

$$
\hat{\beta}^{i+1} = \hat{\beta}^i + \left[ \sum_{t=1}^{T} \frac{x_t x'_t}{\sigma_t^2} + \frac{1}{2} \left( \frac{d_t}{\sigma_t^2} \right) \left( \frac{d_t}{\sigma_t^2} \right)' \right]^{-1} \left[ \sum_{t=1}^{T} \frac{x_t \epsilon_t}{\sigma_t^2} + \frac{1}{2} \left( \frac{d_t}{\sigma_t^2} \right) \epsilon_t \right]
$$

which has been referred to as a double-length regression. The update vector $h_i$ is the vector of slopes in an augmented or double-length generalized regression,

$$
h_i = [C'\Omega^{-1}C]^{-1}[C'\Omega^{-1}a],
$$

where $C$ is a $2T \times K$ matrix whose first $T$ rows are the $X$ from the original regression model and whose next $T$ rows are $(1/\sqrt{2})d_t'/\sigma_t^2$, $t = 1, ..., T$; $a$ is a $2T \times 1$ vector whose first $T$ elements are $\epsilon_t$ and whose next $T$ elements are $(1/\sqrt{2})v_t/\sigma_t^2$, $t = 1, ..., T$; and $\Omega$ is a diagonal matrix with $1/\sigma_t^2$ in positions $1, ..., T$ and ones below observations $T$.

### 11.8.4 Testing for GARCH Effects

The simplest approach is to examine the squares of the least squares residuals. The autocorrelations (correlations with lagged values) of the squares of the residuals provide evidence about ARCH effects. An LM test of ARCH($q$) against the hypothesis of no ARCH effects [ARCH(0), the classical model] can be carried out by computing $\chi^2 = TR^2$ in the regression of $e_t^2$ on a constant and $q$ lagged values. Under the null hypothesis of no ARCH effects, the statistic has a limiting chi-squared distribution with $q$ degrees of freedom. Values larger than the critical table value give evidence of the presence of ARCH (or GARCH) effects.

Bollerslev suggests a Lagrange multiplier statistic that is, in fact, surprisingly simple to compute. The LM test for GARCH($p, 0$) against GARCH($p, q$) can be carried out by referring $T$ times the $R^2$ in the linear regression defined in

$$
\hat{\gamma}^{i+1} = \hat{\gamma}^i + [W_p'W_s]^{-1}W_p'v_s = \hat{\gamma}^i + [W_p'W_s]^{-1} \left( \frac{\partial \ln L}{\partial \gamma} \right)
$$

to the chi-squared critical value with $q$ degrees of freedom. The test for ARCH($q$) against GARCH($p, q$) is exactly the same as that for ARCH($p$) against ARCH($p + q$).
Chapter 12

Serial Correlation

12.1 Introduction

Time-series data often display autocorrelation, or serial correlation of the disturbances across periods.

12.2 The Analysis of Time-Series Data

A time-series model will typically describe the path of a variable $y_t$ in terms of contemporaneous (and perhaps lagged) factors $x_t$, disturbances (innovations), $\epsilon_t$ and its own past, $y_{t-1},...$. For example

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 + y_{t-1} + \epsilon_t$$

The sequence of observations, $\{y_t\}_{t=-\infty}^{\infty}$ is a time-series process which is characterized by its time ordering and its systematic correlation between observations in the sequence.

In the current context, this distribution of $\epsilon_t$ is said to be covariance stationary or weakly stationary. It can be said, for example, that $\epsilon_t$ is generated by a time-series process whose mean and variance are not changing over time.

Except in very special cases, we would expect all the elements in the $T$ component random vector $(y_1,...,y_T)$ to be correlated. In this instance, said correlation is called "autocorrelation".

12.3 Disturbance Processes

12.3.1 Characteristics of Disturbance Processes

In the usual time-series setting, the disturbances are assumed to be homoscedastic but correlated across observations, so that

$$E[\epsilon\epsilon'|X] = \sigma^2 \Omega,$$

where $\sigma^2 \Omega$ is a full, positive definite matrix with a constant $\sigma^2 = \text{Var}[\epsilon_t|X]$ on the diagonal.

We define the autocovariances:

$$\text{Cov}[\epsilon_t, \epsilon_{t-s}|X] = \text{Cov}[\epsilon_{t+1}, \epsilon_t|X] = \sigma^2 \Omega_{t,t-s} = \gamma_s = \gamma_{-s}$$
Note that $\sigma^2 \Omega_{tt} = \gamma_0$. The correlation between $\epsilon_t$ and $\epsilon_{t-s}$ is their autocorrelation,

$$\text{Corr}[\epsilon_t, \epsilon_{t-s} | X] = \frac{\text{Cov}[\epsilon_t, \epsilon_{t-s} | X]}{\sqrt{\text{Var}[\epsilon_t | X] \text{Var}[\epsilon_{t-s} | X]}} = \frac{\gamma_s}{\gamma_0} = \rho_s = \rho_{-s}$$

We can then write

$$E[\epsilon \epsilon' | X] = \Gamma = \gamma_0 R$$

where $\Gamma$ is an autocovariance matrix and $R$ is an autocorrelation matrix—the $ts$ element is an autocorrelation coefficient

$$\rho_{ts} = \frac{\gamma_{|t-s|}}{\gamma_0}$$

Different types of processes imply different patterns in $R$. The most frequently analyzed process is a first-order autoregression or AR(1) process,

$$\epsilon_t = \rho \epsilon_{t-1} + u_t,$$

where $u_t$ is a stationary, nonautocorrelated (“white noise”) process and $\rho$ is a parameter. Higher-Order autoregressive processes of the form

$$\epsilon_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_p \epsilon_{t-p} + u_t$$

imply more involved patterns, including, for some values of the parameters, cyclical behavior of the autocorrelations.

### 12.3.2 AR(1) Disturbances

The first-order autoregressive disturbance, or AR(1) process, is represented in the autoregressive form as

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

where

$$E[u_t] = 0,$$
$$E[u_t^2] = \sigma_u^2,$$

and

$$\text{Cov}[u_t, u_s] = 0, \quad \text{if} \; t \neq s$$

By repeated substitution, we have

$$\epsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + ...$$

Since the successive values of $u_t$ are uncorrelated, the variance of $\epsilon_t$ is

$$\text{Var}[\epsilon_t] = \sigma_u^2 + \rho^2 \sigma_u^2 + \rho^4 \sigma_u^2 + ...$$

To proceed, a restriction must be placed on $\rho$,

$$|\rho| < 1$$
and

\[ \text{Var}[\epsilon_t] = \frac{\sigma_u^2}{1 - \rho^2} = \sigma_t^2 \]

With the stationarity assumption, there is an easier way to obtain the variance:

\[ \text{Var}[\epsilon_t] = \rho^2 \text{Var}[\epsilon_{t-1}] + \sigma_u^2 \]

By repeated substitution, we see that for any \( s \),

\[ \epsilon_t = \rho^s \epsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i} \]

It follows that

\[ \text{Cov}[\epsilon_t, \epsilon_{t-s}] = E[\epsilon_t \epsilon_{t-s}] = \rho^s \sigma_u^2 \frac{1}{1 - \rho^2} \]

Dividing by \( \gamma_0/(1 - \rho^2) \) provides the autocorrelations:

\[ \text{Corr}[\epsilon_t, \epsilon_{t-s}] = \rho_s = \rho^s. \]

### 12.4 Some Asymptotic Results for Analyzing Time Series Data

As we say in our analysis of heteroscedasticity, whether least squares is consistent or not, depends on the matrices

\[ Q_T = (1/T)X'X \]

and

\[ Q_T^* = (1/T)X'\Omega X \]

In our earlier analyses, we were able to argue for convergence of \( Q_T \) to a positive definite matrix of constants, \( Q \), by invoking laws of large numbers. But, these theorems assume that the observations in the sums are independent, which as suggested is surely not the case here. Thus, we require a different tool for this result. We can expand the matrix \( Q_T^* \) as

\[ Q_T^* = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \rho_{ts} x_t x_s' \]

where \( x_t' \) and \( x_s' \) are rows of \( X \) and \( \rho_{ts} \) is the autocorrelation between \( \epsilon_t \) and \( \epsilon_s \). Sufficient conditions for this matrix to converge are that \( Q_T \) converge and that the correlations between disturbances die off reasonably rapidly as the observations become further apart in time.

#### 12.4.1 Convergence of Moments – The Ergodic Theorem

**DEFINITION Strong Stationarity**

A time series process, \( \{z_t\}_{t=-\infty}^{\infty} \) is strongly stationary, or “stationary” if the joint probability distribution of any set of \( k \) observations in the sequence, \( [z_t, z_{t+1}, \ldots, z_{t+k}] \) is the same regardless of the origin, \( t \), in the time scale.
For example, if we add \( u_t \sim N[0, \sigma_u^2] \), then the resulting process \( \{\epsilon_t\}_{t=-\infty}^{\infty} \) can easily be shown to be strongly stationary.

**DEFINITION Weak Stationarity**
A time series process, \( \{z_t\}_{t=-\infty}^{\infty} \) is weakly stationary, or “covariance stationary” if \( E[z_t] \) is finite and is the same for all \( t \) and if the covariances between any two observations, \( \text{Cov}[z_t, z_{t-k}] \), is a finite function only of model parameters and their distance apart in time, \( k \), but not of the absolute location of either observation on the time scale.

Weak stationary is obviously implied by strong stationary, though it requires less since the distribution can, at least in principle, be changing on the time axis.

**DEFINITION Ergodicity**
A time series process, \( \{z_t\}_{t=-\infty}^{\infty} \) is ergodic if for any two bounded functions that map vectors in the \( a \) and \( b \) dimensional real vector spaces to real scalars, \( f: \mathbb{R}^a \rightarrow \mathbb{R}^1 \) and \( g: \mathbb{R}^b \rightarrow \mathbb{R}^1 \),

\[
\lim_{k \rightarrow \infty} |E[f(z_t, z_{t+1}, \ldots, z_{t+a})g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b})]| = |E[f(z_t, z_{t+1}, \ldots, z_{t+a})]| |E[g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b})]|
\]

The definition states essentially that if events are separated far enough in time, then they are “asymptotically independent”.

**THEOREM The Ergodic Theorem**
If \( \{z_t\}_{t=-\infty}^{\infty} \) is a time-series process which is stationary and ergodic and \( E[|z_t|] \) is a finite constant and \( E[z_t] = \mu \), and if \( \bar{z}_T = (1/T) \sum_{t=1}^{T} z_t \), then \( \bar{z}_T \xrightarrow{a.s.} \mu \). Note that the convergence is almost surely, not in probability or in mean square.

What we have in The Ergodic Theorem is, for sums of dependent observations, a counterpart to the laws of large numbers that we have used at many points in the preceding chapters.

**THEOREM Ergodicity of Functions**
If \( \{z_t\}_{t=-\infty}^{\infty} \) is a time-series process which is stationary and ergodic and if \( y_t = f(z_t) \) is a measurable function in the probability space that defines \( z_t \), then \( y_t \) is also stationary and ergodic. Let \( \{z_t\}_{t=-\infty}^{\infty} \) define a \( K \times 1 \) vector valued stochastic process–each element of the vector is an ergodic and stationary series and the characteristics of ergodicity and stationarity apply to the joint distribution of the elements of \( \{z_t\}_{t=-\infty}^{\infty} \). Then the ergodic Theorem applies to functions of \( \{z_t\}_{t=-\infty}^{\infty} \).

In particular, our minimal assumptions about the data are

**ASSUMPTION Ergodic Data Series:**
In the regression model, \( y_t = x_t' \beta + \epsilon_t \), \( \{x_t, \epsilon_t\}_{t=-\infty}^{\infty} \) is jointly stationary and ergodic process.

By analyzing terms element by element we can use these results directly to assert that averages of \( w_t = x_t \epsilon_t, Q_t = x_t x'_t \) and \( Q^*_t = \epsilon^2_t x_t x_t' \) will converge to their population counterparts, 0, \( Q \) and \( Q^* \).
12.4.2 Convergence To Normality–A Central Limit Theorem

As noted earlier, we cannot invoke the familiar central limit theorem because the observations in the sum are not independent. But, with the assumptions already made, we do have an alternative result. Some needed preliminaries are as follows:

**DEFINITION Martingale Sequence**
A vector sequence \( \mathbf{z}_t \) is a martingale sequence if
\[
E[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, ...] = \mathbf{z}_{t-1}
\]

An important example of a martingale sequence is the random walk,
\[
\mathbf{z}_t = \mathbf{z}_{t-1} + \mathbf{u}_t
\]
where \( \text{Cov} [\mathbf{u}_t, \mathbf{u}_s] = 0 \) for all \( t \neq s \). Then
\[
E[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, ...] = \mathbf{z}_{t-1} + 0 = \mathbf{z}_{t-1}
\]

**DEFINITION Martingale Difference Sequence**
A vector sequence \( \mathbf{z}_t \) is a martingale difference sequence if
\[
E[\mathbf{z}_t - \mathbf{z}_{t-1} | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, ...] = 0.
\]

With this definition, we have the following broadly encompassing result:

**THEOREM Martingale Difference Central Limit Theorem**
If \( \mathbf{z}_t \) is a vector valued stationary and ergodic martingale difference sequence, with \( E[\mathbf{z}_t \mathbf{z}_t'] = \Sigma \), where \( \Sigma \) is a finite positive definite matrix, and if \( \bar{z}_T = (1/T) \sum_{t=1}^{T} \mathbf{z}_t \), then
\[
\sqrt{T} \bar{z}_T \xrightarrow{d} N[0, \Sigma]
\]

Suppose as before that \( \{\mathbf{z}_t\}_{t=-\infty}^{\infty} \) is a stationary and ergodic stochastic process. We consider \( \sqrt{T} \bar{z}_T \). The following conditions are assumed:

1. **Summability of autocovariances:** With dependent observations,
   \[
   \lim_{T \to \infty} \text{Var}[\sqrt{T} \bar{z}_T] = \sum_{t=0}^{\infty} \sum_{s=0}^{\infty} \text{Cov}[\mathbf{z}_t \mathbf{z}_s'] = \sum_{k=\text{infty}}^{\infty} \Gamma_k = \Gamma^*.
   \]
   To begin, we will need to assume that this matrix is finite, a condition called **summability**. If the sum is to be finite, then the \( k = 0 \) term must be finite, which gives us a necessary condition
   \[
   E[\mathbf{z}_t \mathbf{z}_s'] = \Gamma_0, \text{ a finite matrix}
   \]

2. **Asymptotic uncorrelatedness:** \( E[\mathbf{z}_t | \mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, ...] \) converges in mean square to zero as \( k \to \infty \)

3. **Asymptotic negligibility of innovations:** Let
   \[
   r_{tk} = E[\mathbf{z}_t | \mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, ...] - E[\mathbf{z}_t | \mathbf{z}_{t-k-1}, \mathbf{z}_{t-k-2}, ...]
   \]
   An observation \( \mathbf{z}_t \) may be viewed as the accumulated information that has entered the process since it began up to time \( t \). Thus, it can be shown that
   \[
   \mathbf{z}_t = \sum_{s=0}^{\infty} r_{ts}
   \]
   The vector \( r_{tk} \) can be viewed as the information in this accumulated sum that entered the process at time \( t - k \). In words, condition (3) states that information eventually becomes negligible as it fades far back in time from the current observation.
With all this machinery in place, we now have the theorem we will need:

**THEOREM**  **Gordin’s Central Limit Theorem**

If condition (1) to (3) listed above are met, then $\sqrt{T} \bar{z}_T \xrightarrow{d} N[0, \Gamma^*]$

### 12.5 Least Squares Estimation

Broadly, least squares fares better in data which have long periods and little cyclical variation, such as aggregate output series. As might be expected, the greater is the autocorrelation in $\epsilon$, the greater will be the benefit to using generalized least squares.

#### 12.5.1 Asymptotic Properties of Least Squares

There is an important exception to the results in the preceding paragraph. If the regression contains any lagged values of the dependent variable, then least squares will no longer be unbiased or consistent. To take the simplest case, suppose that

$$ y_t = \beta y_{t-1} + \epsilon_t $$

and assume $|\beta| < 1, |\rho| < 1$. Rearrange by subtracting $\rho y_{t-1}$ from $y_t$. Then,

$$ y_t = (\beta + \rho) y_{t-1} - \beta \rho y_{t-2} + u_t $$

Since $u_t$ is an innovation in period $t$, it is uncorrelated with both regressors, and least squares regression of $y_t$ on $(y_{t-1}, y_{t-2})$ estimates $\rho_1 = (\beta + \rho)$ and $\rho_2 = -\beta \rho$. By stationarity, $\text{Var}[y_t] = \text{Var}[y_{t-1}]$, and $\text{Cov}[y_t, y_{t-1}] = \text{Cov}[y_{t-1}, y_{t-2}]$, and so on. These imply the following relationships

$$ \gamma_0 = \rho_1 \gamma_1 + \rho_2 \gamma_2 + \sigma^2_u $$

$$ \gamma_1 = \rho_1 \gamma_0 + \rho_2 \gamma_1 $$

$$ \gamma_2 = \rho_1 \gamma_1 + \rho_2 \gamma_0 $$

Solving the three equations for $\gamma_0, \gamma_1$ and $\gamma_2$ in terms of $\rho_1, \rho_2$ and $\sigma^2_u$ produces

$$ \text{plimb} = \frac{\beta + \rho}{1 + \beta \rho} $$

Therefore, least squares is inconsistent unless $\rho$ equals zero.

#### 12.5.2 Estimating the Variance of the Least Squares Estimator

Nonetheless, as a general proposition, one would normally not want to rely on $s^2(X'X)^{-1}$ as an estimator of the asymptotic covariance matrix of the least squares estimator.

In view of this situation, if one is going to use least squares, then it is desirable to have an appropriate estimator of the covariance matrix of the least squares estimator. If the form of the autocorrelation is known, then one can estimate the parameters of $\Omega$ directly and compute a consistent estimator.
Suppose that the form of the autocorrelation is unknown. Then, a direct estimator of $\Omega$ or $\Omega(\theta)$ is not available. The problem is estimation of

$$
\Sigma = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \rho_{|t-s|} x_t x'_s
$$

Following White’s suggestion for heteroscedasticity, Newey and West’s (1987a) robust, consistent estimator for autocorrelated disturbances with an unspecified structure is

$$
S_* = S_0 + \frac{1}{T} \sum_{t=1}^{T} \sum_{s=t+1}^{T} \left(1 - \frac{j}{L+1}\right) e_t e_{t-j} [x_t x'_{t-j} + x_{t-j} x'_t]
$$

The maximum lag $L$ must be determined in advance to be large enough that autocorrelations at lags longer than $L$ are small enough to ignore.

### 12.6 GMM Estimation

The GMM estimator in the regression model with autocorrelated disturbances is produces by the empirical moment equations

$$
\frac{1}{T} \sum_{t=1}^{T} x_t (y_t - x'_t \hat{\beta}_{GMM}) = \frac{1}{T} X' \hat{\epsilon}(\hat{\beta}_{GMM}) = \bar{m}(\hat{\beta}_{GMM}) = 0
$$

The estimator is obtained by minimizing

$$
q = \bar{m}(\hat{\beta}_{GMM}) W \bar{m}(\hat{\beta}_{GMM})
$$

where $W$ is a positive definite weighting matrix. The optimal weighting matrix would be

$$
W = \{\text{Asy.Var}[\sqrt{T} \bar{m}(\beta)]\}^{-1}
$$

which is the inverse of

$$\text{Asy.Var}[\sqrt{T} \bar{m}(\beta)] = \text{Asy.Var} \left[ \frac{1}{\sqrt{T}} \sum_{i=1}^{n} x_i \epsilon_i \right] = \text{plim}_{n \to \infty} \frac{1}{T} \sum_{j=1}^{T} \sum_{t=j+1}^{T} \sigma^2 \rho_{ts} x_t x'_s = \sigma^2 Q^*
$$

The optimal weighting matrix would be $[\sigma^2 Q^*]^{-1}$.

### 12.7 Testing for Autocorrelation

The available tests for autocorrelation are based on the principle that if the true disturbances are autocorrelated, then this fact can be detected through the autocorrelations of the least squares residuals. The simplest indicator is the slope in the artificial regression

$$
e_t = r e_{t-1} + v_t
$$

$$
e_t = y_t - x'_t b
$$

$$
r = \left( \sum_{i=2}^{T} e_t e_{t-1} \right) / \left( \sum_{i=1}^{T} e_i^2 \right)
$$

If there is autocorrelation, then the slope in this regression will be an estimator of $\rho = \text{Corr}[\epsilon_t, \epsilon_{t-1}]$. As a first approximation, treating the model above as a classical linear model and using a $t$ or $F$ (squared $t$) test to test the hypothesis is a valid way to proceed based on the Lagrange multiplier principle.
12.7.1 Lagrange Multiplier Test

The Breusch (1978)-Godfrey (1978) test is a Lagrange multiplier test of $H_0$: no autocorrelation versus $H_1: \epsilon_t = AR(P)$ or $\epsilon_t = MA(P)$. The test statistic is

$$LM = T \left( \frac{e'e}{e'e - 1} \right) X_0'(X_0'X_0)^{-1}X_0'e$$

where $X_0$ is the original $X$ matrix augmented by $P$ additional columns containing the lagged OLS residuals, $e_{t-1}, ..., e_{t-P}$. The test can be carried out simply by regressing the ordinary least squares residuals $e_t$ on $x_{t0}$ and referring $TR_0^2$ to the tabled critical value for the chi-squared distribution with $P$ degrees of freedom.

12.7.2 Box and Pierce’s Test and Ljung’s Refinement

An alternative test which is asymptotically equivalent to the LM test when the null hypothesis, $\rho = 0$, is true and when $X$ does not contain lagged values of $y$ is due to Box and Pierce (1970). The Q test is carried out by referring

$$Q = \sum_{j=1}^P r_j^2$$

where $r_j = (\sum_{t=j+1}^{T} e_t e_{t-j})/(\sum_{t=1}^{T} e_t^2)$ to the critical values of the chi-squared table with $P$ degrees of freedom. A refinement suggest by Ljung and Box (1979) is

$$Q' = T(T+2) \sum_{j=1}^P \frac{r_j^2}{T-j}$$

The essential difference between the Godfrey-Breusch and the Box-Pierce tests is the use of partial correlations in the former and simple correlations int he latter.

12.7.3 The Durbin-Watson Test

The Durbin-Watson statistic was the first formal procedure developed for testing for autocorrelation using the least squares residuals. The test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = 2(1 - r) - \frac{e_t^2 + e_{T}^2}{\sum_{t=1}^{T} e_t^2}$$

where $r$ is the same first order autocorrelation which underlies the preceding two statistics. If the sample is reasonably large, then the last term will be negligible, leaving $d \approx 2(1-r)$. The statistic takes this form because the authors were able to determine the exact distribution of this transformation of the autocorrelation and could provide tables of critical values. The one-sided test for $H_0: \rho = 0$ against $H_1: \rho > 0$ is carried out by comparing $d$ to values $d_L(T, K)$ and $d_U(T, K)$. If $d < d_L$ the null hypothesis is rejected; if $d > d_U$, the hypothesis is not rejected. If $d$ lies between $d_L$ and $d_U$, then no conclusion is drawn.
12.7.4 Testing in the Presence of A Lagged Dependent Variables

The Durbin-Watson test is not likely to be valid when there is a lagged dependent variable in the equation. Durbin (1970) derived a Lagrange multiplier test that is appropriate in the presence of a lagged dependent variable. The test may be carried out by referring

\[ h = r \sqrt{T/(1 - T s^2_c)} \]

where \( s^2_c \) is the estimated variance of the least squares regression coefficient on \( y_{t-1} \), to the standard normal tables.

12.8 Efficient Estimation When \( \Omega \) is Known

If the parameters of \( \Omega \) are known, then the GLS estimator

\[ \hat{\beta} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y) \]

and the estimate of its sampling variance,

\[ \text{Est. Var}[\hat{\beta}] = \hat{\sigma}^2(T-X'\Omega^{-1}X)^{-1} \]

where

\[ \hat{\sigma}^2 = \frac{(y-X\hat{\beta})'\Omega^{-1}(y-X\hat{\beta})}{T} \]

can be computed in one step. For the AR(1) case, data for the transformed model are

\[ y_* = \begin{bmatrix} \sqrt{1 - \rho^2}y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}, \quad X_* = \begin{bmatrix} \sqrt{1 - \rho^2}x_1 \\ x_2 - \rho x_1 \\ x_3 - \rho x_2 \\ \vdots \\ x_T - \rho x_{T-1} \end{bmatrix} \]

These transformations are variously labeled partial differences, quasi differences, or pseudodifferences.

The variance of the transformed disturbance is

\[ \text{Var}[\epsilon_t - \rho \epsilon_{t-1}] = \text{Var}[u_t] = \sigma_u^2 \]

The variance of the first disturbance is also \( \sigma_u^2 \). This can be estimated using \( (1 - \rho^2)\hat{\sigma}^2 \).

For the AR(2) model

\[ \epsilon_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + u_t \]

the transformed data for generalized least squares are obtained by

\[ z_{t+1} = \left[ (1 + \theta_2)(1 - \theta_2)^2 - \theta_1^2 \right]^{1/2} z_1, \]
\[ z_{t+2} = (1 - \theta_2)z_2 - \frac{\theta_1(1 - \theta_2)^{1/2}}{1 - \theta_2} z_1 \]
\[ z_{st} = z_t - \theta_1 z_{t-1} - \theta_2 z_{t-2}, \quad t > 2, \]

where \( z_t \) is used for \( y_t \) or \( x_t \). The transformation becomes progressively more complex for higher-order processes.
12.9 Estimation When Ω Is Unknown

12.9.1 AR(1) Disturbances

The most common procedure is to begin FGLS with a natural estimator of ρ, the autocorrelation of the residuals. Since b is consistent, we can use r. Others that have been suggested include Theil’s(1971) estimator, \( r((T - K)/(T - 1)) \) and Durbin’s(1970), the slope on \( y_{t-1} \) in a regression of \( y_t \) on \( y_{t-1} \), \( x_t \) and \( x_{t-1} \). The second step is FGLS based on estimation in above section. This is Prais and Winsten (1954) estimator. The Cochrane and Orcutt (1949) estimator omits the first observation.

Maximum likelihood estimator can be obtained by maximizing the log-likelihood with respect to \( β, σ_u^2 \) and \( ρ \). The log-likelihood function may by written

\[
\ln L = -\sum_{t=1}^{T} \frac{u_t^2}{2σ_u^2} + \frac{1}{2} \ln(1 - ρ^2) - \frac{T}{2}(\ln 2Π + \ln σ_u^2)
\]

where, as before, the first observation is computed differently from the others using

\[
y_* = \begin{bmatrix}
√{1 - ρ^2} y_1 \\
y_2 - ρy_1 \\
y_3 - ρy_2 \\
\vdots \\
y_T - ρy_{T-1}
\end{bmatrix}, \quad X_* = \begin{bmatrix}
√{1 - ρ^2} x_1 \\
x_2 - ρx_1 \\
x_3 - ρx_2 \\
\vdots \\
x_T - ρx_{T-1}
\end{bmatrix}
\]

For a given value of \( ρ \), the maximum likelihood estimators of \( β \) and \( σ_u^2 \) are the usual ones, GLS and the mean squared residual using the transformed data. The problem is estimation of \( ρ \). One possibility is to search the range \(-1 < ρ < 1\) for the value that with the implied estimates of the other parameters maximizes \( \ln L \). Beach and MacKinnon (1978a) argue that this way to do the search is very inefficient and have devised a much faster algorithm. Omitting the first observation and adding an approximation at the lower right corner produces the standard approximations to the asymptotic variances of the estimators.

12.9.2 AR(2) Disturbances

Maximum likelihood procedures for most other disturbance processes are exceedingly complex. Beach and MacKinnon (1978b) have derived an algorithm for AR(2) disturbances. For models of the form,

\[
ε_t = \theta_1 ε_{t-1} + \theta_2 ε_{t-2} + ... + \theta_p ε_{t-p} + u_t
\]

a simple approach for estimation of the autoregressive parameters is to use the following method: Regress \( ε_t \) on \( ε_{t-1}, ..., ε_{t-p} \), to obtain consistent estimates of the autoregressive parameters. With the estimates of \( ρ_1, ..., ρ_p \) in hand, the Cochrane-Orcutt estimator can be obtained. If the model is an AR(2), the full FGLS procedure can be used instead. The least squares computations for the transformed data provide the appropriate estimates of \( σ_u^2 \) and covariance matrix of \( \hat{β} \).
12.9.3 Estimation with A Lagged Dependent Variable

Hatanaka (1974, 1976) has devised an efficient two-step estimator. We consider estimation of the model

\[ y_t = x_t' \beta + \gamma y_{t-1} + \epsilon_t \]
\[ \epsilon_t = \rho \epsilon_{t-1} + u_t \]

To get to the second step of FGLS, we require a consistent estimator of the slope parameters. These estimates can be obtained using an IV estimator, where the column of \( Z \) corresponding to \( y_{t-1} \) is the only one that need be different from that of \( X \). An appropriate instrument can be obtained by using the fitted values in the regression of \( y_t \) on \( x_t \) and \( x_{t-1} \). The residuals from the IV regression are then used to construct

\[ \hat{\rho} = \frac{\sum_{t=3}^{T} \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum_{t=3}^{T} \hat{\epsilon}_t^2} \]

where

\[ \hat{\epsilon}_t = y_t - b'_{IV} x_t - c_{IV} y_{t-1} \]

FGLS estimates may now be computed by regressing \( y_{st} = y_t - \hat{\rho} y_{t-1} \) on

\[ x_{st} = x_t - \hat{\rho} x_{t-1}, \]
\[ y_{st} = y_{t-1} - \hat{\rho} y_{t-2}, \]
\[ \hat{\epsilon}_{t-1} = y_{t-1} - b'_{IV} x_{t-1} - c_{IV} y_{t-2}. \]

Let \( d \) be the coefficient on \( \hat{\epsilon}_{t-1} \) in this regression. The efficient estimator of \( \rho \) is

\[ \hat{\rho} = \hat{\rho} + d \]

12.10 Common Factors

The assumption that the correctly specified model is

\[ y_t = x_t' \beta + \epsilon_t, \quad \epsilon_t = \rho \epsilon_{t-1} + u_t, \quad t = 1, \ldots, T \]

implies the "reduced form",

\[ M_0 : y_t = \rho y_{t-1} + (x_t - \rho x_{t-1})' \beta + u_t, \quad t = 2, \ldots, T \]

where \( u_t \) is free from serial correlation. The second of these is actually a restriction on the model

\[ M_1 : y_t = \rho y_{t-1} + x_t' \beta + x_{t-1} \alpha + u_t, \quad t = 2, \ldots, T \]

The second model contains \( 2K + 1 \) parameters, but if the model is correct, then \( \alpha = -\rho \beta \) and there are only \( K+1 \) parameters and \( K \) restrictions. One might then test the restrictions of \( M_0 \) using an \( F \) test.
12.11 Forecasting in the Presence of Autocorrelation

For purposes of forecasting, we refer first to the transformed model,

\[ y_{st} = x_{st}' \beta + \epsilon_{st} \]

Suppose that the process generating \( \epsilon_t \) is an AR(1) and that \( \rho \) is known. The optimal forecast of \( y_{sT+1}^0 \), given \( x_{T+1}^0 \) and \( x_T \), is

\[ \hat{y}_{sT+1}^0 = x_{sT+1}' \hat{\beta} \]

Disassembling \( \hat{y}_{sT+1}^0 \), we find that

\[ \hat{y}_{T+1}^0 - \rho y_T = x_{T+1}' \hat{\beta} - \rho x_T' \hat{\beta} \]

or

\[ \hat{y}_{T+1}^0 = x_{T+1}' \hat{\beta} + \rho (y_T - x_T' \hat{\beta}) = x_{T+1}' \hat{\beta} + \rho \epsilon_T \]

Thus, we carry forward a proportion \( \rho \) of the estimated disturbance in the preceding period. This step can be justified by reference to

\[ E[\epsilon_{T+1}|\epsilon_T] = \rho \epsilon_T \]

Moving average models are somewhat simpler, as the autocorrelation lasts for only \( Q \) periods. For an MA(1) model, for the first postsample period,

\[ \hat{y}_{T+1}^0 = x_{T+1}' \hat{\beta} + \hat{\epsilon}_{T+1}, \]

where

\[ \hat{\epsilon}_{T+1} = \hat{\epsilon}_{T+1} - \lambda \hat{\epsilon}_{T+1} \]

Therefore, a forecast of \( \epsilon_{T+1} \) will use all previous residuals. One way to proceed is to accumulate \( \hat{\epsilon}_{T+1} \) from the recursion

\[ \hat{\epsilon}_t = \hat{\epsilon}_t + \lambda \hat{\epsilon}_{t-1} \]

with \( \hat{\epsilon}_{T+1} = \hat{\epsilon}_0 = 0 \) and \( \hat{\epsilon}_t = (y_t - x_t' \hat{\beta}) \). After the first postsample period,

\[ \hat{\epsilon}_{T+n} = \hat{\epsilon}_{T+n} - \lambda \hat{\epsilon}_{T+n-1} = 0 \]

If the parameters of the disturbance process are known, then the variances for the forecast errors can be computed. For an AR(1) disturbance, the estimated variance would be

\[ s_f^2 = \hat{\sigma}_\epsilon^2 + (x_t - \rho x_{t-1})'/\{\text{Est.Var}[\hat{\beta}]\}(x_t - \rho x_{t-1}) \]
Chapter 13

Models for Panel Data

13.1 Panel Data Models

Many recent studies have analyzed panel, or longitudinal, data sets.

The fundamental advantage of a panel data set over a cross section is that it will allow the researcher great flexibility in modeling differences in behavior across individuals. The basic framework for this discussion is a regression model of the form

$$y_{it} = x_{it}'\beta + z_{i}'\alpha + \epsilon_{it}$$

There are $K$ regressors in $x_{it}$, not including a constant term. The heterogeneity, or individual effect is $z_{i}'\alpha$ where $z_{i}$ contains a constant term and a set of individual or group specific variables, which may be observed, such as race, sex, location, and so on or unobserved, such as family specific characteristics, individual heterogeneity in skill or preferences, and so on, all of which are taken to be constant over time $t$. The various cases we will consider are:

1. Pooled Regression: If $z_{i}$ contains only a constant term, then ordinary least squares provides consistent and efficient estimates of the common $\alpha$ and the slope vector $\beta$.

2. Fixed Effects: if $z_{i}$ is unobserved, but correlated with $x_{it}$, then the least squares estimator of $\beta$ is biased and inconsistent as a consequence of an omitted variable. However, in this instance, the model

$$y_{it} = x_{it}'\beta + \alpha_{i} + \epsilon_{it}$$

where $\alpha_{i} = z_{i}'\alpha$, embodies all the observable effects and specifies an estimable conditional mean. This fixed effects approach takes $\alpha_{i}$ to be a group-specific constant term in the regression model. It would be noted that the term “fixed” as used here indicates that the term does not vary over time, not that it is nonstochastic, which need not be the case.

3. Random Effects: If the unobservable individual heterogeneity, however formulated, can be assumed to be uncorrelated with the included variables, then the model may be formulated as

$$y_{it} = x_{it}'\beta + E[z_{i}'\alpha] + \{z_{i}'\alpha - E[z_{i}'\alpha]\} + \epsilon_{it}$$

$$= x_{it}'\beta + \alpha + u_{i} + \epsilon_{it}$$
that is, a as a linear regression model with a compound disturbance that may be consistently, albeit inefficiently, estimated by least squares. This random effects approach specifies that \( u_i \) is a group specific random element, similar to \( \epsilon_{it} \) except that for each group, there is but a single draw that enters the regression identically in each period.

4. Random Parameters: The random effects model can be viewed as a regression model with a random constant term. With a sufficiently rich data set, we may extend this idea to a model in which the other coefficients vary randomly across individuals as well. The extension of the model might appear as

\[
y_{it} = x_{it}'(\beta + h_i) + (\alpha + u_i) + \epsilon_{it}
\]

where \( h_i \) is a random vector which induces the variation of the parameters across individuals.

5. Covariance Structure: In some settings, researchers have concluded that a preferable approach to modeling heterogeneity in the regression model is to layer it into the variation around the conditional mean, rather than in the placement of the mean.

13.2 Fixed Effects

This formulation of the model assumes that differences across units can be captured in differences in the constant term. Each \( \alpha_i \) is treated as an unknown parameter to be estimated. Let \( y_i \) and \( X_i \) be the \( T \) observations for the \( i \)th unit, \( i \) be a \( T \times 1 \) column of ones, and let \( \epsilon_i \) be associated \( T \times 1 \) vector of disturbances. Then,

\[
y_i = X_i\beta + i\alpha_i + \epsilon_i
\]

Collecting these terms gives

\[
y = [X \quad d_1 \quad d_2 \quad ... \quad d_n][\begin{array}{c} \beta \\ \alpha \end{array}] + \epsilon,
\]

where \( d_i \) is a dummy variable indicating the \( i \)th unit. Let the \( nT \times n \) matrix \( D = [d_1 \quad d_2 \quad ... \quad d_n] \). Then, assembling all \( nT \) rows gives

\[
y = X\beta + D\alpha + \epsilon
\]

This model is usually referred to as the least squares dummy variable (LSDV) model. We write the least squares estimator of \( \beta \) as

\[
b = [X'M_DX]^{-1}[X'M_Dy],
\]

where

\[
M_D = I - D(D'D)^{-1}D'
\]
13.2.1 Testing the Significance of the Group Effects

If we are interested in differences across groups, then we can test the hypothesis that the constant terms are all equal with an \( F \) test. Under the null hypothesis of equality, the efficient estimator is pooled least squares. The \( F \) ratio used for this test is

\[
F(n-1, nT - n - K) = \frac{(R^2_{LSDV} - R^2_{Pooled})/(n - 1)}{(1 - R^2_{LSDV})/(nT - n - K)}
\]

where \( LSDV \) indicates the dummy variable model and \( Pooled \) indicates the pooled or restricted model with only a single overall constant term.

13.2.2 The Within- And Between-Groups Estimators

We can formulated a pooled regression model in three ways. First, the original formulation is

\[
y_{it} = x'_{it}\beta + \alpha + \epsilon_{it}
\]

In terms of deviations from the group means,

\[
y_{it} - \bar{y}_i = (x'_{it} - \bar{x}_i)\beta + \epsilon_{it} - \bar{\epsilon}_i,
\]

while in terms of the group means,

\[
\bar{y}_i = \bar{x}'_i\beta + \alpha + \bar{\epsilon}_i
\]

All three are classical regression models, and in principle, all three could be estimated, at least consistently if not efficiently, by ordinary least squares.

In first case, the moments would accumulated variation about the overall means, \( \bar{y} \) and \( \bar{x} \), and we would use the total sums of squares and cross products,

\[
S_{xx}^{total} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x})(x_{it} - \bar{x})'
\quad \text{and} \quad
S_{xy}^{total} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x})(y_{it} - \bar{y})
\]

For second case, since the data are in deviations already, the means of \( (y_{it} - \bar{y}_i) \) and \( (x_{it} - \bar{x}_i) \) are zero. The moment matrices are within-groups sums of squares and cross products,

\[
S_{xx}^{within} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'
\quad \text{and} \quad
S_{xy}^{within} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)
\]

Finally, for the third case, the mean of group means is the overall mean. The moment matrices are the between-groups sums of squares and cross products—that is, the variation of the group means around the overall means:

\[
S_{xx}^{between} = \sum_{i=1}^{n} T(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'
\quad \text{and} \quad
S_{xy}^{between} = \sum_{i=1}^{n} T(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})
\]

It is easy to verify that

\[
S_{xx}^{total} = S_{xx}^{within} + S_{xx}^{between}
\quad \text{and} \quad
S_{xy}^{total} = S_{xy}^{within} + S_{xy}^{between}
\]
Therefore, there are three possible least squares estimators of $\beta$ corresponding to the decomposition. The least squares estimator is

$$b^{total} = [S_{xx}^{total}]^{-1}S_{xy}^{total} = [S_{xx}^{within} + S_{xx}^{between}]^{-1}[S_{xy}^{within} + S_{xy}^{between}]$$

The within-groups estimator is

$$b^{within} = [S_{xx}^{within}]^{-1}S_{xy}^{within}$$

This is the LSDV estimator computed earlier. An alternative estimator would be the between-groups estimator

$$b^{between} = [S_{xx}^{between}]^{-1}S_{xy}^{between}$$

sometimes called the group means estimator. From the preceding expressions

$$S_{xy}^{between} = S_{xx}^{within}b^{within}$$ and $$S_{xy}^{within} = S_{xx}^{between}b^{between}$$

We see that the least squares estimator is a matrix weighted average of the within- and between-groups estimators:

$$b^{total} = F^{within}b^{within} + F^{between}b^{between}$$

where

$$F^{within} = [S_{xx}^{within} + S_{xx}^{between}]^{-1}S_{xx}^{within} = I - F^{between}$$

The resemblance is more than passing; it can be shown that

$$F^{within} = \left\{ [\text{Asy. Var}(b^{within})]^{-1} + [\text{Asy. Var}(b^{between})]^{-1} \right\}^{-1} [\text{Asy. Var}(b^{within})]^{-1}$$

### 13.2.3 Fixed Time and Group Effects

The least squares dummy variable approach can be extended to include a time-specific effect as well. One way to formulated the extended model is simply to add the time effect, as in

$$y_{it} = x_{it}' \beta + \alpha_i + \gamma_t + \epsilon_{it}$$

This model is obtained from the preceding one by the inclusion of an additional $T - 1$ dummy variables. A symmetric form of the model is

$$y_{it} = x_{it}' \beta + \mu + \alpha_i + \gamma_t + \epsilon_{it}$$

where a full $n$ and $T$ effects are included, but the restrictions

$$\sum_i \alpha_i = \sum_i \gamma_t = 0$$

are imposed. Least squares estimates of the slopes in this model are obtained by regression of

$$y_{sit} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$$

on

$$x_{sit} = x_{it} - \bar{x}_i - \bar{x}_t + \bar{x}$$
where the period-specific and overall means are
\[
\bar{y}_{t} = \frac{1}{n} \sum_{i=1}^{n} y_{it} \quad \text{and} \quad \bar{y} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} y_{it}
\]
and likewise for \( \bar{x}_{t} \) and \( \bar{x} \). The overall constant and the dummy variable coefficients can then be recovered from the normal equations.

The estimated asymptotic covariance matrix for \( b \) is computed using the sums of squares and cross products of \( x_{it} \)
\[
s^{2} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - x_{it}'b - m - a_{i} - c_{t})^{2}}{nT - (n-1) - (T-1) - K - 1}
\]

### 13.2.4 Unbalanced Panels and Fixed Effects

Missing data are very common in panel data sets. These panels are called unbalanced panels. A modification to allow unequal group sizes is quite simple. First, the full sample size is \( \sum_{i=1}^{n} T_{i} \) instead of \( nT \), which calls for minor modifications in the computations of \( s^{2} \), \( \text{Var}[b] \), \( \text{Var}[a_{i}] \), and the \( F \) statistic. Second, group means must be based on \( T_{i} \), which varies across groups. The overall means for the regressors are
\[
\bar{\bar{x}} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T_{i}} x_{it}}{\sum_{i=1}^{n} T_{i}} = \frac{\sum_{i=1}^{n} T_{i} \bar{x}_{i}}{\sum_{i=1}^{n} T_{i}} = \sum_{i=1}^{n} f_{i} \bar{x}_{i}
\]
where \( f_{i} = T_{i}/(\sum_{i=1}^{n} T_{i}) \). The within groups moment matrix is
\[
\sum_{i=1}^{n} X_{i}'M_{i}^{0}X_{i} = \sum_{i=1}^{n} \left( \sum_{t=1}^{T} (x_{it} - \bar{x}_{i})(x_{it} - \bar{x}_{i})' \right)
\]
The other moments, \( S_{xy}^{\text{within}} \) and \( S_{yy}^{\text{within}} \), are computed likewise.

### 13.3 Random Effects

The fixed effects model allows the unobserved individual effects to be correlated with the included variables.

Consider, then a reformulation of the model
\[
y_{it} = x_{it}'\beta + (\alpha + u_{i}) + \epsilon_{i}
\]
where there are \( K \) regressors including a constant and now the single constant term is the mean of the unobserved heterogeneity, \( E[z_{i}'\alpha] \). The component \( u_{i} \) is the random heterogeneity specific to the \( i \)th observation and is constant through time; recall \( u_{i} = [z_{i}'\alpha - E(z_{i}'\alpha)] \). For these \( T \) observations, let
\[
\eta_{it} = \epsilon_{it} + u_{i}
\]
and
\[
\eta_{i} = [\eta_{1}, \eta_{2}, ..., \eta_{T}]'
\]
In view of this form of $\eta_{it}$, we have what is often called an “error components model”. For this model,

$$E[\eta_{it}^2 | X] = \sigma^2 + \sigma_u^2$$
$$E[\eta_{it}\eta_{is} | X] = \sigma_u^2, \quad t \neq s$$
$$E[\eta_{it}\eta_{is} | X] = 0, \quad \text{for all } t \text{ and } s \text{ if } i \neq j$$

### 13.3.1 Generalized Least Squares

The generalized least squares estimator of the slope parameters is

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y = \left(\sum_{i=1}^{n} X'_i\Omega^{-1}X_i\right)^{-1} \left(\sum_{i=1}^{n} X'_i\Omega^{-1}y_i\right)^{-1}$$

To compute this estimator, we will require $\Omega^{-1/2} = [I_n \otimes \Sigma]^{-1/2}$. We need only find $\Sigma^{-1/2}$, which is

$$\Sigma^{-1/2} = \frac{1}{\sigma_\epsilon^2} \left[ I - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_u^2}} \right]$$

where

$$\theta = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_u^2}}$$

For the data set as a whole, then, generalized least squares is computed by the regression of these partial deviations of $y_{it}$ on the same transformations of $x_{it}$.

It can be shown that the GLS estimator is, like the OLS estimator, a matrix weighted average of the within- and between-units estimators:

$$\hat{\beta} = \hat{F}_{\text{within}}\hat{b}_{\text{within}} + (I - \hat{F}_{\text{within}})\hat{b}_{\text{between}}$$

where now,

$$\hat{F}_{\text{within}} = [S_{xx}^{\text{within}} + \lambda S_{xx}^{\text{between}}]^{-1}S_{xx}^{\text{within}}$$

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_u^2} = (1 - \theta)^2$$

If $\lambda$ equals 1, then generalized least squares is identical to ordinary least squares. This situation would occur if $\sigma_u^2$ were zero, in which case a classical regression model would apply. If $\lambda$ equals zero, then the estimator is the dummy variable estimator we used in the fixed effects setting. These are two possibilities. If $\sigma_\epsilon^2$ were zero, then all variation across units would be due to the different $u_i$s, which, because they are constant across time, would be equivalent to the dummy variables we used in the fixed-effects model. The other case is $T \rightarrow \infty$.

Unbalanced panels add a layer of difficulty in the random effects model. The matrix $\Omega$ is no longer $I \otimes \Sigma$ because the diagonal blocks in $\Omega$ are of different sizes. There is also group wise heteroscedasticity, because the $i$th diagonal block in $\Omega^{-1/2}$ is

$$\Omega_i^{-1/2} = I_{T_i} - \frac{\theta_i}{T_i}t_i' T_i, \quad \theta_i = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T_i\sigma_u^2}}$$
13.3.2 Feasible Generalized Least Squares When $\Sigma$ is Unknown

A heuristic approach to estimation of the variance components is as follows:

$$y_{it} = x_{it}' \beta + \alpha + \epsilon_{it} + u_i$$

and

$$\bar{y}_i = \bar{x}_i' \beta + \bar{\alpha} + \bar{\epsilon}_i + u_i$$

Therefore, taking deviations from the group means removes the heterogeneity:

$$y_{it} - \bar{y}_i = [x_{it} - \bar{x}_i]' \beta + [\epsilon_{it} - \bar{\epsilon}_i]$$

Since

$$E \left[ \sum_{t=1}^{T} (\epsilon_{it} - \bar{\epsilon}_i)^2 \right] = (T-1)\sigma^2_\epsilon$$

if $\beta$ were observed, then an unbiased estimator of $\sigma^2_\epsilon$ based on $T$ observations in group $i$ would be

$$\hat{\sigma}^2_\epsilon(i) = \frac{\sum_{t=1}^{T} (\epsilon_{it} - \bar{\epsilon}_i)^2}{T - 1}$$

Since $\beta$ must be estimated, implies that the LSDV estimator is consistent, indeed, unbiased in general–we make the degrees of freedom correction and use the LSDV residuals in

$$s^2_e(i) = \frac{\sum_{t=1}^{T} (\epsilon_{it} - \bar{\epsilon}_i)^2}{T - K - 1}$$

We have $n$ such estimators, so we average them to obtain

$$s^2_e = \frac{1}{n} \sum_{i=1}^{n} s^2_e(i) = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\sum_{t=1}^{T} (\epsilon_{it} - \bar{\epsilon}_i)^2}{T - K - 1} \right] = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (\epsilon_{it} - \bar{\epsilon}_i)^2}{nT - nK - n}$$

We propose the unbiased estimator

$$\hat{\sigma}^2_\epsilon = s^2_{LSDV} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (\epsilon_{it} - \bar{\epsilon}_i)^2}{nT - n - K}$$

This is the variance estimator in the LSDV model appropriately corrected for degrees of freedom.

It remains to estimate $\sigma^2_u$. Return to the original model specification. In spite of the correlation across observations, this is a classical regression model in which the ordinary least squares slopes and variance estimators are both consistent and, in most cases, unbiased. Therefore, using the ordinary least squares residuals from the model with only a single overall constant, we have

$$\text{plims} s^2_{Pooled} = \lim_{n \to \infty} \frac{e'e}{nT - K - 1} = \sigma^2_\epsilon + \sigma^2_u$$

This provides the two estimators needed for the variance components; the second would be

$$\hat{\sigma}^2_u = s^2_{Pooled} - s^2_{LSDV}$$
13.3.3 Testing for Random Effects

Breusch and Pagan (1980) have devised a Lagrange multiplier test for the random effects model based on the OLS residuals. For

\[ H_0 : \sigma_u^2 = 0 \quad \text{(or Corr}[\eta_{it}, \eta_{is}] = 0) \]
\[ H_1 : \sigma_u^2 \neq 0 \]

the test statistic is

\[ LM = \frac{nT}{2(T-1)} \left[ \frac{\sum_{i=1}^{n} [\sum_{t=1}^{T} e_{it}]^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2 = \frac{nT}{2(T-1)} \left[ \frac{\sum_{i=1}^{n} (T \bar{e}_i)^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2 \]

Under the null hypothesis, LM is distributed as chi-squared with one degree of freedom.

13.3.4 Hausman’s Specification Test for the Random Effects Model

At various points, we have made the distinction between fixed and random effects models. The fixed effects approach has one considerable virtue. There is little justification for treating the individual effects as uncorrelated with the other regressors, as is assumed in the random effects model. The random effects treatment, therefore, may suffer from the inconsistency due to this correlation between the included variables and the random effect.

The specification test devised by Hausman (1978) is used to test for orthogonality of the random effects and the regressors. The test is based on the idea that under the hypothesis of no correlation, both OLS in the LSDV model and GLS are consistent, but OLS is inefficient, whereas under the alternative, OLS is consistent, but GLS is not. Therefore, under the null hypothesis, the two estimates should not differ systematically, and a test can be based on the difference. The other essential ingredient for the test is, the covariance matrix of the difference vector, \([b - \hat{\beta}]\):

\[ \text{Var}[b - \hat{\beta}] = \text{Var}[b] + \text{Var}[\hat{\beta}] - \text{Cov}[b, \hat{\beta}] - \text{Cov}[b, \hat{\beta}] \]

Hausman’s essential result is that the covariance of an efficient estimator with its difference from an inefficient estimator is zero, which implies that

\[ \text{Cov}[(b - \hat{\beta}), \hat{\beta}] = \text{Cov}[b, \hat{\beta}] - \text{Var}[\hat{\beta}] = 0 \]

or that

\[ \text{Cov}[b, \hat{\beta}] = \text{Var}[\hat{\beta}] \]

produces the required covariance matrix for the test,

\[ \text{Var}[b - \hat{\beta}] = \text{Var}[b] - \text{Var}[\hat{\beta}] = \Psi \]

The chi-squared test is based on the Wald criterion:

\[ W = \chi^2[K - 1] = (b - \hat{\beta})' \Psi^{-1} (b - \hat{\beta}) \]

Under the null hypothesis, \(W\) has a limiting chi-squared distribution with \(K - 1\) degrees of freedom.
13.4 Instrumental Variables Estimation of the Random Effects Model

Recall the original specification of the linear model for panel data

\[ y_{it} = x'_{it} \beta + z'_{i} \alpha + \epsilon_{it} \]

The random effects model is based on the assumption that the unobserved person specific effects, \( z_i \), are uncorrelated with the included variables, \( x_{it} \). However, the random effects treatment does allow the model to obtain observed time invariant characteristics. **Hausman and Taylor’s estimator** for the random effects model suggests a way to overcome the first of these while accommodating the second.

Their model is of the form:

\[ y_{it} = x'_{1it} \beta_1 + x'_{2it} \beta_2 + z'_{1i} \alpha_1 + z'_{2i} \alpha_2 + \epsilon_{it} + u_i \]

where \( \beta = (\beta_1', \beta_2')' \) and \( \alpha = (\alpha_1', \alpha_2')' \). Hausman and Taylor define four sets of observed variables in the model:

- \( x_{1it} \) is \( K_1 \) variables that are time varying and uncorrelated with \( u_i \)
- \( z_{1i} \) is \( L_1 \) variables that are time invariant and uncorrelated with \( u_i \)
- \( x_{2it} \) is \( K_2 \) variables that are time varying and correlated with \( u_i \)
- \( z_{2i} \) is \( L_2 \) variables that are time invariant and correlated with \( u_i \)

The assumptions about the random terms in the model are

\[
\begin{align*}
E[u_i] &= E[u_i | x_{1it}, z_{1i}] = 0 \quad \text{though} \quad E[u_i] = E[u_i | x_{2it}, z_{2i}] \neq 0 \\
\text{Var}[u_i | x_{1it}, z_{1i}, x_{2it}, z_{2i}] &= \sigma_u^2 \\
\text{Cov}[\epsilon_{it}, u_i | x_{1it}, z_{1i}, x_{2it}, z_{2i}] &= 0, \\
\text{Var}[\epsilon_{it} + u_i | x_{1it}, z_{1i}, x_{2it}, z_{2i}] &= \sigma^2 = \sigma_\epsilon^2 + \sigma_u^2 \\
\text{Corr}[\epsilon_{it} + u_i, \epsilon_{is} + u_i | x_{1it}, z_{1i}, x_{2it}, z_{2i}] &= \rho = \sigma_u^2 / \sigma^2
\end{align*}
\]

By construction, any OLS or GLS estimators of this model are inconsistent when the model contains variables that are correlated with the random effects. Hausman and Taylor have proposed an instrumental variables estimator that uses only the information within the model. The strategy for estimation is based on the following logic: First, by taking deviations from group means, we find that

\[ y_{it} - \bar{y}_i = (x_{1it} - \bar{x}_{1i})' \beta_1 + (x_{2it} - \bar{x}_{2i})' \beta_2 + \epsilon_{it} - \bar{\epsilon}_i \]

which implies that \( \beta \) can be consistently estimated by least squares, in spite of the correlation between \( x_2 \) and \( u \). Hausman and Taylor show that the group mean deviations can be used as \( (K_1 + K_2) \) instrumental variables for estimation of \( (\beta, \alpha) \). The authors show that the group means for \( x_1 \) can serve as these remaining instruments, and the model will be identified so long as \( K_1 \) is greater than or equal to \( L_2 \). The authors propose the following set of steps for consistent and efficient estimation:

1. Obtain the LSDV (fixed effects) estimator of \( \beta = (\beta_1, \beta_2)' \) based on \( x_1 \) and \( x_2 \).
2. Form the within groups residuals, \( \epsilon_{it} \), from the LSDV regression at step 1.
3. The residual variance in the regression in step 2 is a consistent estimator of \( \sigma^2 = \sigma_u^2 + \sigma^2 / T \). From this estimator and the estimator of \( \sigma^2 \) in step 1, we deduce an estimator of \( \sigma_u^2 = \sigma^2 - \sigma^2 / T \). We then form the weight for feasible GLS in this model by forming the estimate of

\[
\theta = \sqrt{\frac{\sigma^2}{\sigma^2 + T \sigma_u^2}}
\]

4. The final step is a weighted instrumental variable estimator. Let the full set of variables in the model be

\[
w'_{it} = (x_{1it}, x_{2it}, z_{1i}, z_{2i})
\]

Collect these \( nT \) observations in the rows of data matrix \( W \). The transformed variables for GLS are, as before when we first fit the random effects model,

\[
w^*_{it} = w_{it} - (1 - \hat{\theta}) \bar{w}_i \quad \text{and} \quad y^*_{it} = y_{it} - (1 - \hat{\theta}) \bar{y}_i
\]

where \( \hat{\theta} \) denotes the sample estimate of \( \theta \). The transformed data are collected in the rows data matrix \( W^* \) and in column vector \( y^* \). The instrumental variables are

\[
v'_{it} = [(x_{1it} - \bar{x}_{1i})', (x_{2it} - \bar{x}_{2i})', z_{1i}\bar{x}_{1i}]
\]

These are stacked in the rows of the \( nT \times (K_1 + K_2 + L_1 + K_1) \) matrix \( V \). The instrumental variable estimator would be

\[
(\hat{\beta}', \hat{\alpha}')_{IV} = [(W'^*V)(V'V)^{-1}(VW'^*y^*)]^{-1}[(W'^*V)(V'V)^{-1}(Vy^*)]
\]

### 13.5 GMM Estimation of Dynamic Panel Data Models

Panel data are well suited for examining dynamic effects, as in the first-order model

\[
y_{it} = x_{it}'\beta + \gamma y_{i,t-1} + \alpha_i + \epsilon_{it}
\]

where the set of right hand side variables, \( w_{it} \) now includes the lagged dependent variable, \( y_{i,t-1} \). Assume for now that \( T \geq K + 1 \) where \( K \) is the number of variables in \( x_{it} \). Then, from

\[
\hat{\delta} = \left[ \sum_{i=1}^{n} W_i' M^0 W_i \right]^{-1} \left[ \sum_{i=1}^{n} W_i' M^0 y_i \right]
\]

\[
= \left[ \sum_{i=1}^{n} W_i' M^0 W_i \right]^{-1} \left[ \sum_{i=1}^{n} W_i' M^0 W_i d_i \right]
\]

\[
= \sum_{i=1}^{n} F_i d_i
\]

where the rows of the \( T \times (K + 1) \) matrix \( W_i \) are \( w_{it}' \) and \( M^0 \) is the \( T \times T \) matrix that creates deviations from group means. Each group specific estimator, \( d_i \) is inconsistent, as it is biased in finite samples and its variance does not go to zero as \( n \) increases.
Neither of these results renders the model inestimable, but they do make necessary some technique other than our familiar LSDV or FGLS estimator. The general approach, which has been developed in several stages in the literature, relies on instrumental variables estimators and, most recently on a GMM estimator.

We extend the Hausman and Taylor (HT) formulation of the random effects model to include the lagged dependent variable;

\[ y_{it} = \gamma y_{i,t-1} + x_{1it}' \beta_1 + x_{2it}' \beta_2 + z_{1i}' \alpha_1 + z_{2i}' \alpha_2 + \epsilon_{it} + u_i \]

\[ = \delta' w_{it} + \epsilon_{it} + u_i \]

where

\[ w_{it}' = [y_{i,t-1}, x_{1it}', x_{2it}', z_{1i}', z_{2i}']' \]

is now a \((1 + K_1 + K_2 + L_1 + L_2) \times 1\) vector. Arellano et al. suggest a GMM estimator, and show that efficiency gains are available by using a larger set of moment conditions. The set of moment conditions we used to formulate the instrumental variables were

\[ E \left[ \begin{pmatrix} x_{1it} \\ x_{2it} \\ z_{1i} \\ \bar{x}_{ii} \end{pmatrix} (\eta_{it} - \bar{\eta}_i) \right] = E \left[ \begin{pmatrix} x_{1it} \\ x_{2it} \\ z_{1i} \\ \bar{x}_{ii} \end{pmatrix} (\epsilon_{it} - \bar{\epsilon}_i) \right] = 0 \]

This moment condition is used to produce the instrumental variable estimator. In order to proceed, as noted, we can include the lagged dependent variable in \(x_{2it}\). This set of instrumental variables can be used to construct the estimator, actually whether the lagged variable is present or not. Let

\[ W_i = \begin{bmatrix} w_{i1}' \\ w_{i2}' \\ \vdots \\ w_{iT_i}' \end{bmatrix} = \text{the full set of rhs data for group } i \text{ and } y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT_i} \end{bmatrix} \]

Note that \(W_i\) is assume to be a \(T \times (1 + K_1 + K_2 + L_1 + L_2)\) matrix. Since there is a lagged dependent variable in the model, it must be assumed that there are actually \(T + 1\) observations available on \(y_{it}\). We will form a matrix \(V_i\) consisting of \(T_i - 1\) rows constructed the same way for \(T_i - 1\) observations and a final row that will be different, as discussed below. The matrix will be of the form

\[ V_i = \begin{bmatrix} v_{i1}' & 0' & \ldots & 0' \\ 0' & v_{i2}' & \ldots & 0' \\ \vdots & \vdots & \ddots & \vdots \\ 0' & 0' & \ldots & a_i' \end{bmatrix} \]

The instrumental variable sets contained in \(v_{it}'\) which have been suggested might include the following form within the model:

\(x_{it}\) and \(x_{i,t-1}, x_{i1}, \ldots, x_{iT}, x_{i1}, \ldots, x_{iT}\)
The final row of $V_i$ is important to the construction. Two possibilities have been suggested:

\[
a'_i = [z'_{1i} \ x_{i1}] \text{(produces the Hausman and Taylor estimator)}
\]

\[
a'_i = [z'_{1i} \ x_{i1}, x_{i12}, ..., x_{i1T}] \text{(produces the Amemiya and MaCurdy’s estimator)}
\]

To construct the estimator, we will require a transformation matrix, $H$ constructed as follows. Let $M^{01}$ denote the first $T - 1$ rows of $M^0$, the matrix that creates deviations from group means. Then,

\[
H = \left( M^{01} \right)
\]

Thus, $H$ replaces the last row of $M^0$ with a row of $1/T$. The effect is as follows: if $q$ is $T$ observations on a variable, then $Hq$ produces $q^*$ in which the first $T - 1$ observations are converted to deviations from group means and the last observation is the group mean. In particular, let the $T \times 1$ column vector of disturbances

\[
\eta_i = [\eta_{i1}, \eta_{i2}, ..., \eta_{iT}] = [(\epsilon_{i1} + u_i), (\epsilon_{i2} + u_i), ..., (\epsilon_{iT} + u_i)],'
\]

then

\[
H = \begin{bmatrix}
\eta_{i1} - \bar{\eta}_i \\
\vdots \\
\eta_{iT-1} - \bar{\eta}_i \\
\bar{\eta}_i
\end{bmatrix}
\]

We can now construct the moment conditions

\[
E[V_i'H\eta_i] = E[g_i] = 0
\]

Write this as

\[
\text{plim} \frac{1}{n} \sum_{i=1}^{n} m_i = \text{plimm} = 0
\]

The GMM estimator $\hat{\delta}$ is then obtained by minimizing

\[
q = \bar{m}'A\bar{m}
\]

with an appropriate choice of the weighting matrix, $A$. The optimal weighting matrix will be the inverse of the asymptotic covariance matrix of $\sqrt{n}\hat{m}$. With a consistent estimator of $\delta$ in hand, this can be estimated empirically using

\[
\text{Est.Asy.Var}[\sqrt{n}\hat{m}] = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_i\hat{m}_i' = \frac{1}{n} \sum_{i=1}^{n} V_i'H\hat{\eta}_i\hat{\eta}_i'H'V_i
\]

use the residuals in

\[
\hat{\eta}_i = y_i - W_i\hat{\delta}
\]

to estimate $\sigma_u^2$ and $\sigma^2$ and then $\Sigma$, which produces

\[
\text{Est.Asy.Var}[\sqrt{n}\hat{m}] = \frac{1}{n} \sum_{i=1}^{n} V_i'\hat{\Sigma}\hat{\Sigma}'V_i
\]
We now have the full set of results needed to compute the GMM estimator. The solution to the optimization problem of minimizing \( q \) with respect to the parameter vector \( \delta \) is

\[
\hat{\delta}_{GMM} = \left[ \left( \sum_{i=1}^{n} W_i' HV_i \right) \left( \sum_{i=1}^{n} V_i' \hat{\Sigma} HV_i \right) \right]^{-1}
\]

\[
\times \left( \sum_{i=1}^{n} W_i' HV_i \right) \left( \sum_{i=1}^{n} V_i' \hat{\Sigma} HV_i \right) \left( \sum_{i=1}^{n} V_i' Hy_i \right)
\]

The estimator of the asymptotic covariance matrix for \( \hat{\delta} \) is the inverse matrix in brackets.

13.6 Nonspherical Disturbances and Robust Covariance Estimation

In this section, we will consider robust estimation of the asymptotic covariance matrix for least squares.

13.6.1 Robust Estimation of the Fixed Effects Model

In the fixed effects model, the full regressor matrix is \( Z = [X, D] \). The White heteroscedasticity consistent covariance matrix for OLS—that is, for the fixed effect estimator—is the lower right block of the partitioned matrix.

\[
\text{Est.Asy.Var}[b,a] = (Z'Z)^{-1}Z'E^2Z(Z'Z)^{-1}
\]

where \( E \) is a diagonal matrix of least squares (fixed effects estimator) residuals.

Arellano (1987) has taken this analysis a step further. If one takes the \( i \)th group as a whole, then we can treat the observations in

\[
y_i = X_i \beta + \alpha_i i_T + \epsilon_i
\]

as a generalized regression model with disturbance covariance matrix \( \Omega_i \). As before, let \( X_{is} \) denote the data in group mean deviation form. The counterpart to \( X'\Omega X \) here is

\[
X'_i \Omega X_s = \sum_{i=1}^{n} (X'_{is} \Omega_i X_{is})
\]

We can consider estimating \( \Omega_i \) with the sample of one, \( e_i e'_i \). As before, it is not consistent estimation of the individual \( \Omega_i \)s that is at issue, but estimation of the sum. If \( n \) is large enough, then we could argue that

\[
\text{plim} \frac{1}{nT} X'_i \Omega X_s = \text{plim} \frac{1}{nT} \sum_{i=1}^{n} X'_i \Omega_i X_{si}
\]

\[
= \text{plim} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T} X'_i e_i e'_i X_{si}
\]

\[
= \text{plim} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} e_i e_{is} x_{sit} x'_{sis} \right)
\]

The result is a combination of the White and Newey-West estimators.
13.6.2 Heteroscedasticity in the Random Effects Model

Since the random effects model is a generalized regression model with a known structure, the GLS estimator is efficient whereas the OLS estimator is not. Since pooled OLS is still consistent, OLS provides a usable set of residuals. Using the OLS residuals for the specific groups, we would have, for each group,

\[ \hat{\sigma}^2_{e_i} + u_i^2 = \frac{\epsilon'_i \epsilon_i}{T} \]

The residuals from the dummy variable model are purged of the individual specific effect, \(u_i\), so \(\hat{\sigma}^2_{e_i}\) may be consistently (in \(T\)) estimated with

\[ \hat{\sigma}^2_e = \frac{e'_i \text{LSDV}_i e_i \text{LSDV}_i}{T} \]

where \(e'_i \text{LSDV}_i = y_{it} - x'_i \text{b}_{\text{LSDV}} - a_i\). Combining terms, then,

\[ \hat{\sigma}^2_u = \frac{1}{n} \sum_{i=1}^{n} \left[ (\frac{e'_i \text{OLS}_i e_i \text{OLS}_i}{T} - \frac{e'_i \text{LSDV}_i e_i \text{LSDV}_i}{T}) \right] = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \]

We can now compute the FGLS estimator as before.

13.6.3 Autocorrelation in Panel Data Models

In the random effects model, as before, there are additional complications. The regression model is

\[ y_{it} = x'_i \beta + \alpha + \epsilon_{it} + u_i \]

If \(\epsilon_{it}\) is produced by an AR(1) process, \(\epsilon_{it} = \rho \epsilon_{i,t-1} + v_{it}\), then the familiar partial differencing procedure we used before would produce

\[ y_{it} - \rho y_{i,t-1} = \alpha (1 - \rho) + (x_{it} - \rho x_{i,t-1})' \beta + \epsilon_{it} - \rho \epsilon_{i,t-1} + u_i (1 - \rho) \]

Two issues remain: (1) how is the estimate of \(\rho\) obtained and (2) how does one treat the first observation? For the first of these, the first autocorrelation coefficient of the LSDV residuals is a simple expedient. The second question is more difficult. If the panel contains many groups (large \(n\)), then omitting the first observation is not likely to cause the inefficiency that it would in a single time series. One can apply the Prais-Winsten transformation to the first observation in each group instead [multiply by \((1 - \rho^2)^{1/2}\)], but then an additional complication arises at the second (FGLS) step when the observations are transformed a second time.

We could allow the autocorrelation to differ across groups. An estimate of each \(\rho_i\) is computable using the group mean mean deviation data.
Parameter heterogeneity across individuals or groups can be modeled as stochastic variation. Suppose that we write
\[ y_i = X_i \beta_i + \epsilon_i \]
where
\[ \beta_i = \beta + u_i, \]
and
\[ E[u_i|X_i] = 0, \]
\[ E[u_i u_j'|X_i] = \Gamma. \]
Assume for now that there is no autocorrelation or cross-sectional correlation. Thus, the \( \beta_i \) that applies to a particular cross-sectional unit is the outcome of a random process with mean vector \( \beta \) and covariance matrix \( \Gamma \). Expanding the result, we find that \( \Omega \) is a block diagonal matrix with
\[ \Omega_{ii} = E[(y_i - X_i \beta)(y_i - X_i \beta)'|X_i] = \sigma^2 I_T + X_i \Gamma X_i' \]
We can write the GLS estimator as
\[ \hat{\beta} = (X' \Omega^{-1} X)^{-1} (X' \Omega^{-1} y) = \sum_{i=1}^{n} W_i b_i \]
where
\[ W_i = \left[ \sum_{i=1}^{n} (\Gamma + \sigma^2_i (X_i'X_i)^{-1})^{-1} \right]^{-1} (\Gamma + \sigma^2_i (X_i'X_i)^{-1})^{-1} \]
Empirical implementation of this model requires an estimator of \( \Gamma \). One approach is to use the empirical variance of the set of \( n \) least squares estimates, \( b_i \) minus the average value of \( s_i^2 (X_i'X_i)^{-1} \). This matrix may not be positive definite, however, one might drop the second term. The more difficult obstacle is that panels are often short and there may be too few observations to compute \( b_i \).

Recent research in a number of fields have extended the random parameters model to a “multilevel” model or “hierarchical regression” model by allowing the means of the coefficients to vary with measured covariates. In this formulation,
\[ \beta_i = \beta + \Delta z_i + u_i \]
This model retains the earlier stochastic specification, but adds the measurement equation to the generation of the random parameters. The regression equation would now become
\[ y_i = X_i \beta + X_i \Delta z_i + (\epsilon_i + X_i u_i) \]
which can still be fit by least squares.
13.8 Covariance Structures For Pooled Time-Series Cross-Sectional Data

In the models we shall examine in this section, the data set consists of \( n \) cross-sectional units, denoted \( i = 1, \ldots, n \), observed at each of \( T \) time periods, \( t = 1, \ldots, T \). We have a total of \( nT \) observations. We will assume that \( n \) is fixed.

The framework for this analysis is the generalized regression model:

\[
y_{it} = x_{it}' \beta + \epsilon_{it}
\]

We have assumed that \( \beta_1 = \beta_2 = \cdots = \beta_n \). It is useful to stack the \( n \) time series,

\[
y_i = X_i \beta + \epsilon_i, \; i = 1, \ldots, n
\]

Each submatrix or subvector has \( T \) observations. We also specify

\[
E = [\epsilon_i | X] = 0
\]

and

\[
E[\epsilon_i \epsilon_j | X] = \sigma_{ij} \Omega_{ij}
\]

The terms above, we have the full specification,

\[
E[\epsilon | X] = 0
\]

and

\[
E[\epsilon \epsilon' | X] = \Omega = \begin{pmatrix}
\sigma_{11} \Omega_{11} & \sigma_{12} \Omega_{12} & \cdots & \sigma_{1n} \Omega_{1n} \\
\sigma_{21} \Omega_{21} & \sigma_{22} \Omega_{22} & \cdots & \sigma_{2n} \Omega_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} \Omega_{n1} & \sigma_{n2} \Omega_{n2} & \cdots & \sigma_{nn} \Omega_{nn}
\end{pmatrix}
\]

13.8.1 Generalized Least Squares Estimation

To begin, we assume that there is no correlation across periods, which implies that \( \Omega_{ij} = I \).

The generalized least squares estimator of \( \beta \) is based on a known \( \Omega \) would be

\[
\hat{\beta} = [X' \Omega^{-1} X]^{-1} [X' \Omega^{-1} y]
\]

The matrix \( \Omega \) can be written as

\[
\Omega = \Sigma \otimes I,
\]

where \( \Sigma \) is the \( n \times n \) matrix \([\sigma_{ij}]\). Then,

\[
\Omega^{-1} = \Sigma^{-1} \otimes I = \begin{pmatrix}
\sigma^{11} \Omega_{11} & \sigma^{12} \Omega_{12} & \cdots & \sigma^{1n} \Omega_{1n} \\
\sigma^{21} \Omega_{21} & \sigma^{22} \Omega_{22} & \cdots & \sigma^{2n} \Omega_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma^{n1} \Omega_{n1} & \sigma^{n2} \Omega_{n2} & \cdots & \sigma^{nn} \Omega_{nn}
\end{pmatrix}
\]

where \( \sigma^{ij} \) denotes the \( ij \)th element of \( \Sigma^{-1} \). This provides a specific form for the estimator,

\[
\hat{\beta} = \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma^{ij} X_i' X_j \right]^{-1} \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma^{ij} X_i' y_j \right]
\]

The asymptotic covariance matrix of the GLS estimator is the inverse matrix in brackets.
13.8.2 Feasible GLS Estimation

To compute the FGLS estimators for this mode, we require the full set of sample moments, \( y_j'y_j, X_j'X_j, \) and \( X_j'y_j \) for all pairs of cross-sectional units. With \( \hat{\sigma}_{ij} \) in hand, FGLS may be computed using

\[
\hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y
\]

There is an important consideration to note in feasible GLS estimation of this model. The computation requires inversion of the matrix \( \hat{\Sigma} \). This matrix is \( n \times n \). It is computed from the least squares residuals using

\[
\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} e_t e_t' = \frac{1}{T} E'E
\]

where \( e_t' \) is a \( 1 \times n \) vector containing all \( n \) residuals for the \( n \) groups at time \( t \), placed as the \( t \)th row of the \( T \times n \) matrix of residuals, \( E \). The rank of this matrix cannot be larger than \( T \).

13.8.3 Heteroscedasticity and the Classical Model

Two special cases of this model are of interest. The groupwise heteroscedastic model results if the off diagonal terms in \( \Sigma \) all equal zero. Then,

\[
\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y = \left[ \sum_{i=1}^{n} \frac{1}{\sigma_i^2} X_i'y_i \right]^{-1} \left[ \sum_{i=1}^{n} \frac{1}{\sigma_i^2} X_i'y_i \right]
\]

\( \sigma_i^2 \), are known, so the two-step FGLS method noted earlier, now based only on the diagonal elements of \( \Sigma \) would be used. The second special case is the classical regression model, which adds the further restriction \( \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 \). We would now stack the data in the pooled regression model in

\[
y = X\beta + \epsilon
\]

For this model, the GLS estimator reduces to pooled ordinary least squares.

13.8.4 Specification Tests

Under the null hypothesis of a common variance, \( \sigma^2 \) the Wald statistic for testing the null hypothesis against the alternative of the group wise heteroscedasticity model would be

\[
W = \sum_{i=1}^{n} \frac{(\hat{\sigma}_i^2 - \sigma^2)^2}{\text{Var}[\hat{\sigma}_i^2]}
\]

If the null hypothesis is correct,

\[
W \overset{d}{\longrightarrow} \chi^2[n]
\]

By hypothesis,

\[
\text{plim} \hat{\sigma}^2 = \sigma^2
\]

where \( \hat{\sigma}^2 \) is the disturbance variance estimator from the pooled OLS regression. We may estimate \( \text{Var}[\hat{\sigma}_i^2] \) with

\[
f_{ii} = \frac{1}{T} \frac{1}{T-1} \sum_{t=1}^{T} (e_{it}^2 - \hat{\sigma}_i^2)^2
\]

98
The modified Wald statistic is then

\[ W' = \sum_{i=1}^{n} \frac{(\hat{\sigma}_i^2 - \sigma_i^2)^2}{f_{ii}} \]

White’s general test is an alternative. To use White’s test, we would regress the squared OLS residuals on the \( P \) unique variables in \( x \) and the squares and cross products, including a constant. The chi-squared statistic, which has \( P - 1 \) degrees of freedom, is \((nT)R^2\).

### 13.8.5 Autocorrelation

The preceding discussion dealt with heteroscedasticity and cross-sectional correlation. It is simplest to begin with the assumption that

\[ \text{Corr}[\epsilon_{it}, \epsilon_{js}] = 0, \quad \text{if} \quad i \neq j. \]

That is, the disturbances between cross-sectional units are uncorrelated. Then,

\[ \epsilon_{it} = \rho_i \epsilon_{i,t-1} + u_{it}, \]

\[ \text{Var}[\epsilon_{it}] = \sigma_i^2 = \frac{\sigma_{ui}^2}{1 - \rho_i^2} \]

Then, if we take each time series \([y_i, X_i]\) separately, we can transform the data using the Prais-Winsten transformation:

\[
y_{si} = \begin{bmatrix} \sqrt{1 - r_i^2} y_{i1} \\ y_{i2} - r_i y_{i1} \\ y_{i3} - r_i y_{i2} \\ \vdots \\ y_{iT} - r_i y_{iT-1} \end{bmatrix}, \quad X_{si} = \begin{bmatrix} \sqrt{1 - r_i^2} x_{i1} \\ x_{i2} - r_i x_{i1} \\ x_{i3} - r_i x_{i2} \\ \vdots \\ x_{iT} - r_i x_{iT-1} \end{bmatrix}
\]

In terms of the transformed data \(y_{si}\) and \(X_{si}\), the model is now only heteroscedastic; the transformation has removed the autocorrelation. We may now use weighted least squares, as described earlier. This requires a second least squares estimate. The first, OLS regression produces initial estimates of \(\rho_i\). The transformed data are then used in a second least squares regression to obtain consistent estimators,

\[ \hat{\sigma}_{ui}^2 = \frac{e_{si}'e_{si}}{T} = \frac{(y_{si} - X_{si}\hat{\beta})'(y_{si} - X_{si}\hat{\beta})}{T} \]

At the end of the calculation, the moment matrix used in the last regression gives the correct asymptotic covariance matrix for the estimator, now \(\hat{\beta}\). If desired, then a consistent estimator of \(\sigma_{ei}^2\) is

\[ \hat{\sigma}_{ei}^2 = \frac{\sigma_{ui}^2}{1 - r_i^2} \]

The remaining question is how to obtain the initial estimates \(r_i\). The natural choice would be

\[ r_i = \frac{\sum_{t=2}^{T} e_{it}e_{i,t-1}}{\sum_{t=1}^{T} e_{it}^2} \]
One which is analogous to that used in the single equation case is
\[ r = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T} e_{it} e_{i,t-1}}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} \]

Finally, one may wish to allow for cross-sectional correlation across units. The preceding
has a natural generalization. If we assume that
\[ \text{Cov}[u_{it}, u_{jt}] = \sigma_{ij} \]
then we obtain the original model in which the off-diagonal blocks of \( \Omega \), are
\[ \sigma_{ij} \Omega_{ij} = \frac{\sigma_{uij}}{1 - \rho_i \rho_j} \]

Initial estimates of \( \rho_i \) are required, as before. Estimates of \( \sigma_{eij} \) can be obtained from the
least squares residual covariances obtained from the transformed data:
\[ \tilde{\sigma}_{eij} = \frac{\hat{\epsilon}_{uij}}{1 - r_i r_j} \]

where \( \hat{\epsilon}_{uij} = e_{i}^{t} e_{s}^{j} / T \).

### 13.8.6 Maximum Likelihood Estimation

Consider the general model with groupwise heteroscedasticity and cross group correlation. We now assume that the \( n \) disturbances at time \( t \), \( \epsilon_t \) have a multivariate normal distribution with zero mean and this \( n \times n \) covariance matrix. Taking logs and summing over the \( T \) periods gives the log-likelihood for the sample,
\[ \ln L(\beta, \Sigma | \text{data}) = -\frac{nT}{2} \ln 2\pi - \frac{T}{2} \ln |\Sigma| - \frac{1}{2} \sum_{t=1}^{T} \epsilon_t' \Sigma^{-1} \epsilon_t \]
\[ \epsilon_{it} = y_{it} - x_{it}' \beta, \quad i = 1, ..., n \]

The result is that the maximum likelihood estimator of \( \beta \) is the generalized least squares estimator. The maximum likelihood estimator of \( \Sigma \) is
\[ \hat{\sigma}_{ij} = \frac{(y_i' - X_i \hat{\beta}_h)'(y_i' - X_i \hat{\beta}_h)}{T} = \frac{\hat{\epsilon}_i' \hat{\epsilon}_j}{T} \]

based on the MLE of \( \beta \). Hypothesis test about \( \beta \) may be done using the familiar Wald statistic.

For testing the hypothesis that the off-diagonal elements of \( \Sigma \) are zero—that is, that there is no correlation across firms–there are three approaches. The likelihood ratio test is based on the statistic
\[ \lambda_{LR} = T(\ln |\hat{\Sigma}_{\text{heteroscedastic}}| - \ln |\hat{\Sigma}_{\text{general}}|) = T \left( \sum_{i=1}^{n} \ln \hat{\sigma}_{ii}^2 - \ln |\hat{\Sigma}| \right) \]
where $\hat{\sigma}_i^2$ are the estimates of $\sigma_i^2$ obtained from the maximum likelihood estimates of the groupwise heteroscedastic model and $\bar{\Sigma}$ is the maximum likelihood estimator in the unrestricted model. The large-sample distribution of the statistic is chi-squared with $n(n-1)/2$ degrees of freedom. The Lagrange multiplier test developed by Breusch and Pagan (1980) provides an alternative. The general form of the statistic is

$$\lambda_{LM} = T \sum_{i=2}^{n} \sum_{j=1}^{i-1} r_{ij}^2$$

where $r_{ij}^2$ is the $ij$th residual correlation coefficient.

For the groupwise heteroscedasticity model, ML estimation reduces to groupwise weighted least squares. The maximum likelihood estimator of $\beta$ is feasible GLS.

For testing the heteroscedasticity assumption of the model, the full set of test strategies that we have used before is available. The Lagrange multiplier test is probably the most convenient test, since it does not require another regression after the pooled last squares regression. It is convenient to rewrite

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{T}{2\sigma^2} \left[ \frac{\hat{\sigma}_i^2}{\sigma^2} - 1 \right]$$

where $\hat{\sigma}_i^2$ is the $i$th unit-specific estimate of $\sigma_i^2$ based on the true disturbances. Under the null hypothesis $\sigma_i^2 = \sigma^2, i = 1, ..., n$ the first derivative of the log-likelihood function with respect to this common $\sigma^2$ is

$$\frac{\partial \log L_R}{\partial \sigma^2} = -\frac{nT}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \epsilon'_i \epsilon_i$$

Equating this derivative to zero produces the restricted maximum likelihood estimator

$$\hat{\sigma}^2 = \frac{1}{nT} \sum_{i=1}^{n} \epsilon'_i \epsilon_i = \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}_i^2$$

which is the simple average of the $n$ individual consistent estimators. Using the least squares residuals at the restricted solution, we obtain $\hat{\sigma}^2 = (1/nT) \epsilon' \epsilon$ and $\sigma_i^2 = (1/T) \epsilon'_i \epsilon_i$, the Lagrange multiplier statistic reduces to

$$\lambda_{LM} = \sum_{i=1}^{n} \left[ \frac{T}{2\sigma^2} \left( \frac{\sigma_i^2}{\sigma^2} - 1 \right) \right] \left( \frac{2\sigma^4}{T} \right) = \frac{T}{2} \sum_{i=1}^{n} \left( \frac{\sigma_i^2}{\sigma^2} - 1 \right)^2$$

The statistic has $n - 1$ degrees of freedom. If we assume normality, then the asymptotic variance of each variance estimator is $2\sigma_i^4/T$ and the variances are asymptotically uncorrelated. Therefore, the Wald statistic to test the hypothesis of a common variance $\sigma^2$, using $\hat{\sigma}_i^2$ to estimate $\sigma_i^2$, is

$$W = \sum_{i=1}^{n} (\hat{\sigma}_i^2 - \sigma^2)^2 \left( \frac{2\sigma_i^4}{T} \right)^{-1} = \frac{T}{2} \sum_{i=1}^{n} \left( \frac{\hat{\sigma}_i^2}{\sigma^2} - 1 \right)^2$$

Note the similarity to the Lagrange multiplier statistic.
We can also carry out a likelihood ratio test. The appropriate statistic is

$$\lambda_{LR} = T(\ln |\hat{\Sigma}_{\text{homoscedastic}}| - \ln |\hat{\Sigma}_{\text{heteroscedastic}}|) = (nT) \ln \hat{\sigma}^2 - \sum_{i=1}^{n} T \ln \hat{\sigma}_i^2$$

where

$$\hat{\sigma}^2 = \frac{e'e}{nT} \quad \text{and} \quad \sigma_i^2 = \frac{\hat{\epsilon}_i^2}{T}$$

with all residuals computed using the maximum likelihood estimators. This chi-squared statistic has $n - 1$ degrees of freedom.
Chapter 14

Systems of Regression Equations

14.1 Introduction

It makes sense to consider the several models jointly.

14.2 The Seemingly Unrelated Regressions Model

The seemingly unrelated regressions (SUR) model is

\[ y_i = X_i \beta + \epsilon_i, \quad i = 1, ..., M \]

where

\[ \epsilon = [\epsilon'_1, \epsilon'_2, ..., \epsilon'_M]' \]

and

\[
E[\epsilon|X_1, X_2, ..., X_M] = 0, \\
E[\epsilon'\epsilon'|X_1, X_2, ..., X_M] = \Omega
\]

We assume that a total of \( T \) observations are used in estimating the parameters of the \( M \) equations. Each equation involves \( K_m \) regressors, for a total of \( K = \sum_{i=1}^{n} K_i \). We will require \( T > K \). We also assume that disturbances are uncorrelated across observations.

\[
E[\epsilon_i \epsilon_j | X_1, X_2, ..., X_M] = \sigma_{ij}, \quad \text{if } t = s \text{ and } 0 \text{ otherwise.}
\]

The disturbance formulation is

\[
E[\epsilon_i \epsilon'_j | X_1, X_2, ..., X_M] = \sigma_{ij} I_T
\]

or

\[
E[\epsilon' \epsilon | X_1, X_2, ..., X_M] = \Omega = \begin{bmatrix}
\sigma_{11} I_{n1} & \sigma_{12} I_{n2} & \cdots & \sigma_{1n} I_{nn} \\
\sigma_{21} I_{n1} & \sigma_{22} I_{n2} & \cdots & \sigma_{2n} I_{nn} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} I_{n1} & \sigma_{n2} I_{n2} & \cdots & \sigma_{nn} I_{nn}
\end{bmatrix}
\]
14.2.1 Generalized Least Squares

The generalized regression model applies to the stacked model,

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_M
\end{bmatrix}
= \begin{bmatrix}
  X_1 & 0 & \cdots & 0 \\
  0 & X_2 & \cdots & 0 \\
  \vdots \\
  0 & 0 & \cdots & X_M
\end{bmatrix}
\begin{bmatrix}
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_M
\end{bmatrix}
+ \begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_M
\end{bmatrix}
= X\beta + \epsilon
\]

Therefore, the efficient estimator is generalized least squares. For the \( t \)th observation, the \( M \times M \) covariance matrix of the disturbances is

\[
\Sigma = \begin{bmatrix}
  \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\
  \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\
  \vdots \\
  \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM}
\end{bmatrix}
\]

so,

\[
\Omega = \Sigma \otimes I
\]

and

\[
\Omega^{-1} = \Sigma^{-1} \otimes I
\]

Denoting the \( ij \)th element of \( \Sigma^{-1} \) by \( \sigma^{ij} \), we find that the GLS estimator is

\[
\hat{\beta} = [X'\Omega^{-1}X]^{-1}X'y = [X'\Omega^{-1}X]^{-1}X'(\Sigma^{-1} \otimes I)y
\]

14.2.2 Seemingly Unrelated Regressions with Identical Regressors

The case of identical regressors is quite common. In this special case, generalized least squares is equivalent to equation by equation ordinary least squares. Impose the assumption that \( X_i = X_j = X \), so that \( X_iX_i = X'X \) for all \( i \) and \( j \). The inverse matrix on the right-hand side now becomes \( [\Sigma^{-1} \otimes X'X]^{-1} \), which equals \( [\Sigma \otimes (X'X)^{-1}] \). Also on the right-hand side, each term \( X_i'Y_j \) equals \( X'Y_j \), which, in turn equals \( X'Xb_j \). We obtain

\[
\hat{\beta} = \begin{bmatrix}
  \sigma_{11}(X'X)^{-1} & \sigma_{12}(X'X)^{-1} & \cdots & \sigma_{1M}(X'X)^{-1} \\
  \sigma_{21}(X'X)^{-1} & \sigma_{22}(X'X)^{-1} & \cdots & \sigma_{2M}(X'X)^{-1) \\
  \vdots \\
  \sigma_{M1}(X'X)^{-1} & \sigma_{M2}(X'X)^{-1} & \cdots & \sigma_{MM}(X'X)^{-1}
\end{bmatrix}
\begin{bmatrix}
  (X'X)\sum_{l=1}^{M} \sigma^{1l}b_l \\
  (X'X)\sum_{l=1}^{M} \sigma^{2l}b_l \\
  \vdots \\
  (X'X)\sum_{l=1}^{M} \sigma^{Ml}b_l
\end{bmatrix}
\]

Now, we isolate one of the subvectors, say the first, from \( \hat{\beta} \).

\[
\hat{\beta} = \sum_{j=1}^{M} \sigma_{1j} \sum_{l=1}^{M} \sigma^{lj}b_l = b_1 \left( \sum_{j=1}^{M} \sigma_{1j}\sigma^{j1} \right) + b_2 \left( \sum_{j=1}^{M} \sigma_{1j}\sigma^{j2} \right) + \ldots + b_M \left( \sum_{j=1}^{M} \sigma_{1j}\sigma^{jM} \right)
\]

The terms in parentheses are the elements of the first row of \( \Sigma \Sigma^{-1} = I \), so the end result is \( \hat{\beta}_1 = b_1 \). For the remaining subvectors, which are obtained the same way, \( \hat{\beta}_i = b_i \), which is the result we sought.
14.2.3 Feasible Generalized Least Squares

The preceding discussion assumes that $\Sigma$ is known. The least squares residuals may be used to estimate consistently the element of $\Sigma$ with

$$\hat{\sigma}_{ij} = s_{ij} = \frac{e_i'e_j}{T}$$

The consistency of $s_{ij}$ follows from that of $b_i$ and $b_j$. A degrees of freedom correction in the divisor is occasionally suggested. Two possibilities are

$$s^*_{ij} = \frac{e_i'e_j}{[(T - K_i)(T - K_j)]^{1/2}}$$

and

$$s^{**}_{ij} = \frac{e_i'e_j}{T - \max(K_i, K_j)}$$

The second is unbiased only if $i$ equals $j$ or $K_i$ equals $K_j$, whereas the first is unbiased only if $i$ equals $j$.

Goodness-of-fit measures for the system have been devised. For instance, McElroy (1977) suggested the systemwide measure

$$R_*^2 = 1 - \frac{\hat{\epsilon}'\hat{\Omega}^{-1}\hat{\epsilon}}{\sum_{i=1}^M \sum_{j=1}^M \hat{\sigma}_{ij} \left[ \sum_{t=1}^T (y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j) \right]} = 1 - \frac{M}{\text{tr}(\hat{\Sigma}^{-1}S_{yy})}$$

where $\hat{\cdot}$ indicates the FGLS estimate.

For testing a hypothesis about $\beta$, a statistic analogous to the $F$ ratio in multiple regression analysis is

$$F[J, MT - K] = \frac{(R\hat{\beta} - q)'[R(X'\hat{\Omega}^{-1}X)^{-1}R'](R\hat{\beta} - q)/J}{\hat{\epsilon}'\hat{\Omega}^{-1}\hat{\epsilon}/(MT - K)}$$

The computation requires the unknown $\Omega$. If we insert the FGLS estimate $\hat{\Omega}$ and use the result that the denominator converges to one, then, in large samples, the statistic will behave the same as

$$\hat{F} = \frac{1}{J} (R\hat{\beta} - q)'[R\text{Var}^{\hat{\beta}}R'](R\hat{\beta} - q)$$

This can be referred to the standard $F$ table.

An alternative test statistic that has a limiting chi-squared distribution with $J$ degrees of freedom when the hypothesis is true is

$$JF = (R\hat{\beta} - q)'[R\text{Var}^{\hat{\beta}}R'](R\hat{\beta} - q)$$

This can be recognized as a Wald statistic that measures the distance between $R\hat{\beta}$ and $q$.

14.2.4 Maximum Likelihood Estimation

We wish to arrange this observation horizontally instead of vertically.

$$[y_1, y_2, ..., y_M]_t = [x_1^*]'[\pi_1, \pi_2, ..., \pi_M] + [\epsilon_1, \epsilon_2, ..., \epsilon_M]_t = [x_1^*]'\Pi' + E,$$

where $x_1^*$ is the full set of all $K^*$ different independent variables that appear in the model. The parameter matrix then has one column for each equation, but the columns are not the
same as $\beta_i$ unless every variable happens to appear in every equation. Otherwise, in the $i$th equation, $\pi_i$ will have a number of zeros in it, each one imposing an exclusion restriction.

This vector is one observation. Let $\epsilon_t$ be the vector of $M$ disturbances for this observation arranged, for now, in a column. Then $E[\epsilon_t \epsilon_t'] = \Sigma$. The log of the joint normal density of these $M$ disturbances is

$$\log L_t = -\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \epsilon_t' \Sigma^{-1} \epsilon_t$$

The log-likelihood for a sample of $T$ joint observations is the sum of these over $t$:

$$\log L = \sum_{t=1}^{T} \log L_t = -\frac{MT}{2} \log(2\pi) - \frac{T}{2} \log |\Sigma| - \frac{1}{2} \sum_{t=1}^{T} \epsilon_t' \Sigma^{-1} \epsilon_t$$

We can always permute the matrices in a trace, so

$$\sum_{t=1}^{T} \epsilon_t' \Sigma^{-1} \epsilon_t = \sum_{t=1}^{T} \text{tr}(\epsilon_t' \Sigma^{-1} \epsilon_t) = \sum_{t=1}^{T} \text{tr}(\Sigma^{-1} \epsilon_t' \epsilon_t)$$

Combining all three steps, we obtain

$$\sum_{t=1}^{T} \text{tr}(\Sigma^{-1} \epsilon_t' \epsilon_t) = T \text{tr}\left[\Sigma^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \epsilon_t \epsilon_t'\right)\right] = T \text{tr}(\Sigma^{-1} W)$$

where

$$W_{ij} = \frac{1}{T} \sum_{t=1}^{T} \epsilon_{ti} \epsilon_{tj}$$

Inserting this result in the log-likelihood, we have

$$\log L = -\frac{T}{2} [M \log(2\pi) + \log |\Sigma| + \text{tr}(\Sigma^{-1} W)]$$

We now consider maximizing this function. It has been shown that

$$\frac{\partial \log L}{\partial \Pi'} = \frac{T}{2} X^* E \Sigma^{-1}$$

$$\frac{\partial \log L}{\partial \Sigma'} = -\frac{T}{2} \Sigma^{-1} (\Sigma - W) \Sigma^{-1}$$

where the $x^*_t$ is row $t$ of $X^*$. If we insert our solution for $\Sigma$ in the likelihood function, then we obtain the concentrated log-likelihood,

$$\log L_c = -\frac{T}{2} [M(1 + \log(2\pi)) + \log |W|]$$

We have shown, therefore, that the criterion for choosing the maximum likelihood estimator of $\beta$ is

$$\hat{\beta}_{ML} = \min_{\beta} \frac{1}{2} \log |W|$$

The likelihood ratio statistic is an alternative to the $F$ statistic discussed earlier for testing hypotheses about $\beta$. The likelihood ratio statistic is

$$\lambda = -2(\log L_r - \log L_u) = T(\log |\hat{W}_r| - \log |\hat{W}_u|)$$

106
where ˆ\(W_r\) and ˆ\(W_u\) are the residual sums of squares and cross-product matrices using the constrained and unconstrained estimators, respectively. The likelihood ratio statistic is asymptotically distributed as chi-squared with degrees of freedom equal to the number of restrictions.

It may also be of interest to test whether Σ is a diagonal matrix. The likelihood ratio statistic would be

\[ \lambda_{LR} = T \left( \sum_{i=1}^{M} \log \hat{\sigma}_i^2 - \log |\hat{\Sigma}| \right) \]

where \(\hat{\sigma}_i^2\) is \(e'_i e_i / T\) from the individual least squares regressions and \(\hat{\Sigma}\) is the maximum likelihood estimator of Σ. This statistic has a limiting chi-squared distribution with \(M(M - 1)/2\) degrees of freedom under the hypothesis. The alternative suggested by Breusch and Pagan (1980) is the Lagrange multiplier statistic,

\[ \lambda_{LM} = T \sum_{i=2}^{T} \sum_{j=1}^{i-1} r_{ij}^2 \]

where \(r_{ij}\) is the estimated correlation \(\hat{\sigma}_{ij}/[\hat{\sigma}_{ii}\hat{\sigma}_{jj}]^{1/2}\). This statistic also has a limiting chi-squared distribution with \(M(M - 1)/2\) degrees of freedom. This test has the advantage that it does not require computation of the maximum likelihood estimator of Σ, since it is based on the OLS residuals.

14.2.5 An Application From Financial Econometrics: the Capital Asset Pricing Model

The capital asset pricing model (CAPM) is one of the foundations of that field and is a frequent subject of econometric analysis.

The Sharpe and Lintner analyses produce the following model for the following model for the expected excess return from an asset \(i\):

\[ E[R_i] - R_f = \beta_i (E[R_m] - R_f) \]

where \(R_i\) is the return on asset \(i\), \(R_f\) is the return on a “risk-free” asset, \(R_m\) is the return on the market’s optimal portfolio, and \(\beta_i\) is the asset market “beta”

\[ \beta_i = \frac{\text{Cov}[R_i, R_m]}{\text{Var}[R_m]} \]

Black (1972) considered the more general case in which there is no risk-free asset. In this instance, the observed \(R_f\) is replaced by the unobservable return on a “zero-beta” portfolio, \(E[R_0] = \gamma\).

The empirical counterpart to the Sharpe and Lintner model is a seemingly unrelated regressions (SUR) model, which we cast in the form

\[
[y_1, y_2, ..., y_N] = [1, z_t] \begin{bmatrix}
\alpha_1 & \alpha_2 & \cdots & \alpha_N \\
\beta_1 & \beta_2 & \cdots & \beta_N 
\end{bmatrix} + [\epsilon_1, \epsilon_2, ..., \epsilon_N]_t = x_t' \Pi + \epsilon_t
\]

where \(y_{it}\) is \(R_{it} - R_{ft}\), the observed excess return on asset \(i\) in period \(t\); \(z_t\) is \(R_{mt} - R_{ft}\), the market excess return in period \(t\); and disturbances \(\epsilon_{it}\) are the deviations from the conditional means. We define the \(T \times 2\) matrix \(X = ([1, z_t], t = 1, ..., T)\). The assumptions of the seemingly unrelated regressions model are
1. \( E[\epsilon_t|X] = E[\epsilon_t] = 0 \),
2. \( \text{Var}[\epsilon_t|X] = E[\epsilon_t \epsilon'_t|X] = \Sigma \), a positive definite \( N \times N \) matrix,
3. \( \epsilon_t|X \sim N(0, \Sigma) \)

The data are also assumed to be “well behaved” so that

4. \( \text{plim} \bar z = E[z_t] = \mu_z \),
5. \( \text{plim} s^2_z = \text{plim} \left( \frac{1}{T} \sum_{t=1}^{T} (z_t - \bar z)^2 \right) = \text{Var}[z_t] = \sigma^2_z \)

We can proceed to the maximum likelihood estimators of \( \Pi \) and \( \Sigma \). We know from our results that the GLS and maximum likelihood estimators are simply equation by equation ordinary least squares and that the estimator of \( \Sigma \) is just \( S \), the sample covariance matrix of the least squares residuals. The asymptotic covariance matrix for the \( 2N \times 1 \) estimator \([a, b]'\) will be

\[
\text{Asy.Var}[a, b]' = \frac{1}{T} \text{plim} \left[ \left( \frac{X'X}{T} \right)^{-1} \otimes \Sigma \right] = \frac{1}{T \sigma^2_z} \left[ \frac{\sigma^2_z + \mu^2_z}{\mu_z} \mu_z \right] \otimes \Sigma,
\]

which we will estimate with \( (X'X)^{-1} \otimes S \).

The model above does not impose the Markowitz-Sharpe-Lintner hypothesis, \( H_0: \alpha = 0 \). A Wald test of \( H_0 \) can be based on the unrestricted least squares estimates:

\[
W = (a - 0)' \{\text{Est.Asy.Var}[a - 0]\}^{-1} (a - 0) = a'[(X'X)^{11} S]^{-1} a = \left( \frac{T s^2_z}{s^2_z + \bar z^2} \right) a' S^{-1} a
\]

Under the null hypothesis, the statistic has a limiting chi-squared distribution with \( N \) degrees of freedom. An alternative that is likely to be better behaved is \( \frac{(T - N - 1)/N}{W} \), which is exactly distributed as \( F[N, T - N - 1] \) under the null hypothesis. The restricted estimator of \( \Sigma \) is, as before, the matrix of mean squares and cross products of the residuals, now \( S_0 \). The chi-squared statistic for the likelihood ratio test is

\[
\lambda = N(\ln |S_0| - \ln |S|).
\]

To compute the LM statistic, we will require the derivatives of the unrestricted log-likelihood function, evaluated at the restricted estimators. For this model, they may be written

\[
\frac{\partial \ln L}{\partial \alpha_i} = \sum_{j=1}^{n} \sigma^{ij} \left( \sum_{t=1}^{T} \epsilon_{jt} \right) = \sum_{j=1}^{n} \sigma^{ij} (T \bar \epsilon_j),
\]

where \( \sigma^{ij} \) is the \( ij \)th element of \( \Sigma^{-1} \), and

\[
\frac{\partial \ln L}{\partial \beta_i} = \sum_{j=1}^{n} \sigma^{ij} \left( \sum_{t=1}^{T} z_t \epsilon_{jt} \right) = \sum_{j=1}^{n} \sigma^{ij} (z'_j \bar \epsilon_j),
\]

The first vector of first derivatives can be written as

\[
\frac{\partial \ln L}{\partial \alpha} = \Sigma^{-1} E' i = \Sigma^{-1} (T \bar \epsilon)
\]

108
where \( i \) is a \( T \times 1 \) vector of 1s, \( E \) is a \( T \times N \) matrix of disturbances, and \( \bar{\epsilon} \) is the \( N \times 1 \) vector of means of asset specific disturbances. Combining terms and inserting the restricted estimates, we obtain

\[
LM = \left[ T \bar{e}_0' S_0^{-1} : 0 \right]' \left[ X' X \otimes S_0^{-1} \right]^{-1} \left[ T \bar{e}_0' S_0^{-1} : 0 \right]'
\]

\[
= T^2 \left( s_x^2 + \frac{z^2}{s_x^2} \right) \bar{e}_0' S_0^{-1} \bar{e}_0
\]

Under the null hypothesis, the limiting distribution of LM is chi-squared with \( N \) degrees of freedom.

The model formulation gives \( E[R_{it}] = R_{ft} + \beta_i(E[R_{mt} - R_{ft}] \right). If there is no risk-free asset but we write the model in terms of \( \gamma \), the unknown return on a zero-beta portfolio, then we obtain

\[
R_{it} = \gamma + \beta_i(R_{mt} - \gamma) + \epsilon_{it}
\]

\[
= (1 - \beta_i)\gamma + \beta_i R_{mt} + \epsilon_{it}
\]

The nonlinear restrictions will complicate finding the maximum likelihood estimators. Although maximization over \((\beta, \gamma)\) remains complicated, maximization over \(\beta\) for known \(\gamma\) is trivial. The full set of maximum likelihood estimators may be found just by scanning over the admissible range of \(\gamma\) to locate the value that maximizes

\[
\ln L_c = -\frac{1}{2} \ln |S(\gamma)|
\]

where

\[
s_{ij}(\gamma) = \frac{\sum_{t=1}^{T} \left\{ R_{it} - \gamma [1 - \tilde{\beta}_i(\gamma)] - \tilde{\beta}_i(\gamma) R_{mt} \right\} \left\{ R_{jt} - \gamma [1 - \tilde{\beta}_j(\gamma)] - \tilde{\beta}_j(\gamma) R_{mt} \right\}}{T}
\]

and

\[
\tilde{\beta}_i(\gamma) = \frac{\sum_{t=1}^{T} (R_{it} - \gamma)(R_{mt} - \gamma)}{\sum_{t=1}^{T} (R_{mt} - \gamma)^2}
\]

The log-likelihood for this model is

\[
\ln L = -\frac{T}{2} \left[ N \ln 2\pi + \ln |\Sigma| \right] - \frac{1}{2} \sum_{t=1}^{T} \epsilon_t' \Sigma^{-1} \epsilon_t
\]

The derivatives of the log-likelihood can be written

\[
\frac{\partial \ln L}{\partial [\beta' - \gamma]'} = \sum_{t=1}^{T} \left( (R_{mt} - \gamma) \Sigma^{-1} \epsilon_t \right)' (i - \beta)' \Sigma^{-1} \epsilon_t = \sum_{t=1}^{T} g_t
\]

With the derivatives in this form, we have

\[
E[g_t g_t'] = \begin{bmatrix}
(R_{mt} - \gamma)^2 \Sigma^{-1} & (R_{mt} - \gamma) \Sigma^{-1} (i - \beta) \\
(R_{mt} - \gamma) (i - \beta)' \Sigma^{-1} & (i - \beta)' \Sigma^{-1} (i - \beta)
\end{bmatrix}
\]

109
Sum this expression over \( t \) and use the result that
\[
\sum_{t=1}^{T} (R_{mt} - \gamma)^2 = \sum_{t=1}^{T} (R_{mt} - \bar{R}_m)^2 + T(\bar{R}_m - \gamma)^2 = T[s_{Rm}^2 + (\bar{R}_m - \gamma)^2]
\]
to obtain the negative of the expected Hessian,
\[
-E \left[ \frac{\partial^2 \ln L}{\partial (\beta_i) \partial (\beta_j)} \right] = T \left[ \begin{array}{c}
[s_{Rm}^2 + (\bar{R}_m - \gamma)^2]^{-1} (\bar{R}_m - \gamma) \Sigma^{-1}(i - \beta)
(\bar{R}_m - \gamma)(i - \beta) \Sigma^{-1} (i - \beta)
\end{array} \right]
\]
The inverse of this matrix provides the estimator for the asymptotic covariance matrix.

Once the estimates of \( \beta_i \) and \( \gamma \) are obtained, the implied estimates of \( \alpha_i \) are given by \( \alpha_i = (1 - \beta_i)\gamma \). With these estimates in hand, the LM statistic is exactly what it was before, although now all \( 2N \) derivatives will be required and \( X \) is \([i \ R_m]\). The subscript * indicates computation at the restricted estimates;
\[
LM = T \left( \frac{S_{Rm}^2 + \bar{R}_m^2}{s_{Rm}^2} \right) \varepsilon_*' S_*^{-1} \varepsilon_* + \left( \frac{1}{T s_{Rm}^2} \right) R_m' E_* S_*^{-1} E'_* R_m - \left( \frac{2\bar{R}_m}{s_z^2} \right) R_m' E_* S_*^{-1} \varepsilon_*
\]
A Wald test of the Black model would be based on the unrestricted estimators. Write this set of \( N - 1 \) functions as \( c(\alpha, \beta) = 0 \). The Wald statistic based on the least squares estimates would then be
\[
W = c(a, b)' \{ \text{Est. Asy. Var}[c(a, b)] \}^{-1} c(a, b)
\]
Recall in the unrestricted model that \( \text{Asy. Var}[a, b] = (1/T) \text{plim}(X'X/T)^{-1} \otimes \Sigma = \Delta \). Using the delta method the asymptotic covariance matrix for \( c(a, b) \) would be
\[
\text{Asy. Var}[c(a, b)] = \Gamma \Delta \Gamma'
\]
where
\[
\Gamma = \frac{\partial c(\alpha, \beta)}{\partial (\beta, \beta)}
\]
The \( i \)th row of the \( 2N \times 2N \) matrix \( \Gamma \) has four only nonzero elements, on each in \( i \)th and \( N \)th positions of each of the two subvectors.

### 14.2.6 Maximum Likelihood Estimation of the Seemingly Unrelated Regressions Model with A Block of Zeros in the Coefficient Matrix

In this section, we consider a related case in which the coefficient matrix contains a block of zeros. The block of zeros is created by excluding the same subset of the regressors from some of but not all the equations in a model that without the exclusion restriction is a SUR with the same regressors in all equations.

The model we have described can be formulated as follows
\[
y_t = \Pi x_t + \epsilon_t
\]
Now, consider once again a particular observation and partition the set of dependent variables into two groups of \( M_1 \) and \( M_2 \) variables and the set of regressors into two sets of \( K_1 \) and \( K_2 \) variables. The equation system is now
\[
\begin{pmatrix}
y_1 \\
y_2 
\end{pmatrix}_t =
\begin{pmatrix}
\Pi_{11} & \Pi_{12} \\
\Pi_{21} & \Pi_{22}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 
\end{pmatrix}_t +
\begin{pmatrix}
\epsilon_1 \\
\epsilon_2 
\end{pmatrix}_t,
E \left[ \begin{array}{c}
\epsilon_1
\epsilon_2
\end{array} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix},
\text{Var} \left[ \begin{array}{c}
\epsilon_1
\epsilon_2
\end{array} \right] = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}
\]
The case we are interested in here is the restricted model, with \( \Pi_{12} = 0 \), which has the effect of excluding \( x_2 \) from all the equations for \( y_1 \).
1. The maximum likelihood estimator of \( \Pi_{11} \) when \( \Pi_{12} = 0 \) is equation-by-equation least squares regression of the variables in \( y_1 \) on \( x_1 \) alone.

2. The effect of the restriction on the likelihood function can be isolated to its effect on the smaller set of equations. Thus, the hypothesis can be tested without estimating the larger set of equations.

The log-likelihood function for this multivariate regression model is

\[
\ln L = \sum_{t=1}^{T} \ln f(y_{1t}, y_{2t} | x_{1t}, x_{2t})
\]

where \( f(y_{1t}, y_{2t} | x_{1t}, x_{2t}) \) is the joint normal density of the two vectors. We will now write this joint normal density as the product of a marginal and a conditional:

\[
f(y_{1t}, y_{2t} | x_{1t}, x_{2t}) = f(y_{1t} | x_{1t}, x_{2t})f(y_{2t} | y_{1t}, x_{1t}, x_{2t})
\]

The mean and variance of the marginal distribution for \( y_{1t} \) are just the upper portions of the preceding partitioned matrices:

\[
E[y_{1t} | x_{1t}, x_{2t}] = \Pi_{11}x_{1t} + \Pi_{12}x_{2t}, \quad \text{Var}[y_{1t} | x_{1t}, x_{2t}] = \Sigma_{11}
\]

Collecting terms, we have

\[
E[y_{2t} | y_{1t}, x_{1t}, x_{2t}] = [\Pi_{21} - \Sigma_{21}\Sigma_{11}^{-1}\Pi_{11}]x_{1t} + [\Pi_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Pi_{12}]x_{2t} + [\Sigma_{21}\Sigma_{11}^{-1}]y_{1t}
\]

\[
= \Lambda_{21}x_{1t} + \Lambda_{22}x_{2t} + \Gamma y_{1t}
\]

\[
= \text{Var}[y_{2t} | y_{1t}, x_{1t}, x_{2t}] = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Omega_{22}
\]

The objective of this partitioning is to partition the log-likelihood function likewise;

\[
\ln L = \sum_{t=1}^{T} \ln f(y_{1t}, y_{2t} | x_{1t}, x_{2t})
\]

\[
= \sum_{t=1}^{T} \ln f(y_{1t} | x_{1t}, x_{2t})f(y_{2t} | y_{1t}, x_{1t}, x_{2t})
\]

\[
= \sum_{t=1}^{T} \ln f(y_{1t} | x_{1t}, x_{2t}) + \sum_{t=1}^{T} \ln f(y_{2t} | y_{1t}, x_{1t}, x_{2t})
\]

With no restrictions on any of the parameters, we can maximize this log-likelihood by maximizing its parts separately.

Now, consider estimation of the same system subject to the restriction \( \Pi_{12} = 0 \). The second equation system is still completely unrestricted, so maximum likelihood estimates of its parameters, \( \Lambda_{21}, \Lambda_{21} \) (which now equals \( \Pi_{22} \)), \( \Gamma \), and \( \Omega_{22} \), are still obtained by equation-by-equation least squares. The equation systems have no parameters in common, so maximum likelihood estimators of the first set of parameters are obtained by maximizing the first part of the log-likelihood estimators of the first set of parameters are obtained by maximizing the first part of the log-likelihood, once again, by equation-by-equation ordinary least squares. Thus, our first result is established. To establish the second result, we must
obtain the two parts of the log-likelihood. The log-likelihood function for this model is given in
\[ \log L = -\frac{T}{2} \left[ M \log(2\pi) + \log |\Sigma| + \text{tr}(\Sigma^{-1}W) \right] \]
Since each of the two sets of equations is estimated by least squares, in each case, for each part, the term in the log-likelihood is the concentrated log-likelihood give in
\[ \log L_c = -\frac{T}{2} [M(1 + \log(2\pi)) + \log |W|] \]
where \( W_{jj} \) is \((1/T)\) times the matrix of sums of squares and cross products of least squares residuals. The second set of equations is estimated by regressions on \( x_1, x_2 \) and \( y_1 \) with or without the restriction \( \Pi_{12} = 0 \). So, the second part of the log-likelihood is always the same
\[ \log L_{2c} = -\frac{T}{2} [M_2(1 + \log(2\pi)) + \log |W_{22}|] \]
The concentrated log-likelihood for the first set of equations equals
\[ \log L_{1c} = -\frac{T}{2} [M_1(1 + \log(2\pi)) + \log |W_{11}|] \]
At the maximum likelihood estimators, the log-likelihood for the whole system is
\[ \ln L_c = \ln L_{1c} + \ln L_{2c} \]
The likelihood ratio statistic is
\[ \lambda = -2[(\ln L_c|\Pi_{12} = 0) - (\ln L_c)] = T[\ln |W_{11}(\Pi_{12} = 0)| - \ln |W_{11}|] \]
This establishes our second result, since \( W_{11} \) is based only on the first set of equations.

14.2.7 Autocorrelation and heteroscedasticity

The seemingly unrelated regressions model can be extended to allow for autocorrelation. To reiterate, suppose that
\[ y_i = X_i\beta_i + \epsilon_i \]
\[ \epsilon_{it} = \rho \epsilon_{i,t-1} + u_{it} \]
where \( u_{it} \) is uncorrelated across observations. The treatment developed by Parks (1967) is the one we used earlier. It calls for a three step approach:

1. Estimate each equation in the system by ordinary least squares. Compute any consistent estimators of \( \rho \). For each equation, transform the data by the Prais-Winsten Transformation to remove the autocorrelation.
2. Using the transformed data, use ordinary least squares again to estimate \( \Sigma \)
3. Use FGLS based on the estimated \( \Sigma \) and the transformed data
There is no benefit to iteration. After the last step, Σ should be reestimated with the GLS estimates. The estimated covariance matrix for ε can then be reconstructed using

\[ \hat{\sigma}_{mn}(\epsilon) = \frac{\hat{\sigma}_{mn}}{1 - r_m r_n} \]

In principle, the SUR model can accommodate heteroscedasticity as well as autocorrelation. Bartels and Feibig (1991) suggested the generalized SUR model, \( \Omega = A |\Sigma \otimes I| A' \) where \( A \) is a block diagonal matrix. Ideally, \( A \) is made a function of measured characteristics of the individual and a separate parameter vector, \( \theta \), so that the model can be estimated in stages. In a first step, OLS residuals could be used to form a preliminary estimator of \( \theta \), then the data are transformed to homoscedasticity, leaving \( \Sigma \) and \( \beta \) to be estimated at subsequent steps using transformed data.

14.3 Systems of Demand Equations Singular Systems

Most of the recent applications of the multivariate regression model have been in the context of systems of demand equations, either commodity demands or factor demands in studies of production.

Some special problems arise in these settings. First, the parameters of the systems are generally constrained across equations. A second intrinsic feature of many of these models is that the disturbance covariance matrix \( \Sigma \) is singular.

14.3.1 Cobb-Douglas Cost Function

Consider a Cobb-Douglas production function,

\[ Y = \alpha_0 \prod_{i=1}^{M} x_i^{\alpha_i} \]

Profit maximization with an exogenously determined output price calls for the firm to maximize output for a given cost level \( C \). The Lagrangean for the maximization problem is

\[ \Lambda = \alpha_0 \prod_{i=1}^{M} x_i^{\alpha_i} + \lambda (C - p'x) \]

where \( p \) is the vector of \( M \) factor prices. The necessary conditions for maximizing this function are

\[ \frac{\partial \Lambda}{\partial x_i} = \frac{\alpha_i Y}{x_i} - \lambda p_i = 0 \quad \text{and} \quad \frac{\partial \Lambda}{\partial \lambda} = C - p'x = 0 \]

The joint solution provides \( x_i(Y,p) \) and \( \lambda(Y,p) \). The total cost of production is

\[ \sum_{i=1}^{M} p_i x_i = \sum_{i=1}^{M} \frac{\alpha_i Y}{\lambda} \]

The cost share allocated to the \( i \)th factor is

\[ \frac{p_i x_i}{\sum_{i=1}^{M} p_i x_i} = \frac{\alpha_i}{\sum_{i=1}^{M} \alpha_i} = \beta_i \]

113
The full model is

\[ \ln C = \beta_0 + \beta_y \ln Y + \sum_{i=1}^{M} \beta_i \ln p_i + \epsilon_i, \]

\[ s_i = \beta_i + \epsilon_i, \quad i = 1, \ldots, M \]

By construction, \( \sum_{i=1}^{M} \beta_i = 1 \) and \( \sum_{i=1}^{M} s_i = 1 \). It therefore follows that \( \sum_{i=1}^{M} \epsilon_i = 0 \) at every data point, so the system is singular. Let the \( M \times 1 \) disturbance vector from the shares be \( \epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_M]' \). Since \( \epsilon'Y = 0 \), where \( Y \) is a column of 1s, it follows that \( E[\epsilon'Y] = \Sigma i = 0 \), which implies that \( \Sigma \) is singular. Therefore the methods of the previous sections cannot be used here.

The solution to the singularity problem appears to be to drop one of the equations, estimate the remainder, and solve for the last parameter from the other \( M - 1 \). If we impose the constraint

\[ \beta_M = 1 - \beta_1 - \beta_2 - \cdots - \beta_{M-1} \]

then the system is reduced to a nonsingular one:

\[ \log \left( \frac{C}{p_M} \right) = \beta_0 + \beta_y \log Y + \sum_{i=1}^{M-1} \beta_i \log \left( \frac{p_i}{p_M} \right) + \epsilon, \]

\[ s_i = \beta_i + \epsilon_i, \quad i = 1, \ldots, M - 1 \]

This system provides estimates of \( \beta_0, \beta_y \) and \( \beta_1, \ldots, \beta_{M-1} \).

14.3.2 Flexible Functional Forms: The Translog Cost Function

Suppose that production is characterized by a production function, \( Y = f(x) \). The solution to the problem of minimizing the cost of producing a specified output rate given a set of factor prices produces the cost-minimizing set of factor demands \( x_i = x_i(Y, p) \). The total cost of production is given by the cost function,

\[ C = \sum_{i=1}^{M} p_i x_i(Y, p) = C(Y, p) \]

If there are constant returns to scale, then it can be shown that \( C = Yc(p) \) or

\[ C/Y = c(p) \]

where \( c(p) \) is the unit or average cost function. The cost-minimizing factor demands are obtained by applying Shephard’s (1970) lemma, which states that if \( C(Y, p) \) gives the minimum total cost of production, then the cost-minimizing set of factor demands is given by

\[ x_i^* = \frac{\partial C(Y, p)}{\partial p_i} = \frac{Y \partial c(p)}{\partial p_i} \]

By differentiating logarithmically, we obtain the cost-minimizing factor cost shares:

\[ s_i = \frac{\partial \log C(Y, p)}{\partial \log p_i} = \frac{p_i x_i}{C} \]
With constant returns to scale, \( \ln C(Y, p) = \log Y + \log c(p) \), so

\[
s_i = \frac{\partial \log c(p)}{\partial \log p_i}
\]

In many empirical studies, the objects of estimation are the elasticities of factor substitution and the own price elasticities of demand, which are given by

\[
\theta_{ij} = \frac{c(\partial^2 c/\partial p_i \partial p_j)}{(\partial c/\partial p_i)(\partial c/\partial p_j)}
\]

and

\[
\eta_{ii} = s_i \theta_{ii}
\]

The transcendental logarithmic, or translog, function is the most frequently used flexible function in empirical work. By expanding \( \log c(p) \) in a second-order Taylor’s series about the point \( \log p = 0 \), we obtain

\[
\log c = \beta_0 + \sum_{i=1}^{M} \left( \frac{\partial \log c}{\partial \log p_i} \right) \log p_i + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \left( \frac{\partial^2 \log c}{\partial \log p_i \partial \log p_j} \right) \log p_i \log p_j
\]

If we identify these derivatives as coefficients and impose the symmetry of the cross-price derivatives, then the cost function becomes

\[
\log c = \beta_0 + \beta_1 \log p_1 + \cdots + \beta_M \log p_M + \delta_{11} \left( \frac{1}{2} \log^2 p_1 \right) + \delta_{12} \log p_1 \log p_2 + \cdots + \delta_{MM} \left( \frac{1}{2} \log^2 p_M \right)
\]

This is the translog cost function. If \( \delta_{ij} \) equals zero, then it reduces to the Cobb-Douglas function we looked at earlier. The cost shares are given by

\[
s_1 = \frac{\partial \log c}{\partial \log p_1} = \beta_1 + \delta_{11} \log p_1 + \delta_{12} \log p_2 + \cdots + \delta_{1M} \log p_M
\]

\[
s_2 = \frac{\partial \log c}{\partial \log p_2} = \beta_2 + \delta_{12} \log p_1 + \delta_{22} \log p_2 + \cdots + \delta_{2M} \log p_M
\]

\[
\vdots
\]

\[
s_M = \frac{\partial \log c}{\partial \log p_M} = \beta_M + \delta_{1M} \log p_1 + \delta_{2M} \log p_2 + \cdots + \delta_{MM} \log p_M
\]

The systems of share equations provides a seemingly unrelated regressions model that can be used to estimate the parameters of the model.

14.4 Nonlinear Systems and GMM Estimation

We now consider estimation of nonlinear systems of equations. We briefly consider two cases in this section, maximum likelihood (or FGLS) estimation and GMM estimation.
Consider estimation of the parameters of the equation system

\[ \begin{align*}
    y_1 &= h_1(\beta, X) + \epsilon_1 \\
    y_2 &= h_2(\beta, X) + \epsilon_2 \\
        &\vdots \\
    y_M &= h_M(\beta, X) + \epsilon_M 
\end{align*} \]

There are \( M \) equations in total, to be estimated with \( t = 1, \ldots, T \) observations. There are \( K \) parameters in the model. The disturbances are assume to have zero means and contemporaneous covariance matrix \( \Sigma \).

### 14.4.1 GLS Estimation

In the multivariate regression model, if \( \Sigma \) is known, then the generalized least squares estimator of \( \beta \) that minimizes the generalized sum of squares

\[ \epsilon(\beta)' \Omega^{-1} \epsilon(\beta) = \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma^{ij} [y_i - h_i(\beta, X)]' [y_i - h_i(\beta, X)] \]

where \( \epsilon(\beta) \) is an \( MT \times 1 \) vector of disturbances obtained by stacking the equations and \( \Omega = \Sigma \otimes I \). The first-order condition for minimizing this sum of squares is

\[ \frac{\partial \epsilon(\beta)' \Omega^{-1} \epsilon(\beta)}{\partial \beta} = \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma^{ij} [2 X_i^0(\beta)' \epsilon_j(\beta)] = 0 \]

where \( \sigma^{ij} \) is the \( ij \)th element of \( \Sigma^{-1} \) and \( X_i^0(\beta) \) is a \( T \times K \) matrix of pseudoregressors from the linearization of the \( i \)th equation. If any of the parameters in \( \beta \) do not appear in the \( i \)th equation, then the corresponding column of \( X_i^0(\beta) \) will be a column of zeros.

Once the FGLS estimator is obtained, the appropriate asymptotic covariance matrix is estimated with

\[ \text{Est. Asy. Var}[\hat{\beta}] = \left( \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma^{ij} X_i^0(\beta)' X_j^0(\beta) \right)^{-1} \]

It is easy to show that the preceding defines a GMM estimator. We can use this result to devise an alternative, simple strategy. For our first step, we can find \( \beta \) to minimize

\[ \epsilon(\beta)' \epsilon(\beta) = \sum_{i=1}^{M} [y_i - h_i(\beta, X)]' [y_i - h_i(\beta, X)] = \sum_{i=1}^{M} \sum_{t=1}^{T} [y_{it} - h_i(\beta, x_{it})]^2 \]

### 14.4.2 Maximum Likelihood Estimation

With normally distributed disturbances, the log-likelihood function for this model is still given by

\[ \log L = \sum_{t=1}^{T} \log L_t = -\frac{MT}{2} \log(2\pi) - \frac{T}{2} \log |\Sigma| - \frac{1}{2} \sum_{t=1}^{T} \epsilon_t' \Sigma^{-1} \epsilon_t \]

Therefore, estimation of \( \Sigma \) is done exactly as before. Likewise, the concentrated log-likelihood and the criterion function are unchanged. Therefore, one approach to maximum
likelihood estimation is iterated FGLS, based on the results in Section 14.2.3. This method will require two levels of iteration, however, since for each estimated \( \Sigma(\beta_l) \), written as a function of the estimates of \( \beta \) obtained at iteration \( l \), a nonlinear, iterative solution is required to obtain \( \beta_{l+1} \). The iteration then returns to \( S \). Convergence is based either on \( S \) or \( \hat{\beta} \); if one stabilizes, then the other will also.

Suppose then that there is one underlying parameter vector \( \beta \) and that we formulated each equation as

\[
h_{it} = h_i(\gamma_i, x_{it}) + \epsilon_{it}
\]

Then the derivatives of the log-likelihood function are built up from

\[
\frac{\partial \ln |S(\gamma)|}{\partial \gamma_i} = d_i = -\frac{1}{T} \sum_{t=1}^{T} \left( \sum_{j=1}^{M} s^{ij} x^0_{it}(\gamma_i) e_{jt}(\gamma_j) \right), \quad i = 1, ..., M
\]

Since each \( \gamma_i \) is built up just by extracting elements from \( \beta \), the relevant derivative with respect to \( \beta \) is just a sum of those with respect to \( \gamma \).

\[
\frac{\partial \ln L_c}{\partial \beta_k} = \sum_{i=1}^{n} \left[ \frac{\partial \ln L_c}{\partial \gamma_{ig}} 1_{(\gamma_{ig} = \beta_k)} \right]
\]

where \( 1_{(\gamma_{ig} = \beta_k)} \) equals 1 if \( \gamma_{ig} \) equals \( \beta_k \) and 0 of not. Define the matrix \( F \) with \( G \) rows and \( K \) columns. Then let \( F_{ij} = 1 \) if \( \gamma_g = \beta_j \) and 0 otherwise. Let \( d \) be the \( G \times 1 \) vector of derivatives obtained by stacking \( d_i \) from equation above, then

\[
\frac{\partial \ln L_c}{\partial \beta} = F'd
\]

The Hessian is likewise computed as a simple sum of terms. We can construct it in blocks using

\[
H_{ij} = \frac{\partial^2 \ln L_c}{\partial \gamma_i \partial \gamma_j'} = -\sum_{t=1}^{T} s^{ij} x^0_{it}(\gamma_i) x^0_{jt}(\gamma_j)'
\]

The asymptotic covariance matrix for \( \hat{\beta} \) is once again a sum of terms:

\[
\text{Est. Asy. Var}[\hat{\beta}] = V = [-F'HF]^{-1}
\]

14.4.3 GMM Estimation

All the preceding estimation techniques can be obtained as GMM estimators. Suppose as well that there are a set of instrumental variables \( z_t \) such that

\[
E[z_t \epsilon_{it}] = 0, \quad t = 1, ..., T \quad \text{and} \quad i = 1, ..., M
\]

The sample analog is

\[
\frac{1}{T} \sum_{t=1}^{T} z_t [y_{it} - h_i(\beta, x_{it})] = 0, \quad i = 1, ..., M
\]

Let

\[
\frac{1}{T} Z' \Omega_{ij} Z = E \left[ \frac{Z' \epsilon_i \epsilon_j' Z}{T} \right]
\]
The GMM criterion for estimation in this setting is

$$q = \sum_{i=1}^{M} \sum_{j=1}^{M} \left[(y_i - h_i(\beta, X))^\prime Z/T \right] \left[ Z' \Omega_{ij} Z/T \right]^{ij} \left[ Z'(y_i - h_i(\beta, X))/T \right]$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{M} \epsilon_i(\beta)^\prime Z/T \left[ Z' \Omega_{ij} Z/T \right]^{ij} \left[ Z' \epsilon_j(\beta)/T \right]$$

where $[Z' \Omega_{ij} Z/T]^{ij}$ denotes the $ij$th block of the inverse of the matrix with the $ij$th block equal to $Z' \Omega_{ij} Z/T$. The first-order conditions are

$$\frac{\partial q}{\partial \beta} = \sum_{i=1}^{M} \sum_{j=1}^{M} [X_i^0(\beta)^\prime Z/T] \left[ Z' W_{ij} Z/T \right]^{ij} \left[ Z' \epsilon_j(\beta)/T \right] = 0$$

At completion, the asymptotic covariance matrix for the GMM estimator is estimated with

$$V_{GMM} = \frac{1}{T} \left[ \sum_{i=1}^{M} \sum_{j=1}^{M} [X_i^0(\beta)^\prime Z/T] \left[ Z' W_{ij} Z/T \right]^{ij} \left[ Z' X_j^0(\beta)/T \right] \right]^{-1}$$
Chapter 15

Simultaneous-Equations Models

15.1 Fundamental Issues in Simultaneous-Equations Models

In this section, we describe the basic terminology and statistical issues in the analysis of simultaneous-equations models.

15.1.1 Illustrative Systems of Equations

A familiar example of a system of simultaneous equations is a model of market equilibrium, consisting of the following:

\[
\begin{align*}
\text{demand equation:} & \quad q_{d,t} &= \alpha_1 p_t + \alpha_2 x_t + \epsilon_{d,t} \\
\text{supply equation:} & \quad q_{s,t} &= \beta_1 p_t + \epsilon_{s,t} \\
\text{equilibrium condition:} & \quad q_{d,t} = q_{s,t} = \pi_t
\end{align*}
\]

These equations are structural equations in that they are derived from theory and each purports to describe a particular aspect of the economy. Since the model is one of the joint determination of price and quantity, they are labeled joint dependent or endogenous variables. Income \( x \) is assumed to be determined outside of the model, which makes it exogenous. The disturbances are added to the usual textbook description to obtain an econometric model. All three equations are needed to determine the equilibrium price and quantity in terms of income and the disturbances is, indeed, implied, the system is said to be a complete system of equations. The completeness of the system requires that the number of equations equal the number of endogenous variables.

For simplicity, assume that \( \epsilon_d \) and \( \epsilon_s \) are well behaved, classical disturbances with

\[
\begin{align*}
E[\epsilon_{d,t}|x_t] &= E[\epsilon_{s,t}|x_t] = 0, \\
E[\epsilon_{d,t}^2|x_t] &= \sigma_d^2, E[\epsilon_{s,t}^2|x_t] = \sigma_s^2, \\
E[\epsilon_{d,t}\epsilon_{s,t}|x_t] &= E[\epsilon_{d,t}x_t] = E[\epsilon_{s,t}x_t] = 0
\end{align*}
\]

Solving the equations for \( p \) and \( q \) in terms of \( x \), and \( \epsilon_d \), and \( \epsilon_s \) produces the reduced form of the model

\[
\begin{align*}
p &= \frac{\alpha_2 x}{\beta_1 - \alpha_1} + \frac{\epsilon_d - \epsilon_s}{\beta_1 - \alpha_1} = \pi_1 x + v_1 \\
q &= \frac{\beta_1 \alpha_2 x}{\beta_1 - \alpha_1} + \frac{\beta_1 \epsilon_d - \alpha_1 \epsilon_s}{\beta_1 - \alpha_1} = \pi_2 x + v_2
\end{align*}
\]
It follows that \( \text{Cov}[p, \epsilon_d] = \sigma_d^2/ (\beta_1 - \alpha_1) \) and \( \text{Cov}[p, \epsilon_s] = -\sigma_s^2/ (\beta_1 - \alpha_1) \) so neither the demand nor the supply equation satisfies the assumptions of the classical regression model. The price elasticity of demand cannot be consistently estimated by least squares regression of \( q \) on \( y \) and \( p \).

Suppose that we have \( T \) observations on \( p, q \) and \( y \) such that
\[
\text{plim} (1/T)x'x = \sigma_x^2
\]
Since least squares is inconsistent, we might instead use an instrumental variable estimator. The only variable in the system that is not correlated with the disturbances is \( x \). Consider, then, the IV estimator, \( \hat{\beta}_1 = q'x/p'x \). This estimator has
\[
\text{plim} \hat{\beta}_1 = \text{plim} \frac{q'x/T}{p'x/T} = \beta_1 \frac{\alpha_2/(\beta_1 - \alpha_1)}{\alpha_2/(\beta_1 - \alpha_1)} = \beta_1
\]
In the least squares regression of \( p \) on \( x \), the predicted values are \( \hat{p} = (p'x/x'x)x \). It follows that in the instrumental variable regression the instrument is \( \hat{p} \). That is,
\[
\hat{\beta}_1 = \frac{\hat{p}'q}{\hat{p}'p}
\]
This interpretation defines the two-stage least squares estimator.

15.1.2 Endogeneity and Causality

Engle, Hendry, and Richard (1983) define a set of variables \( x_t \) in a parameterized model to be weakly exogenous if the full model can be written in terms of a marginal probability distribution for \( x_t \) and a conditional distribution for \( y_t | x_t \) such that estimation of the parameters for the conditional distribution is no less efficient than estimation of the full set of parameters of the joint distribution.

With reference to time-series applications, variables \( x_t \) are said to be predetermined in the model if \( x_t \) is independent of all subsequent structural disturbances \( \epsilon_{t+s} \) for \( s > 0 \).

Granger causality is absent when \( f(x_t | x_{t-1}, y_{t-1}) \) equals \( f(x_t | x_{t-1}) \). The definition states that in the conditional distribution, lagged values of \( y_t \) add no information to explanation of movements of \( x_t \) beyond that provided by lagged values of \( x_t \) itself. Finally, if \( x_t \) is weakly exogenous and if \( y_{t-1} \) does not Granger cause \( x_t \), then \( x_t \) is strongly exogenous.

15.1.3 A General Notation for Linear Simultaneous Equations Models

The structural form of the model is
\[
\begin{align*}
\gamma_{11}y_1 + \gamma_{21}y_2 + \cdots + \gamma_{M1}y_M + \beta_{11}x_{t1} + \cdots + \beta_{K1}x_{tK} &= \epsilon_{t1}, \\
\gamma_{12}y_1 + \gamma_{22}y_2 + \cdots + \gamma_{M2}y_M + \beta_{12}x_{t1} + \cdots + \beta_{K2}x_{tK} &= \epsilon_{t2}, \\
&\vdots \\
\gamma_{1M}y_1 + \gamma_{2M}y_2 + \cdots + \gamma_{MM}y_M + \beta_{1M}x_{t1} + \cdots + \beta_{KM}x_{tK} &= \epsilon_{tM}.
\end{align*}
\]
There are \( M \) equations and \( M \) endogenous variables, denoted \( y_1, \ldots, y_M \). There are \( K \) exogenous variables, \( x_1, \ldots, x_K \), that may include predetermined values of \( y_1, \ldots, y_M \) as well. The first element of \( x_t \) will usually be the constant, 1. Finally, \( \epsilon_{t1}, \ldots, \epsilon_{tM} \) are the structural
disturbances. The subscript $t$ will be used to index observations, $t = 1, \ldots, T$. In matrix terms, the system may be written

$$y_t' \Gamma + x_t' B = \epsilon_t'$$

Each column of the parameter matrices is the vector coefficients in a particular equation, whereas each row applies to a specific variable.

If $\Gamma$ is an upper triangular matrix, then the system is said to be triangular. In this case, the model is of the form,

$$y_{t1} = f_1(x_t) + \epsilon_{t1},$$
$$y_{t2} = f_2(y_{t1}, x_t) + \epsilon_{t2},$$
$$\vdots$$
$$y_{tM} = f_M(y_{t1}, y_{t2}, \ldots, y_{tM-1}, x_t) + \epsilon_{tM}$$

The joint determination of the variables in this model is recursive. The first completely determined by the exogenous factors. Then, given the first, the second is likewise determined, and so on.

The solution of the system equations determining $y_t$ in terms of $x_t$ and $\epsilon_t$ is the reduced form of the model

$$y_t = -x_t' B \Gamma^{-1} + \epsilon_t' \Gamma^{-1} = x_t' \Pi + v_t'$$

For this solution to exist, the model must satisfy the completeness condition for simultaneous equations systems: $\Gamma$ must be nonsingular.

The structural disturbances are assumed to be randomly drawn from an $M$-variate distribution with

$$E[\epsilon_t|x_t] = 0 \quad \text{and} \quad E[\epsilon_t' \epsilon_s|x_t] = \Sigma$$

For the present, we assume that

$$E[\epsilon_t \epsilon_s'|x_t, x_s] = 0, \quad \forall t, s$$

It follows that the reduced form disturbances, $v_t' = \epsilon_t' \Gamma^{-1}$ have

$$E[v_t|x_t] = (\Gamma^{-1})'0 = 0,$$
$$E[v_t v_s'|x_t] = (\Gamma^{-1})' \Sigma \Gamma^{-1} = \Omega$$

This implies that

$$\Sigma = \Gamma' \Omega \Gamma$$

In terms of the full set of $T$ observations, the structure is

$$Y \Gamma + X B = E,$$

with

$$E[E|X] = 0 \quad \text{and} \quad E[(1/T) E'E|X] = \Sigma$$
Under general conditions, we can strengthen this structure to
\[ \text{plim}[(1/T)E'E] = \Sigma \]

An important assumption
\[ \text{plim}(1/T)X'X = Q, \quad \text{a finite positive definite matrix} \]
The reduced form is
\[ Y = X\Pi + V, \quad \text{where } V = E\Gamma^{-1} \]
Combining the earlier results, we have
\[
\text{plim} \frac{1}{T} \begin{bmatrix}
Y' \\
X' \\
V'
\end{bmatrix} = \begin{bmatrix}
\Pi'Q & \Pi'Q & \Omega \\
Q & Q & 0' \\
\Omega & 0 & \Omega
\end{bmatrix}
\]

### 15.2 The Problem of Identification

We observe
\[
\begin{align*}
\text{plim}(1/T)X'X &= Q \\
\text{plim}(1/T)X'Y &= \text{plim}(1/T)X'(X\Pi + V) = Q\Pi \\
\text{plim}(1/T)Y'Y &= \text{plim}(1/T)(\Pi'X' + V')(X\Pi + V) = \Pi'Q\Pi + \Omega
\end{align*}
\]
Therefore, \( \Pi \), the matrix of reduced-form coefficients, is observable:
\[
\Pi = \begin{bmatrix}
\text{plim} \left( \frac{X'X}{T} \right)^{-1} \\
\text{plim} \left( \frac{X'Y}{T} \right)
\end{bmatrix}
\]
Since \( \Pi \) is observable, \( \Omega \) is also:
\[
\Omega = \text{plim} \frac{Y'Y}{T} - \text{plim} \begin{bmatrix}
X'X \\
X'Y
\end{bmatrix}^{-1} \begin{bmatrix}
X'X \\
X'Y
\end{bmatrix}
\]
Therefore, \( \Pi \) and \( \Omega \) can be estimated consistently by least squares regression of \( Y \) on \( X \).

The correspondence between the structural and reduced-form parameters is the relationships
\[
\Pi = -B\Gamma^{-1} \quad \text{and} \quad \Omega = E[vv'] = (\Gamma^{-1})'\Sigma\Gamma^{-1}
\]
If \( \Gamma \) were known, then we could deduce \( B \) as \(-\Pi\Gamma \) and \( \Sigma \) as \( \Gamma'\Omega\Gamma \).

The identification question we will pursue can be posed as follows: We can “observe” the reduced form. We must deduce the structure from what we know about the reduced form. If there is more than one structure that can lead to the same reduced form, then we can estimate the structure.

#### 15.2.1 The Rank and Order Conditions for Identification

The additional information comes in several forms.

1. Normalization. In each equation, one variable has a coefficient of 1.
2. Identities. In some models, variable definitions or equilibrium conditions imply that all the coefficients in a particular equation are known.

3. Exclusions. The omission of variables from an equation places zeros in $B$ and $\Gamma$.

4. Linear Restrictions. Restrictions on the structural parameters may also serve to rule out false structures.

5. Restrictions on the disturbance covariance matrix. In the identification of a model, these are similar to restrictions on the slope parameters.

To formalize the identification criteria, we require a notation for a single equation. The coefficients of the $j$th equation are contained in the $j$th columns of $\Gamma$ and $B$. The $j$th equation is

$$y'\Gamma_j + x'B_j = \epsilon_j$$

In this equation, we known that (1) one of the elements in $\Gamma_j$ is one and (2) some variables that appear elsewhere in the model are excluded from this equation.

Equation $j$ may be written

$$y_j = Y'_j\gamma_j + Y'_j\gamma^*_j + x'_j\beta_j + x'_j\beta^*_j + \epsilon_j$$

The exclusions imply that $\gamma^*_j = 0$ and $\beta^*_j = 0$. Thus,

$$\Gamma'_j = \begin{bmatrix} 1 & -\gamma'_j & 0' \end{bmatrix} \quad \text{and} \quad B'_j = \begin{bmatrix} -\beta'_j & 0' \end{bmatrix}$$

The reduced-form coefficient matrix is

$$\Pi = -B\Gamma^{-1}$$

which implies that

$$\Pi\Gamma = -B$$

The $j$th column of this matrix equation applies to the $j$th equation,

$$\Pi\Gamma_j = -B_j$$

yields

$$\begin{bmatrix} \pi_j & \Pi_j & \Pi_j^* \\ \pi_j^* & \Pi_j^* & \Pi_j^* \end{bmatrix} \begin{bmatrix} 1 \\ -\gamma_j \\ 0 \end{bmatrix} = \begin{bmatrix} \beta_j \\ 0 \end{bmatrix}$$

Now extract the two subequations,

$$\pi_j - \Pi_j\Gamma_j = \beta_j \quad (K_j \text{ equations})$$

$$\pi_j^* - \Pi_j^*\Gamma_j = 0 \quad (K_j^* \text{ equations})$$

The second equation may be written

$$\Pi_j^*\gamma_j = \pi_j^*$$

This system is $K_k^*$ equations in $M_j$ unknowns. If they can be solved for $\gamma_j$, then the first equation gives the solution for $\beta_j$ and the equation is identified.
**Definition**  
**Order Condition for Identification of Equation** \( j \)  
\[ K_j^* \geq M_j \]

The order condition is only a counting rule. It ensures that the equations has at least one solution, but it does not ensure that it has only one solution.

**Definition**  
**Rank Condition for Identification**  
\[ \text{rank}[\pi_j^*, \Pi_j^*] = \text{rank}[\Pi_j^*] = M_j \]

This condition imposes a restriction on a submatrix of the reduced-form coefficient matrix.

The rank condition ensures that there is exactly one solution for the structural parameters given the reduced-form parameters. We first rearrange the structural coefficients in the matrix.

\[
A = \begin{bmatrix} \Gamma \\ B \end{bmatrix} = \begin{bmatrix} 1 & A_1 \\ -\gamma_j & A_2 \\ 0 & A_3 \\ -\beta_j & A_4 \\ 0 & A_5 \end{bmatrix} = [a_j \ A_j]
\]

The \( j \)th column in a false structure \([\Gamma F, BF]\) would be \([\Gamma f_j, Bf_j]\), where \( f_j \) is the \( j \)th column of \( F \). Thus, partitioning as previously,

\[
\tilde{a}_j = \begin{bmatrix} 1 & A_1 \\ -\gamma_j & A_2 \\ 0 & A_3 \\ -\beta_j & A_4 \\ 0 & A_5 \end{bmatrix} \begin{bmatrix} f^0 \\ f^1 \end{bmatrix} = \begin{bmatrix} 1 \\ \tilde{\gamma}_j \\ 0 \\ \tilde{\beta}_j \\ 0 \end{bmatrix}
\]

If this hybrid is to have the same variables as the original, then it must have nonzero elements in the same places, which can be ensured by taking \( f^0 = 1 \), and zeros in the same positions as the original \( a_j \). Extracting the third and fifth blocks of rows, if \( \tilde{a}_j \) is to be admissible, then it must meet the requirement

\[
\begin{bmatrix} A_3 \\ A_5 \end{bmatrix} f^1 = 0
\]

so we have the equivalent rank condition,

\[
\text{rank} \begin{bmatrix} A_3 \\ A_5 \end{bmatrix} = M - 1
\]

The corresponding order condition is that the matrix in brackets must have at least as many rows as columns. Thus, \( M_j^* + K_j^* \geq M - 1 \). For a large model, frequently only the order condition is verified. We distinguish three cases:

1. Underidentified. \( K_j^* < M_j \) or rank condition fails.
2. Exactly identified. \( K_j^* = M_j \) and rank condition is met.
3. Overidentified. \( K_j^* > M_j \) and rank condition is met.
15.2.2 Identification Through Other Nonsample Information

The order condition that emerges is

\[ n_j \geq M - 1 \]

where \( n_j \) is the total number of restrictions. Since \( M - 1 = M_j + \hat{M}_j \) and \( n_j \) is the number of exclusions plus \( r_j \), the number of additional restrictions, this condition is equivalent to

\[ r_j + K_j^* + \hat{M}_j \geq M_j + \hat{M}_j \]

or

\[ r_j + K_j^* \geq M_j \]

15.2.3 Identification Through Covariance Restrictions–The Fully Recursive Model

The fully recursive model is an important special case of the preceding result. A triangular system is

\[
\begin{align*}
y_1 &= \beta_1' x + \epsilon_1 \\
y_2 &= \gamma_{12} y_1 + \beta_2' x + \epsilon_2 \\
&\vdots \\
y_M &= \gamma_{1M} y_1 + \gamma_{2M} y_2 + \cdots + \gamma_{M-1,M} y_{M-1} + \beta_M' x + \epsilon_M
\end{align*}
\]

The first equation is identified, since it is already in reduced form. We conclude that without some identifying restrictions, only the parameters of the first equation in a triangular system are identified. But suppose that \( \Sigma \) is diagonal. Then the entire model is identified, as we now prove.

The \( j \)th column of \( F, f_j \), is the coefficients in a linear combination of the equations that will be an imposter for equation \( j \). Many \( f_j \)'s are already precluded.

1. \( f_1 \) must be the first column of an identify matrix. The first equation is identified and normalized on \( y_1 \).

2. In all remaining columns of \( F \), all elements below the diagonal must be zero, since an equation can only involve the \( y \)s in it or in the equations above it.

Without further restrictions, any upper triangular \( F \) is an admissible transformation. But with a diagonal \( \Sigma \), we have more information. Consider the second column. Since \( \Sigma \) must be diagonal, \( f_1' \Sigma f_2 = 0 \). But given \( f_1 \) in 1 above,

\[ f_1' \Sigma f_2 = \sigma_{11} f_{12} = 0 \]

so \( f_{12} = 0 \). The second column of \( F \) is now complete and is equal to the second column of \( I \). Continuing in the same manner, we find that

\[ f_1' \Sigma f_3 = 0 \quad \text{and} \quad f_2' \Sigma f_3 = 0 \]

will suffice to establish that \( f_3 \) is the third column of \( I \). In this fashion, it can be shown that the only admissible \( F \) is \( F = I \), which was to be shown. With \( \Gamma \) upper triangular, \( M(M - 1)/2 \) unknown parameters remained. That is exactly the number of restrictions placed on \( \Sigma \) when it was assumed to be diagonal.
15.3 Methods of Estimation

There are two approaches for direct estimation, both based on the principle of instrumental variables. It is possible to estimate each equation separately using a limited information estimator. But the same principle that suggests that joint estimation brings efficiency gains in the seemingly unrelated regressions setting of the previous chapter is at work here, so we shall also consider full information or system methods of estimation.

15.4 Single Equation: Limited Information Estimation Methods

15.4.1 Ordinary Least Squares

For all \( T \) observations, the nonzero terms in the \( j \)th equation are

\[
y_j = Y_j\gamma_j + X_j\beta_j + \epsilon_j = Z_j\delta_j + \epsilon_j
\]

The \( M \) reduced-form equations are

\[
Y = X\Pi + V.
\]

Likewise, \( V_j \) is \( M_j \) columns of \( V = ET^{-1} \). This least squares estimator is

\[
d_j = [Z_j'Z_j]^{-1}Z_j'y_j = \delta_j + \left[ \begin{bmatrix} Y_j'Y_j & Y_j'X_j \\ X_j'Y_j & X_j'X_j \end{bmatrix} \right]^{-1} \begin{bmatrix} Y_j'\epsilon_j \\ X_j'\epsilon_j \end{bmatrix}
\]

None of the terms in the inverse matrix converge to 0. \( \text{plim}(1/T)Y_j'\epsilon_j \) is nonzero, which means that both parts of \( d_j \) are inconsistent.

An intuitively appealing form of simultaneous equations model is the triangular system

\[
\begin{align*}
y_1 &= x'\beta_1 + \epsilon_1 \\
y_2 &= x'\beta_2 + \gamma_{12}y_1 + \epsilon_2 \\
y_3 &= x'\beta_3 + \gamma_{13}y_1 + \gamma_{23}y_2 + \epsilon_3
\end{align*}
\]

If \( \Gamma \) is triangular and \( \Sigma \) is diagonal, so that the disturbances are uncorrelated, then the system is a fully recursive model. It is easy to see that in this case, the entire system may be estimated consistently by OLS.

15.4.2 Estimation by Instrumental Variables

Returning to the structural form, we first consider direct estimation of the \( j \)th equation,

\[
y_j = Y_j\gamma_j + X_j\beta_j + \epsilon_j = Z_j\delta_j + \epsilon_j
\]

A general method of obtaining consistent estimates is the method of instrumental variables. Let \( W_j \) be a \( T \times (M_j + K_j) \) matrix that satisfies the requirements for an IV estimator,

\[
\text{plim}(1/T)W_j'Z_j = \Sigma_{wz} = \text{a finite nonsingular matrix}
\]

\[
\text{plim}(1/T)W_j'\epsilon_j = 0
\]

\[
\text{plim}(1/T)W_j'W_j = \Sigma_{ww} = \text{a positive definite matrix}
\]

126
Then the IV estimator,
\[ \hat{\delta}_{j,IV} = \left[ W_j'Z_j \right]^{-1} W_j' y_j \]
will be consistent and have asymptotic covariance matrix
\[
\text{Asy.Var}[\hat{\delta}_{j,IV}] = \frac{\sigma_{jj}}{T} \text{plim} \left[ \frac{1}{T} W_j' Z_j \right]^{-1} \left[ \frac{1}{T} W_j' W_j \right]^{-1} \left[ \frac{1}{T} Z_j' W_j \right]^{-1}
\]
A consistent estimator of \( \sigma_{jj} \) is
\[ \hat{\sigma}_{jj} = \frac{(y_j - Z_j \hat{\delta}_{j,IV})' (y_j - Z_j \hat{\delta}_{j,IV})}{T} \]
which is the familiar sum of squares of the estimated disturbances.

### 15.4.3 Two-Stage Least Squares

The method of two-stage least squares is the most common method used for estimating simultaneous-equations models.

The **two-stage least squares (2SLS)** method consists of using as the instruments for \( Y_j \) the predicted values in a regression of \( Y_j \) on all the \( x \)s in the system:
\[ \hat{Y}_j = X[(X'X)^{-1}X'Y_j]XP_j \]

The 2SLS estimator is, thus,
\[ \hat{\delta}_{j,2SLS} = \left[ \hat{Y}_j' Y_j \hat{Y}_j' X_j \right]^{-1} \left[ \hat{Y}_j' y_j \right] \]

It can also be written
\[ \hat{\delta}_{j,2SLS} = \left[ \hat{Y}_j' \hat{Y}_j \hat{Y}_j' X_j \right]^{-1} \left[ \hat{Y}_j' y_j \right] \]

A direct proof of the consistency of the 2SLS estimator requires only that we establish that it is a valid IV estimator. We require
\[
\text{plim} \left[ \frac{\hat{Y}_j' Y_j}{T} \right] = \text{plim} \left[ \frac{P_j' X'(X \Pi_j + V_j) / T}{X_j' (X \Pi_j + V_j) / T} \right] \]
to be a finite nonsingular matrix. We also require that
\[
\text{plim} \frac{1}{T} \left[ \hat{Y}_j' \epsilon_j \right] = \left[ \begin{array}{c} 0 \\ 0 \end{array} \right]
\]
By direct substitution,
\[
\text{plim} \frac{1}{T} \hat{Y}_j' X(X'X)^{-1} X' \epsilon_j = \text{plim} \left( \frac{Y_j' X}{T} \right) \left( \frac{X'X}{T} \right)^{-1} \left( \frac{X' \epsilon_j}{T} \right)
\]
The third part on the right converges to zero, whereas the other two converge to finite matrices, which confirms the result.
15.4.4 GMM Estimation

We assume that
\[ y_{jt} = z_j' \delta_j + \epsilon_{jt} \]
where \( z_{jt} = [Y_{jt}, x_{jt}] \). For the present, we will consider the heteroscedastic case only. The assumptions of the model provide the orthogonality conditions,
\[ E[x_t \epsilon_{jt}] = E[x_t (y_{jt} - z_j' \delta_j)] = 0 \]
If \( x_t \) is taken to be the full set of exogenous variables in the mode, then we obtain the criterion for the GMM estimator,
\[ q = \left[ \frac{e(z_t, \delta_j)' X}{T} \right] W_{jj}^{-1} \left[ \frac{X' e(z_t, \delta_j)}{T} \right] = \bar{m}(\delta_j) W_{jj}^{-1} \bar{m}(\delta_j) \]

where
\[ \bar{m}(\delta_j) = \frac{1}{T} \sum_{t=1}^{T} x_t (y_{jt} - z_j' \delta_j) \quad \text{and} \quad W_{jj}^{-1} = \text{the GMM weighting matrix.} \]

The extension that we can obtain here is to allow for heteroscedasticity of unknown form. The weighting matrix can be estimated with White’s consistent estimator if a consistent estimator of \( \delta_j \) is in hand with which to compute the residuals. Davidson and MacKinnon’s suggested “heteroscedastic 2SLS, or H2SLS,”. The estimator is based on the initial two-stage least squares procedure. Thus,
\[ \hat{\delta}_{j,H2SLS} = [Z_j' X (S_{0,jj})^{-1} X' Z_j]^{-1} [Z_j' X (S_{0,jj})^{-1} X' y_j] \]
where
\[ S_{0,jj} = \sum_{t=1}^{T} x_t x'_t (y_{jt} - z_j' \hat{\delta}_{j,2SLS})^2 \]

The asymptotic covariance matrix is estimated with
\[ \text{Est.Asy.Var}[\hat{\delta}_{j,H2SLS}] = [Z_j' X (S_{0,jj})^{-1} X' Z_j]^{-1} \]

15.4.5 Limited Information Maximum Likelihood and the \( k \) Class of Estimators

The limited information maximum likelihood (LIML) estimator is based on a single equation under the assumption of normally distributed disturbances; LIML is efficient among single-equation estimators.

The LIML, or least variance ratio estimator, can be computed as follows. let
\[ W_j^0 = E_j^0 E_j^0 \]
where
\[ Y_j^0 = [y_j, Y_j] \]
and
\[ E_j^0 = M_j Y_j^0 = [I - X_j (X_j' X_j)^{-1} X_j'] Y_j^0 \]
Each column of $E_0^j$ is a set of least squares residuals in the regression of the corresponding column of $Y_0^j$ on $X_j$. Thus, $W_0^j$ is the matrix of sums of squares and cross products of these residuals. Define,

$$W_1^j = E_1^j E_1^j = Y_0^j [I - X(X'X)^{-1}X'] Y_0^j$$

That is, $W_1^j$ is defined like $W_0^j$ except that the regressions are on all the $x$s in the model, not just the ones in the $j$th equation. Let

$$\lambda_1 = \text{smallest characteristic root of } (W_1^j)^{-1} W_0^j$$

Now partition $W_0^j$ into $W_0^j = [w_{0j}^j w_0^j X_j^j X_j^j w_{0j}^j]$ corresponding to $[y_j, Y_j]$, and partition $W_1^j$ likewise. Then, with these parts in hand,

$$\hat{\gamma}_{j, \text{LIML}} = [W_{1j}^j - \lambda_1 W_{jj}^1]^{-1} (w_{0j}^j - \lambda_1 w_{1j}^j)$$

and

$$\hat{\beta}_{j, \text{LIML}} = [X_j^j X_j^j]^{-1} X_j^j (y_j - Y_j \hat{\gamma}_{j, \text{LIML}})$$

The “$k$ class” of estimators is defined by the following form

$$\hat{\delta}_{j,k} = \left[ \begin{array}{cc} Y_j^j Y_j^1 - k V_j^j V_j^j & Y_j^j X_j^1 \\ X_j^j Y_j^j & X_j^j X_j^1 \end{array} \right]^{-1} \left[ \begin{array}{c} Y_j^j y_j - k V_j^j v_j \\ X_j^j y_j \end{array} \right]$$

We have already considered three members of the class, OLS with $k = 0$, 2SLS with $k = 1$, and, it can be shown, LIML with $k = \lambda_1$.

15.4.6 Two-Stage Least Squares in Models that Are Nonlinear in Variables

The analysis of simultaneous equations becomes considerable more complicated when the equations are nonlinear. A case that is broad enough to include many practical applications is the one analyzed by Kelejian (1971)

$$y_j = \gamma_{1j} f_{1j}(y, x) + \gamma_{2j} f_{2j}(y, x) + \ldots + X_j \beta_j + \epsilon_j$$

Because of the nonlinearity, it may not be possible to solve for the reduced-form equations, $h_{ij}(x) = E[f_{ij} | x]$. Kelejian shows that 2SLS based on a Taylor series approximation to $h_{ij}$, using the linear terms, higher powers, and cross-products of the variables in $x$, will be consistent. The analysis of 2SLS presented earlier then applies to the $Z_j$ consisting of $[\hat{f}_{1j}, \hat{f}_{2j}, \ldots, X_j]$.

15.5 System Methods of Estimation

We may formulate the full system of equations as

$$y = Z \delta + \epsilon,$$

where

$$E[\epsilon | X] = 0, \quad E[\epsilon' \epsilon | X] = \Sigma = \Sigma \otimes I$$
The least squares estimator
\[ d = [Z'Z]^{-1}Z'y \]
is equation-by-equation ordinary least squares and is inconsistent. Assuming that the matrix of instrumental variables, \( \bar{W} \) satisfies the requirements for an IV estimator, a consistent though inefficient estimator would be
\[ \hat{\delta}_{IV} = [\bar{W}'Z]^{-1}\bar{W}'y \]
Analogous to the seemingly unrelated regressions model, a more efficient estimator would be based on the generalized least squares principle,
\[ \hat{\delta}_{IV, GLS} = [\bar{W}'(\Sigma^{-1} \bigotimes I)Z]^{-1}\bar{W}'(\Sigma^{-1} \bigotimes I)y \]
Three techniques are generally used for joint estimation of the entire system of equations: three-stage least squares, GMM, and full information maximum likelihood.

15.5.1 Three-Stage Least Squares
Consider the IV estimator formed from
\[ \bar{W} = \hat{Z} = \text{diag}[X(X'X)^{-1}X'Z_1, ..., X(X'X)^{-1}X'Z_M] \]
The IV estimator
\[ \hat{\delta}_{IV} = [\hat{Z}'Z]^{-1}\hat{Z}'y \]
is simply equation-by-equation 2SLS. We would expect this estimator to be less efficient than a GLS estimator. A natural candidate would be
\[ \hat{\delta}_{3SLS} = [\hat{Z}'(\Sigma^{-1} \bigotimes I)Z]^{-1}\hat{Z}'(\Sigma^{-1} \bigotimes I)y \]
For this estimator to be a valid IV estimator, we must establish that
\[ \text{plim} \frac{1}{T} \hat{Z}'(\Sigma^{-1} \bigotimes I)\epsilon = 0 \]
which is \( M \) sets of equations, each one of the form
\[ \text{plim} \frac{1}{T} \sum_{j=1}^{M} \sigma^{ij} \hat{Z}_j' \epsilon_j = 0 \]
The second requirement, that
\[ \text{plim} \frac{1}{T} \hat{Z}'(\Sigma^{-1} \bigotimes I)Z \neq 0 \]
and that the matrix be nonsingular, can be established along the lines of its counterpart for 2SLS. The appropriate asymptotic covariance matrix for the estimator is
\[ \text{Asy.Var}[\hat{\delta}_{3SLS}] = [\hat{Z}(\Sigma^{-1} \bigotimes I)\hat{Z}]^{-1} \]
where \( \hat{Z} = \text{diag}[X\Pi_j, X_j] \). The three-stage least squares (3SLS) estimator is thus defined as follows:
1. Estimate $\Pi$ by ordinary least squares and compute $\hat{Y}_j$ for each equation.

2. Compute $\hat{\delta}_{j,2\text{SLS}}$ for each equation; then
   \[ \hat{\delta}_{ij} = \frac{(y_i - Z_i\hat{\delta}_i)'(y_i - Z_i\hat{\delta}_i)}{T} \]

3. Compute the GLS estimator according to
   \[ \hat{\delta}_{3\text{SLS}} = \left[ \hat{Z}'(\Sigma^{-1} \otimes I)\hat{Z} \right]^{-1}\hat{Z}'(\Sigma^{-1} \otimes I)y \]
   and an estimate of the asymptotic covariance matrix according to
   \[ \text{Asy.Var}[\hat{\delta}_{3\text{SLS}}] = \left[ \bar{Z}(\Sigma^{-1} \otimes I)\bar{Z} \right]^{-1} \]
   using $\hat{Z}$ and $\hat{\Sigma}$.

15.5.2 Full-Information Maximum Likelihood

The full-information maximum likelihood (FIML) estimator is based on the entire system of equations. With normally distributed disturbances, FIML is efficient among all estimators. We begin with the reduced form

\[ Y = X\Pi + V, \]

where each row of $V$ is assumed to be multivariate normally distributed, with $E[v_t|X] = 0$ and covariance matrix, $E[v_tv'_t|X] = \Omega$. The log-likelihood for this model is precisely that of the seemingly unrelated regressions model of chapter 14. Thus

\[ \ln L = -\frac{T}{2} \left[ M \ln(2\pi) + \ln |\Omega| + \text{tr}(\Omega^{-1}W) \right], \]

where

\[ W_{ij} = \frac{1}{T}(y - X\pi^0_i)'(y - X\pi^0_j) \]

and

\[ \pi^0_j = j\text{th column of } \Pi \]

This function is to be maximized subject to all the restrictions imposed by the structure. Make the substitutions $\Pi = -B\Gamma^{-1}$ and $\Omega = (\Gamma^{-1})\Sigma\Gamma^{-1}$ so that $\Omega^{-1} = \Gamma\Sigma^{-1}\Gamma'$. Thus

\[ \ln L = -\frac{T}{2} \left[ M \ln(2\pi) + \ln |(\Gamma^{-1})\Sigma\Gamma^{-1}| + \text{tr}\left\{ \frac{1}{T}[\Sigma^{-1}\Gamma'(Y + XB\Gamma^{-1})' (Y + XB\Gamma^{-1})]\right\} \right], \]

which can be simplified.

\[ \ln L = -\frac{1}{T} \left[ M \ln(2\pi) - 2 \ln |\Gamma| + \text{tr}(\Sigma^{-1}S) + \ln |\Sigma| \right] \]

where

\[ s_{ij} = \frac{1}{T}(Y\Gamma_i + XB_i)'(Y\Gamma_j + XB_j) \]
The trace may be written in the form
\[
\text{tr}(\Sigma^{-1}S) = \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma^{ij}(y_i - Y_i\gamma_i - X_i\beta_i)'(y_i - Y_i\gamma_i - X_i\beta_i)
\]

Maximizing \(\ln L\) subject to the exclusions above and any other restrictions, if necessary, produces the FIML estimator. The asymptotic covariance matrix for the FIML estimator is the same as that for the 3SLS estimator.

A useful interpretation of the FIML estimator is provided by Dhrymes (1973) and Hausman (1975, 1983). They show that the FIML estimator of \(\delta\) is a fixed point in the equation
\[
\hat{\delta}_{FIML} = [\hat{Z}(\hat{\delta})'(\hat{\Sigma}^{-1} \times I)Z]^{-1} \hat{Z}(\hat{\delta})'(\hat{\Sigma}^{-1} \times I)y = [\hat{Z}'Z]^{-1} \hat{Z}'y
\]
where
\[
\hat{Z}_j = [X\hat{\Pi}_j, X_j]
\]
\(\hat{\Pi}\) is computed from the structural estimates:
\[
\hat{\Pi}_j = M_j \text{ columns of } -\hat{B}\hat{\Gamma}^{-1}
\]
and
\[
\hat{\sigma}_{ij} = \frac{1}{T}(y_i - Z_i\hat{\delta}_i)'(y_i - Z_i\hat{\delta}_i) \quad \text{and} \quad \hat{\sigma}^{ij} = (\hat{\Sigma}^{-1})_{ij}
\]
This result implies that the FIML estimator is also an IV estimator.

**15.5.3 GMM Estimation**

As before, we will consider the case of unknown heteroscedasticity only. The orthogonality conditions are
\[
E[x_t\epsilon_{jt}] = E[x_t(y_{jt} - z'_{jt}\delta_j)] = 0
\]
We obtain the criterion for estimation of all the model’s parameters,
\[
q = \sum_{j=1}^{M} \sum_{l=1}^{M} \left[ \frac{e(z_t, \delta_j)'X}{T} \right] [W]^jl \left[ \frac{X'e(z_t, \delta_j)}{T} \right]
\]
\[
= \sum_{j=1}^{M} \sum_{l=1}^{M} \bar{m}(\delta_j)^jl [W]^jl \bar{m}(\delta_l)
\]
where
\[
\bar{m}(\delta_j) = \frac{1}{T} \sum_{t=1}^{T} x_t(y_{jt} - z'_{jt}\delta_j)
\]
and
\([W]^jl = \text{block } jl \text{ of the weighting matrix, } W^{-1}\)
as before, we consider the optimal weighting matrix obtained as the asymptotic covariance matrix of the empirical moments, \(\bar{m}(\delta_j)\). Then, the \(jl\)th block of \(\text{Asy.Var}[\sqrt{T}\bar{m}(\delta)]\) is
\[
\Phi_{jl} = \text{plim} \left\{ \frac{1}{T} \sum_{t=1}^{T} [x_t x'_t(y_{jt} - z'_{jt}\delta_j)(y_{lt} - z'_{lt}\delta_l)] \right\} = \text{plim} \left( \frac{1}{T} \omega_{jlt} x_t x'_t \right)
\]
For implementation, $\Phi_{jl}$ can be estimated with

$$
\hat{\Phi}_{jl} = \frac{1}{T} \sum_{t=1}^{T} x_t x'_t (y_{jt} - z'_j d_j)(y_{lt} - z'_l d_l)
$$

where $d_j$ is a consistent estimator of $\delta_j$. With this result in hand, the first-order conditions for GMM estimation are

$$
\frac{\partial \hat{q}}{\partial \delta_j} = 2 \sum_{l=1}^{M} \left( \frac{Z_j' X}{T} \right) \hat{\Phi}_{jl} \left[ \frac{X'(y_l - Z_l \delta_l)}{T} \right]
$$

where $\hat{\Phi}_{jl}$ is the $jl$th block in the inverse of the estimate if the center matrix in $q$. The solution is

$$
\begin{bmatrix}
\hat{\delta}_{1,GMM} \\
\hat{\delta}_{2,GMM} \\
\cdots \\
\hat{\delta}_{M,GMM}
\end{bmatrix} = \begin{bmatrix}
Z'_1 \hat{\Phi}^{11} X' Z_1 & Z'_1 \hat{\Phi}^{12} X' Z_2 & \cdots & Z'_1 \hat{\Phi}^{1M} X' Z_M \\
Z'_2 \hat{\Phi}^{21} X' Z_1 & Z'_2 \hat{\Phi}^{22} X' Z_2 & \cdots & Z'_2 \hat{\Phi}^{2M} X' Z_M \\
\vdots & \vdots & \ddots & \vdots \\
Z'_M \hat{\Phi}^{M1} X' Z_1 & Z'_M \hat{\Phi}^{M2} X' Z_2 & \cdots & Z'_M \hat{\Phi}^{MM} X' Z_M
\end{bmatrix}^{-1} \begin{bmatrix}
\sum_{j=1}^{M} Z'_{1j} \hat{\Phi}_{1j} y_j \\
\sum_{j=1}^{M} Z'_{2j} \hat{\Phi}_{2j} y_j \\
\vdots \\
\sum_{j=1}^{M} Z'_{Mj} \hat{\Phi}_{Mj} y_j
\end{bmatrix}
$$

Several of the estimators we have already considered are special cases:

1. If $\hat{\Phi}_{jj} = \hat{\sigma}_{jj} (X'X/T)$ and $\hat{\Phi}_{jl} = 0$ for $j \neq l$, then $\hat{\delta}_j$ is 2SLS.
2. If $\hat{\Phi}_{jl} = 0$ for $j \neq l$, then $\hat{\delta}_j$ is H2SLS, the single-equation GMM estimator.
3. If $\hat{\Phi}_{jl} = \hat{\sigma}_{jl} (X'X/T)$, then $\hat{\delta}_j$ is 3SLS.

As before, the GMM estimator brings efficiency gains in the presence of heteroscedasticity.

### 15.5.4 Recursive Systems and Exactly identified Equations

Finally, there are two special cases worth noting. First, for the fully recursive model,

1. $\Gamma$ is upper triangular, with ones on the diagonal. Therefore, $|\Gamma| = 1$ and $\ln |\Gamma| = 0$.
2. $\Sigma$ is diagonal, so $\ln |\Sigma| = \sum_{j=1}^{M} -j = 1 \ln \Sigma_{jj}$ and the trace in the exponent becomes

$$
\text{tr}(\Sigma^{-1} S) = \sum_{j=1}^{M} \frac{1}{\sigma_{jj} T} (y_j - Y_j \gamma_j - X_j \beta_j)'(y_j - Y_j \gamma_j - X_j \beta_j)
$$

The second interesting special case occurs when every equation is exactly identified. In this case, $K_j^* = M_j$ in every equation. It is straightforward to show that in this case, $2SLS = 3SLS = LIML = FIML$, and $\hat{\delta}_j = (X'Z_j)^{-1} X'y_j$. 

133
15.6 Specification Tests

The first procedure for testing the overidentifying restrictions in a model was developed by Anderson and Rubin (1950). Their likelihood ratio test statistic is a by-product of LIML estimation:

\[ LR = \chi^2[K_j^* - M_j] = T(\lambda_j - 1) \]

where \( \lambda_j \) is the root used to find the LIML estimator. The statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions. An alternative based on the Lagrange multiplier principle was proposed by Hausman (1983). Operationally, the test requires only the calculation of \( TR^2 \), where the \( R^2 \) is the uncentered \( R^2 \) in the regression of \( \hat{\epsilon}_j = y_j - Z_j\hat{\delta}_j \) on all the predetermined variables in the model. The estimated parameters may be computed using 2SLS, LIML, or any other efficient limited-information estimator. The statistic has a limiting chi-squared distribution with \( K^*_j - M_j \) degrees of freedom under the assumed specification of the model.

Suppose that the variable \( x^e \) is in question. The test is based on the existence of two estimators, say \( \hat{\delta} = \hat{\delta}^* \), such that

Under \( H_0 \) : (\( x^e \) is exogenous), both \( \hat{\delta} \) and \( \hat{\delta}^* \) are consistent and \( \hat{\delta}^* \) is asymptotically efficient,

Under \( H_1 \) : (\( x^e \) is endogenous), \( \hat{\delta} \) is consistent, but \( \hat{\delta}^* \) is inconsistent.

Hausman bases his version of the test on \( \hat{\delta} \) being the 2SLS estimator and \( \hat{\delta}^* \) being the 3SLS estimator.

A single-equation version of the test has been devised by Spencer and Berk (1981). We suppose that \( x^e \) appears in equation \( j \), so that

\[
\begin{align*}
y_j &= Y_j\gamma_j + X_j\beta_j + x^e\theta + \epsilon_j \\
&= [Y_j, X_j, x^e]\delta_j + \epsilon_j
\end{align*}
\]

Then \( \hat{\delta}^* \) is the 2SLS estimator, treating \( x^e \) as an exogenous variable in the system, whereas \( \hat{\delta} \) is the IV estimator based on regressing \( y_j \) on \( Y_j, X_j, \hat{x}^e \), where the least squares fitted values are based on all the remaining exogenous variables, excluding \( x^e \). The test statistic is then

\[
w = (\hat{\delta}^* - \hat{\delta})'\{\text{Est.Var}[\hat{\delta}] - \text{Est.Var}[\hat{\delta}^*]\}^{-1}(\hat{\delta}^* - \hat{\delta})
\]

which is the Wald statistic based on the difference of the two estimators. The statistic has one degree of freedom.

15.7 Properties of Dynamic Models

15.7.1 Dynamic Models and Their Multipliers

The structural form of a dynamic model is

\[ y_t'\Gamma + x_t'B + y_{t-1}'\Phi = \epsilon_t' \]

If the model contains additional lags, then we can add additional equations to the system of the form \( y_{t-1}' = y_{t-1}' \). The reduced form is

\[ y_t' = x_t'\Pi + y_{t-1}'\Delta + v_t' \]
where

\[ \Pi = -B \Gamma^{-1} \]

and

\[ \Delta = -\Phi \Gamma^{-1} \]

From the reduced form,

\[ \frac{\partial y_{t,m}}{\partial x_{t,k}} = \Pi_{km} \]

\(\Pi\) is the matrix of impact multipliers. By substituting for \(y_{t-1}\) we obtain,

\[ y_t = x_t' \Pi + x_{t-1}' \Pi \Delta + y_{t-2} \Delta^2 + (v_t' + v_{t-1}' \Delta) \]

Continuing this method for the full \(t\) periods, we obtain

\[ y_t = \sum_{s=0}^{t-1} [x_{t-s}' \Pi \Delta^s] + y_0' \Delta^t + \sum_{s=0}^{t-1} [v_{t-s}' \Delta^s] \]

This shows how the initial conditions \(y_0\) and the subsequent time path of the exogenous variables and disturbances completely determine the current values of the endogenous variables. The coefficient matrices in the bracketed sum are the dynamic multipliers.

\[ \frac{\partial y_{t,m}}{\partial x_{t-s,k}} = (\Pi \Delta^s)_{km} \]

If we let \(s\) go to infinity, we obtain the final form of the model,

\[ y_t = \sum_{s=0}^{\infty} [x_{t-s}' \Pi \Delta^s] + \sum_{s=0}^{\infty} [v_{t-s}' \Delta^s] \]

Then, the matrix of cumulated multipliers in the final form is

\[ \Pi[I + \Delta + \Delta^2 + \ldots] = \Pi[I - \Delta]^{-1} \]

These coefficient matrices are the long-run or equilibrium multipliers. We can also obtain the cumulated multipliers for \(s\) periods as

\[ \text{cumulated multipliers} = \Pi[I - \Delta]^{-1} [I - \Delta^s] \]

Suppose that the values of \(x\) were permanently fixed at \(\bar{x}\). Then the final form shows that if there are no disturbances, the equilibrium value of \(y_t\) would be

\[ \bar{y} = \sum_{s=0}^{\infty} [\bar{x}' \Pi \Delta^s] = \bar{x}' \sum_{s=0}^{\infty} \Pi \Delta^s = \bar{x}' \Pi[I - \Delta]^{-1} \]

Therefore, the equilibrium multipliers are

\[ \frac{\partial \bar{y}_m}{\partial x_k} = [\Pi(I - \Delta)^{-1}]_{km} \]
15.7.2 Stability

It remains to be shown that the matrix of multipliers in the final form converges. Although \( \Delta \) is not a symmetric matrix, it will still have a spectral decomposition of the form

\[
\Delta = C \Lambda C^{-1}
\]

where \( \Lambda \) is a diagonal matrix containing the characteristic roots of \( \Delta \) and each column of \( C \) is a right characteristic vector,

\[
\Delta c_m = \lambda_m c_m
\]

and

\[
\Delta^t = CA^tC^{-1}
\]

It is apparent that whether or not \( \Delta^t \) vanishes as \( t \to \infty \) depends on its characteristic roots. The condition is \( |\lambda_m| < 1 \). For the case of a complex root, \( |\lambda_m| = |a + bi| = \sqrt{a^2 + b^2} \). For a given model, the stability may be established by examining the largest or dominant root.

The model may be written

\[
\begin{bmatrix}
y'_1 \\
y'_2
\end{bmatrix} =
\begin{bmatrix}
x'_0 \Pi_1 & \Pi_2 \\
y'_{t-1,1} & y'_{t-1,2}
\end{bmatrix}
\begin{bmatrix}
\Delta_1 \\
0
\end{bmatrix} +
\begin{bmatrix}
v'_1 \\
v'_2
\end{bmatrix}
\]

The characteristic roots of \( \Delta \) are defined by the characteristic polynomial, \( |\Delta - \lambda I| = 0 \). For the partitioned model, this result is

\[
\begin{vmatrix}
\Delta_1 - \lambda I & \Delta_2 \\
0 & -\lambda I
\end{vmatrix} = 0
\]

Now obtain

\[
|\Delta - \lambda I| = (-\lambda)^{M_2}|\Delta_1 - \lambda I| = 0
\]

where \( M_2 \) is the number of variables in \( y_2 \).

15.7.3 Adjustment to Equilibrium

At time 0, \( y'_0 = x'_0 \Pi + y'_{t-1} \Delta \). But prior to time 0, the system was in equilibrium, so \( y'_0 = x'_0 \Pi + \bar{y}' \Delta \). The initial displacement due to the shock to \( \bar{x} \) is

\[
y'_0 - \bar{y}' = x'_0 \Pi_0 \bar{y}' (I - \Delta)
\]

Substituting \( \bar{x}' \Pi = \bar{y}' (I - \Delta) \) produces

\[
y'_0 - \bar{y}' = (x'_0 - \bar{x}') \Pi
\]

Since \( x_t = \bar{x} \) after period 0,

\[
y'_t = \sum_{s=0}^{t-1} \bar{x}' \Pi \Delta^s + y'_0 \Delta^t
\]

\[
= \bar{x}' \Pi (I - \Delta)^{-1} (I - \Delta^t) + y'_0 \Delta^t
\]

\[
= \bar{y} - \bar{y}' \Delta^t + y'_0 \Delta^t
\]

\[
= \bar{y} + (\bar{y}_0 - \bar{y}') \Delta^t
\]
Then,

\[ y'_t = \bar{y}' + (x'_t - \bar{x}')\Pi \Delta^t \]

Since \( \lim_{t \to \infty} \Delta^t = 0 \), the path back to the equilibrium subsequent to the exogenous shock \((x_0 - \bar{x})\) is defined. The stability condition imposed on \( \Delta \) ensures that if the system is disturbed at some point by a one-time shock, then barring further shocks or disturbances, it will return to its equilibrium. Since \( y_0, \bar{x}, x_0, \) and \( \Pi \) are fixed for all time, the shape of the path is completely determined by the behavior of \( \Delta^t \), which we now examine.

The spectral decomposition of \( \Delta^t \) given may be written

\[ \Delta^t = \sum_{m=1}^{M} \lambda^t_m c_m d'_m \]

where \( c_m \) is the \( m \)th column of \( C \) and \( d'_m \) is the \( m \)th row of \( C^{-1} \). Inserting this result gives

\[ (y_t - \bar{y})' = [(x_0 - \bar{x})'\Pi] \sum_{m=1}^{M} \lambda^t_m c_m d'_m \]

\[ = \sum_{m=1}^{M} \lambda^t_m (x_0 - \bar{x})'\Pi c_m d'_m = \sum_{m=1}^{M} \lambda^t_m g'_m \]

Since \( g_m \) depends only on the initial conditions and the parameters of the model, the behavior of the time path of \((y_t - \bar{y})\) is completely determined by \( \lambda^t_m \). The terms in the sum behave as follows:

- \( \lambda_m \) real \( \neq 0 \), \( \lambda^t_m \) adds a damped exponential term,
- \( \lambda_m \) real \( \neq 0 \), \( \lambda^t_m \) adds a damped sawtooth term,
- \( \lambda_m \) real \( \neq 0 \), \( \lambda^t_m \) adds a damped sinusoidal term,

If we write the complex root \( \lambda_m = a + bi \) in polar form, then \( \lambda = A[\cos B + i \sin B] \), where \( A = [a^2 + b^2]^{1/2} \) and \( B = \arccos(a/A) \) (in radians), the sinusoidal components each have amplitudes \( A^t \) and period \((2\pi/B)\).
Chapter 16

Estimation Frameworks in Econometrics

16.1 Parametric Estimation and Inference

The joint density of a scalar random variable, “y” and a random vector, “x” of interest can be specified by

\[ f(y, x) = g(y|x, \beta) \times h(x|\theta) \]

with unknown parameters \( \beta \) and \( \theta \). Consider the linear regression model with normally distributed disturbances. The assumption produces a full statement of the conditional density that is the population from which an observation is drawn;

\[ y_i|x_i \sim N[x'_i\beta, \sigma^2] \]

The parameter space for the parametric model is the set of allowable values of the parameters which satisfy some prior specification of the model.

16.1.1 Classical Likelihood Based Estimation

When the density of a sample of observations is completely specified, apart from the unknown parameters, then the joint density of those observations, is the likelihood function,

\[ f(y_1, y_2, \ldots, x_1, x_2, \ldots) = \prod_{i=1}^{n} f(y_i, x_i|\beta, \theta) \]

This function contains all the information available in the sample about the population from which those observations where drawn.

The maximum likelihood estimator is that function of the data which maximizes the likelihood function.

16.1.2 Bayesian Estimation

The centerpiece of the Bayesian methodology is Bayes theorem: for events \( A \) and \( B \), the conditional probability of event \( A \) given that \( B \) has occurred is

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]
Paraphrased for our applications here, we would write

\[ P(\text{parameters}|\text{data}) = \frac{P(\text{data}|\text{parameters})P(\text{parameters})}{P(\text{data})} \]

For the purpose of the study, we treat the data as only a fixed set of additional information to be used in updating our beliefs about the parameters. Thus, we write

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

The symbol \( \propto \) means “is proportional to”. The first term on the right is the joint distribution of the observed random variables \( y \), given the parameters. The second term is the prior beliefs of the analyst. The left-hand side is the posterior density of the parameters, given the current body of data, or our revised beliefs about the distribution of the parameters after “seeing” the data.

16.1.3 Bayesian Analysis of the Classical Regression Model

The likelihood for \( \beta \) and \( \sigma^2 \) given the data,

\[ L(\beta, \sigma^2|y, X) = [2\pi \sigma^2]^{-n/2} e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)} \]

Let \( d = n - K \). After a bit of manipulation, the likelihood may be written

\[ L(\beta, \sigma^2|y, X) = [2\pi]^{-d/2} [\sigma^2]^{-d/2} e^{-(d/2)(s^2/\sigma^2)} [2\pi]^{-K/2} [\sigma^2]^{-K/2} e^{-(1/2)(\beta-b)'[\sigma^2(X'X)^{-1}]^{-1}(\beta-b)} \]

Since the data are taken to be constants in the joint density, we may multiply this joint density by the (very carefully chosen), inessential constant function of the observations

\[ A = \frac{(d/2)s^2}{\Gamma(d/2+1)} [2\pi]^{d/2} [X'X]^{-1/2} \]

For convenience, let \( v = d/2 \). Then, multiplying \( L(\beta, \sigma^2|y, X) \) by \( A \) gives

\[ L(\beta, \sigma^2|y, X) \propto [vs^2]^{v+1} \frac{1}{\Gamma(v+1)} \left( \frac{1}{\sigma^2} \right)^v e^{-vs^2(1/\sigma^2)} [2\pi]^{-K/2} [\sigma^2(X'X)^{-1}]^{-1/2} e^{-(1/2)(\beta-b)'[\sigma^2(X'X)^{-1}]^{-1}(\beta-b)} \]

The likelihood function is proportional to the product of a gamma density for \( z = 1/\sigma^2 \) with parameters \( \lambda = vs^2 \) and \( P = v + 1 \) and a \( K \)-variate normal density for \( \beta|\sigma^2 \) with mean vector \( b \) and covariance matrix \( \sigma^2(X'X)^{-1} \).

The departure point for the Bayesian analysis of the model is the specification of a prior distribution. This distribution gives the analyst’s prior beliefs about the parameters of the model. One of two approaches is generally taken. If no prior information is known about the parameters, then we can specify a noninformative prior that reflects that. We do this by specifying a “flat” prior for the parameter in question:

\[ g(\text{parameter}) \propto \text{constant} \]

The second possibility, an informative prior, is treated in the next section.
To begin, we analyze the case in which \( \sigma^2 \) is assumed to be known. Using Bayes’ Theorem, we construct the posterior density,

\[
f(\beta|y, X, \sigma^2) = \frac{L(\beta|y, X, \sigma^2)g(\beta|\sigma^2)}{f(y)} \propto L(\beta|y, X, \sigma^2)g(\beta|\sigma^2)
\]

assuming that the distribution of \( X \) does not depend on \( \beta \) or \( \sigma^2 \). For now, write

\[
f(\beta|y, X, \sigma^2) \propto h(\sigma^2)[2\pi]^{-K/2}|\sigma^2(X'X)^{-1}|^{-1/2}e^{-\frac{1}{2}(\beta-b)'[\sigma^2(X'X)^{-1}]^{-1}(\beta-b)}
\]

where

\[
h(\sigma^2) = \frac{[vs^2]^{v+1}}{\Gamma(v+1)} \left( \frac{1}{\sigma^2} \right)^v e^{-vs^2(1/\sigma^2)}
\]

For the present, we treat \( h(\sigma^2) \) simply as a constant that involves \( \sigma^2 \), not as a probability density. Thus, the posterior density \( f(\beta|y, X, \sigma^2) \) is proportional to a multivariate normal distribution with mean \( b \) and covariance matrix \( \sigma^2(X'X)^{-1} \).

Assuming that \( \beta \) and \( \sigma^2 \) are independent, we now have the noninformative joint prior distribution:

\[
g(\beta, \sigma^2) = g_\beta(\beta)g_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}
\]

We can obtain the joint posterior distribution for \( \beta \) and \( \sigma^2 \) by using

\[
f(\beta, \sigma^2|y, X) = L(\beta|y, X, \sigma^2)g_{\sigma^2}(\sigma^2) \propto L(\beta|y, X, \sigma^2) \times \frac{1}{\sigma^2}
\]

For the same reason as before, we multiply \( g_{\sigma^2}(\sigma^2) \) by a well-chosen constant, this time \( vs^2\Gamma(v+1)/\Gamma(v+2) = vs^2/(v+1) \). Given \( y \) and \( X \):

\[
L(\beta, \sigma^2|y, X) \propto \frac{[vs^2]^{v+2}}{\Gamma(v+2)} \left( \frac{1}{\sigma^2} \right)^{v+1} e^{-vs^2(1/\sigma^2)}[2\pi]^{-K/2}|\sigma^2(X'X)^{-1}|^{-1/2} \times e^{-\frac{1}{2}(\beta-b)'[\sigma^2(X'X)^{-1}]^{-1}(\beta-b)}
\]

By collecting the terms, \( f(\beta, \sigma^2|y, X) \) can be written as

\[
f(\beta, \sigma^2|y, X) \propto A \times \left( \frac{1}{\sigma^2} \right)^{P-1} e^{-\lambda(1/\sigma^2)}
\]

where

\[
A = \frac{[vs^2]^{v+2}}{\Gamma(v+2)}[2\pi]^{-K/2}|\sigma^2(X'X)^{-1}|^{-1/2}
\]

\[
P = v + 2 + K/2 = (n + 4)/2
\]

\[
\lambda = vs^2 + \frac{1}{2}(\beta-b)'X'X(\beta-b)
\]

so the marginal posterior distribution for \( \beta \) is

\[
\int_0^\infty f(\beta, \sigma^2|y, X) d\sigma^2 \propto A \int_0^\infty \left( \frac{1}{\sigma^2} \right)^{P-1} e^{-\lambda(1/\sigma^2)} d\sigma^2
\]
Making the substitution, we obtain
\[
\int_0^\infty f(\beta, \sigma^2|y, X) d\sigma^2 \propto A \int_0^\infty \left( \frac{1}{\sigma^2} \right)^{P-3} e^{-\lambda(1/\sigma^2)} d \left( \frac{1}{\sigma^2} \right) = A \times \frac{\Gamma(P-2)}{\lambda^{P-2}}
\]

Reinserting the expressions for \(A, P\) and \(\lambda\) produces
\[
f(\beta|y, X) \propto \frac{[vs^2/v+2]\Gamma(v+K/2)}{\Gamma(v+2)} \left(\frac{2\pi}{X'X}\right)^{-1/2} \exp\left[-\frac{1}{2} (\beta - b)'X'(\beta - b) \right]^{v+K/2}
\]
This density is proportional to a multivariate \(t\) distribution and is a generalization of the familiar univariate distribution we have used at various points. This distribution has a degrees of freedom parameter, \(d = n - K\), mean \(b\), and covariance matrix \((d/(d - 2)) \times [s^2(X'X)^{-1}]\). Each element of the \(K\)-element vector \(\beta\) has a marginal distribution that is the univariate \(t\) distribution with degrees of freedom \(n - K\), mean \(b_k\), and variance equal to the \(k\)th diagonal element of the covariance matrix given earlier.

### 16.1.4 Interval Estimation

The counterpart to a confidence interval in this setting is an interval of the posterior distribution that contains a specified probability. An interval within which the density function is higher than any points outside it, which justifies the term highest posterior density (HPD) interval.

### 16.1.5 Estimation with an Informative Prior Density

Suppose that we assume that the prior beliefs about \(\beta\) may be summarized in \(K\)-variate normal distribution with mean \(\beta_0\) and variance matrix \(\Sigma_0\). The posterior density of \(\beta\) conditioned on \(\sigma^2\) and the data will be normal with
\[
E[\beta|\sigma^2, y, X] = (\Sigma_0^{-1} + [\sigma^2(X'X)^{-1}]^{-1})^{-1} [\Sigma_0^{-1}\beta_0 + [\sigma^2(X'X)^{-1}]b]
\]
where
\[
F = (\Sigma_0^{-1} + [\sigma^2(X'X)^{-1}]^{-1})^{-1} \Sigma_0^{-1}
\]
Taking \(\sigma^2\) as known, we can write the variance of the posterior normal distribution as
\[
\text{Var}[\beta|y, X, \sigma^2] = (\Sigma_0^{-1} + [\sigma^2(X'X)^{-1}])^{-1}
\]
A conjugate prior for \(\beta\) and \(\sigma^2\) that can be used is
\[
g(\beta, \sigma^2) = g_{\beta|\sigma^2}(\beta|\sigma^2) g_{\sigma^2}(\sigma^2)
\]
where \(g_{\beta|\sigma^2}(\beta|\sigma^2)\) is normal, with mean \(\beta^0\) and variance \(\sigma^2A\) and
\[
g_{\sigma^2}(\sigma^2) = \frac{[m\sigma^2_0]^{m+1}}{\Gamma(m+1)} \left( \frac{1}{\sigma^2} \right)^m e^{-m\sigma^2_0(1/\sigma^2)}
\]
141
This distribution is an **inverted gamma distribution**. It implies that $1/\sigma^2$ has a gamma distribution. The prior mean for $\sigma^2$ is $\sigma_0^2$ and the prior variance is $\sigma_0^4/(m-1)$. By integrating out $\sigma^2$, we would obtain the prior marginal for $\beta$ alone, which would be a multivariate $t$ distribution.

$$E[\beta|y, X] = \{[\sigma^2 A]^{-1} + [\sigma^2 (X'X)^{-1}]^{-1}\}^{-1}\{[\sigma^2 A]^{-1} \beta^0 + [\sigma^2 (X'X)^{-1}]^{-1} b\}^{-1}$$

and

$$\text{Var}[\beta|y, X] = \left(\frac{j}{j-2}\right) \{[\sigma^2 A]^{-1} + [\sigma^2 (X'X)^{-1}]^{-1}\}^{-1}$$

where $j$ is a degrees of freedom parameter and $\bar{\sigma}^2$ is the Bayesian estimate of $\sigma^2$.

### 16.1.6 Hypothesis Testing

The Bayesian approach to hypothesis testing is more appropriately called “comparing hypotheses”.

The Bayesian approach to hypothesis testing bears large similarity to Bayesian estimation. We have formulated two hypotheses, a “null,” denoted $H_0$, and an alternative, denoted $H_1$. Assume that before we begin our experimentation we are able to assign prior probabilities $P(H_0)$ and $P(H_1)$ to the two hypotheses. The **prior odds ratio** is simply the ratio

$$\text{Odds}_{\text{prior}} = \frac{P(H_0)}{P(H_1)}$$

so the posterior is, in general,

$$\text{Odds}_{\text{posterior}} = B_{01} \times \text{Odds}_{\text{prior}}$$

The value $B_{01}$ is called the **Bayes factor** for comparing the two hypotheses. It summarizes the effect of the sample data on the prior odds. We return to our first departure point, the likelihood of the data, given the parameters:

$$f(y|\beta, \sigma^2, X) = (2\pi \sigma^2)^{-n/2} e^{-1/(2\sigma^2)(y-X\beta)'(y-X\beta)}$$

Based on our priors for the parameters, the expected, or average likelihood, assuming that hypothesis $j$ is true ($j = 0, 1$) is

$$f(y|X, H_j) = E_{\beta, \sigma^2}[f(y|\beta, \sigma^2, X, H_j)] = \int_\beta \int_{\sigma^2} f(y|\beta, \sigma^2, X, H_j) g(\beta, \sigma^2) d\beta d\sigma^2$$

The posterior probability is

$$P(H_j|y, X) = \frac{f(y|X, H_j)P(H_j)}{f(y)}$$

The posterior odds ratio is $P(H_0|y, X)/P(H_1|y, X)$, so the Bayes factor is

$$B_{01} = \frac{f(y|X, H_0)}{f(y|X, H_1)}$$
16.1.7 Using Bayes Theorem in A Classical Estimation Problem: The Latent Class Model

We will assume that the data are a panel. Thus, the density of $y_{it}$ when the parameter vector is $\beta_i$ is $f(y_{it}|x_{it}, \beta_i)$. The parameter vector $\beta_i$ is randomly distributed over individuals according to

$$\beta_i = \beta + \Delta z_i + v_i$$

and where $\beta + \Delta z_i$ is the mean of the distribution, which depends on time invariant individual characteristics as well as parameters yet to be estimated, and the random variation comes from the individual heterogeneity, $v_i$. This random vector is assumed to have mean zero and covariance matrix, $\Sigma$. The conditional density of the parameters is

$$g(\beta_i|z_i, \beta, \Delta, \Sigma) = g(v_i + \beta + \Delta z_i, \Sigma),$$

where $g(\cdot)$ is the underlying marginal density of the heterogeneity. The unconditional density for $y_{it}$ is obtained by integrating over $v_i$,

$$f(y_{it}|x_{it}, z_i, \Delta) = E_{\beta_i}[f(y_{it}|x_{it}, \beta_i)] = \int_{v_i} f(y_{it}|x_{it}, \beta_i) g(v_i + \beta + \Delta z_i, \Sigma) dv_i$$

The preceding has assumed $\beta_i$ has a continuous distribution. Suppose that $\beta_i$ is generated from a discrete distribution with $J$ values, or classes, so that the distribution of $\beta_i$ is over these $J$ vectors. The corresponding model formulation is now

$$f(y_{it}|x_{it}, z_i, \Delta) = \sum_{j=1}^{J} p_{ij}(\Delta, z_i) f(y_{it}|x_{it}, \beta_j)$$

The matrix $\Delta$ contains the parameters of the discrete distribution. It has $J$ rows and $M$ columns for the $M$ variable in $z_i$. Therefore, for a group of $T$ observations, the joint density is

$$f(y_{i1}, y_{i2}, ..., y_{iT}|\beta_j, x_{i1}, x_{i2}, ..., x_{iT}) = \prod_{t=1}^{T} f(y_{it}|x_{it}, \beta_j)$$

Inserting this result in the earlier density produces the likelihood function for a panel of data,

$$\ln L = \sum_{i=1}^{n} \ln \left[ \sum_{j=1}^{M} p_{ij}(\Delta, z_i) \prod_{t=1}^{T} g(y_{it}|x_{it}, \beta_j) \right]$$

The class probabilities must be constrained to sum to 1. A simple approach is to reparameterize them as a set of logit probabilities,

$$p_{ij} = \frac{e^{\theta_{ij}}}{\sum_{j=1}^{J} e^{\theta_{ij}}}, \quad j = 1, ..., J, \quad \theta_{iJ} = 0, \quad \theta_{ij} = \delta_{iJ} z_i, \quad (\delta_{iJ} = 0)$$

Note the restriction on $\theta_{iJ}$. This is an identification restriction.

Estimation produces values for the structural parameters, $(\beta_j, \delta_j), j = 1, ..., J$. With these in hand, we can compute the prior class probabilities, $p_{ij}$. For prediction purposes,
one might be more interested in the posterior class probabilities, which we can compute using Bayes theorem as

\[
\text{Prob}(\text{class } j | \text{observation } i) = \frac{f(\text{observation } i | \text{class } j) \text{Prob}(\text{class } j)}{\sum_{j=1}^{M} f(\text{observation } i | \text{class } j) \text{Prob}(\text{class } j)}
\]

This set of probabilities, \( w_i = (w_{i1}, w_{i2}, \ldots, w_{iJ}) \) gives the posterior density over the distribution of values of \( \beta \), that is, \([\beta_1, \beta_2, \ldots, \beta_J]\). The Bayesian estimator of the parameter vector would be the posterior mean

\[
\hat{\beta}_i^p = \bar{E}_j[\beta_j | \text{observation } i] = \sum_{j=1}^{J} w_{ij} \hat{\beta}_j
\]

16.1.8 Hierarchical Bayes Estimation of a Random Parameters Model by Markov Chain Monte Carlo Simulation

For an individual \( i \), the conditional density or the dependent variable in period \( t \) is \( f(y_{it} | x_{it}, \beta_i) \) where \( \beta_i \) is the individual specific data that enter the probability density. For the sequence of \( T \) observations, assuming conditional (on \( \beta_i \)) independence, person \( i \)'s contribution to the likelihood for the sample is

\[
f(y_{i} | X_i, \beta_i) = \prod_{t=1}^{T} f(y_{it} | x_{it}, \beta_i)
\]

where \( y_i = (y_{i1}, \ldots, y_{iT}) \) and \( X_i = [x_{i1}, \ldots, x_{iT}] \). We will suppose that \( \beta_i \) is distributed normally with mean \( \beta \) and covariance matrix \( \Sigma \). The unconditional density would be the expected value over the possible values of \( \beta_i \);

\[
f(y_{i} | X_i, \beta, \Sigma) = \int \prod_{t=1}^{T} f(y_{it} | x_{it}, \beta_i) \phi_K[\beta_i | \beta, \Sigma] d\beta_i
\]

where \( \phi_K[\beta_i | \beta, \Sigma] \) denotes the \( K \) variate normal prior density for \( \beta_i \) given \( \beta \) and \( \Sigma \). For now, we consider the Bayesian approach to estimation of the parameters of this model.

A full matrix \( \Sigma \) is handled by assigning to \( \Sigma \) an inverted Wishart prior density\(^1\) with parameters scalar \( K \) and matrix \( K \times I \). This produces the joint posterior density,

\[
\Gamma(\beta_1, \ldots, \beta_n, \beta, \Sigma | \text{all data}) = \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T} f(y_{it} | x_{it}, \beta_i) \phi_K[\beta_i | \beta, \Sigma] \right\} \times p(\beta, \Sigma)
\]

This gives the joint density of all the unknown parameters conditioned on the observed data. It is at this point that the recently developed techniques of Markov Chain Monte Carlo (MCMC) simulation estimation and the Metropolis Hastings algorithm enter and enable us to do the estimation in a remarkably simple fashion.

\(^1\)The Wishart density is a multivariate counterpart to the Chi-squared distribution. Discussion may be found in Zellner (1971, pp.389-394)
The Gibbs sampler and the Metropolis-Hastings algorithm can be used for sampling from the joint density, \( \Gamma(\beta_1, ..., \beta_n, \beta, \Sigma | \text{all data}) \). This basic principle of the Gibbs sampler is as follows:

1. Draw \( x_j \) from \( f(x|y_j) \),
2. Draw \( y_{j+1} \) from \( f(y|x_j) \),
3. Exit or return to step 1.

If this process is repeated enough times, then at the last step, \((x_j, y_j)\) together are a draw from the joint distribution.

We begin by partitioning the parameters into \( \gamma = (\beta, \Sigma) \) and \( \delta = (\beta_1, ..., \beta_n) \). To obtain a draw from \( \gamma | \delta \), we will use the Gibbs sampler to obtain a draw from the distribution of \( (\beta | \Sigma, \delta) \) then one from the distribution of \( (\Sigma | \beta, \delta) \).

Conditioned on \( \delta \) and \( \Sigma \), \( \beta \) has a \( K \)-variate normal distribution with mean \( \bar{\beta} = \frac{1}{n} \sum_{i=1}^{n} \beta_i \) and covariance matrix \( (1/n)\Sigma \). To sample from this distribution we will first obtain the Cholesky factorization of \( \Sigma = LL' \) where \( L \) is a lower triangular matrix. Let \( v \) be a vector of \( K \) draws from the standard normal distribution. Then, \( \bar{\beta} + LV \) has mean vector \( \bar{\beta} + L \times 0 = \bar{\beta} \) and covariance matrix \( LIL' = \Sigma \) which is exactly what we need.

To obtain a random draw from the distribution of \( \Sigma | \beta, \delta \), we will require a random draw from the inverted Wishart distribution. The marginal posterior distribution of \( \Sigma | \beta, \delta \) is inverted Wishart with parameters scalar \( K + n \) and matrix \( W = (KI + nV) \) where \( V = \frac{1}{n} \sum_{i=1}^{n} (\beta_i - \bar{\beta})(\beta_i - \bar{\beta})' \). Let \( M \) be the lower triangular Cholesky factor of \( W^{-1} \), so \( MM' = W^{-1} \). Obtain \( K + n \) draws of \( v_k = K \) standard normal variates. Then, obtain \( S = M(\sum_{k=1}^{K+n} v_kv_k')M' \). Then, \( \Sigma = S^{-1} \) is a draw from the inverted Wishart distribution.

The difficult step is sampling \( \beta_i \). For this step, we use the Metropolis-Hastings (M-H) algorithm suggested by Chib and Greenberg and Gelman et al. The procedure involves the following steps:

1. Given \( \beta \) and \( \Sigma \) and “tuning constant” \( \tau \), compute \( d = \tau L v \) where \( L \) is the Cholesky factorization of \( \Sigma \) and \( v \) is a vector of \( K \) independent standard normal draws.
2. Create a trial value \( \beta_{i1} = \beta_{i0} + d \) where \( \beta_{i0} \) is the previous value.
3. The posterior distribution for \( \beta_i \) is the likelihood that appears in

\[
 f(y_i|X_i, \beta, \Sigma) = \int_{\beta_i} \prod_{t=1}^{T} f(fit|xit, \beta_i) \phi_K(\beta_i | \beta, \Sigma) d\beta_i
\]

as the joint normal prior density, \( \phi_K(\beta_i | \beta, \Sigma) \). Evaluate this posterior density at the trial value \( \beta_{i1} \) and the previous value \( \beta_{i0} \). Let

\[
 R_{10} = \frac{f(y_i|X_i, \beta_{i1}) \phi_K(\beta_{i1} | \beta, \Sigma)}{f(y_i|X_i, \beta_{i0}) \phi_K(\beta_{i0} | \beta, \Sigma)}
\]

4. Draw one observation, \( u \), from the standard uniform distribution, \( U[0, 1] \).
5. If \( u < R_{10} \), then accept the trial draw. Otherwise, reuse the old one.
This M-H iteration converges to a sequence of draws from the desired density. Overall, then, the algorithm uses the Gibbs sampler and the Metropolis-Hastings algorithm to produce the sequence of draws for all the parameters in the model.

Some practical details remain. The tuning constant, \( \tau \) is used to control the iteration. A smaller \( \tau \) increases the acceptance rate. But at the same time, a smaller \( \tau \) makes new draws look more like old draws so this slows down the process.

### 16.2 Semiparametric Estimation

#### 16.2.1 Least Absolute Deviations Estimation

Least squares can be severely distorted by outlying observations. These have led to the proposal of “robust” estimators that are unaffected by outlying observations. In this section, we will examine one of these, the least absolute deviations, or LAD estimator.

The LAD estimator is the solution to the optimization problem,

\[
\min_{b_0} \sum_{i=1}^{n} |y_i - x'_ib_0|
\]

The LAD estimator is a special case of the quantile regression:

\[
\text{Prob}[y_i \leq x'_i\beta] = q
\]

The LAD estimator estimates the median regression. That is, it is the solution to the quantile regression when \( q = 0.5 \). An estimator for the asymptotic covariance matrix of the quantile regression estimator,

\[
\text{Est. Asy. Var}[b_q] = (X'X)^{-1}X'DX(X'X)^{-1}
\]

where \( D \) is a diagonal matrix containing weights

\[
d_i = \left[ \frac{q}{f(0)} \right]^2 \quad \text{if } y_i - x'_i\beta \text{ is positive and } \left[ \frac{1-q}{f(0)} \right]^2 \text{ otherwise}
\]

and \( f(0) \) is the true density of the disturbances evaluated at 0.

The bootstrap method of inferring statistical properties is well suited for this application. The bootstrap estimator for the asymptotic covariance matrix can be computed as follows:

\[
\text{Est. Var}[b_{LAD}] = \frac{1}{R} \sum_{r=1}^{R} (b_{LAD}(r) - b_{LAD})(b_{LAD}(r) - b_{LAD})'
\]

where \( b_{LAD} \) is the LAD estimator and \( b_{LAD}(r) \) is the \( r \)th LAD estimate of \( \beta \) based on a sample of \( n \) observations, drawn with replacement, form the original data set.

#### 16.2.2 Partially Linear Regression

Consider a regression model in which one variable, \( x \), is of particular interest, and the functional form with respect to \( x \) is problematic. Write the model as

\[
y_i = f(x_i) + z'_i\beta + \epsilon_i
\]
where the data are assumed to be well behaved and save for the functional form, the assumptions of the classical model are met. The function $f(x_i)$ remains unspecified. Suppose the data were such that they consisted of pairs of observations $(y_{j1}, y_{j2}), j = 1, ..., n/2$ in which $x_{j1} = x_{j2}$ within every pair. If so, then estimation of $\beta$ could be based on the simple transformed model

$$y_{j2} - y_{j1} = (z_{j2} - z_{j1})'\beta + (\epsilon_{j2} - \epsilon_{j1}), \quad j = 1, ..., n/2$$

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. Suppose that the observations are sorted so that $x_1 < x_2 < ... < x_n$. Suppose, as well, that this variable is well behaved in the sense that as the sample size increases, this sorted data vector more tightly and uniformly fills the space within which $x_i$ is assumed to vary. A theory is also developed for a better differencing of groups of two or more observations. The transformed observation is $y_{d,i} = \sum_{m=0}^{M} d_m y_{i-m}$ where $\sum_{m=0}^{M} d_m = 0$ and $\sum_{m=0}^{M} d_m^2 = 1$. The pair of weights for $M = 1$ is obviously $\pm \sqrt{t}$, this is just a scaling of the simple difference, 1, -1. This estimator is shown to be consistent, asymptotically normally distributed, and have asymptotic covariance matrix

$$\text{Asy.Var}[\hat{\beta}_d] = \left(1 + \frac{1}{2M}\right) \frac{\sigma^2}{n} E_x[\text{Var}[z|z]]$$

The matrix can be estimated using the sums of squares and cross products of the differenced data. The residual variance is likewise computed with

$$\hat{\sigma}^2_v = \sum_{i=M+1}^{n} (y_{d,i} - z_{d,i}'\hat{\beta}_d)$$

Yatchew suggests that the partial residuals, $y_{d,i} - z_{d,i}'\hat{\beta}_d$ be smoothed with a kernel density estimator to provide an improved estimator of $f(x_i)$.

### 16.3 Nonparametric Estimation

#### 16.3.1 Kernel Density Estimation

The histogram is a crude density estimator. The rectangles in the figure are called bins. By construction, they are of equal width. Let $x_k$ be the midpoint of the $k$th bin and let $h$ be the width of the bin— we will shortly rename $h$ to be the bandwidth for the density estimator. The distance to the left and right boundaries of the bins are $h/2$. The frequency count in each bin is the number of observations in the sample which fall in the range $x_k \pm h/2$. Collecting terms, we have our estimator

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \left(1 - \frac{x_i - x}{h/2} < 1\right)$$

where $1(\text{statement})$ denotes an indicator function which equals 1 if the statement is true and 0 if it is false and $\text{bin}_x$ denotes the bin which has $x$ as its midpoint. The event in the indicator can be rearranged to produce an equivalent form

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \left(\frac{x_i - x}{h} < 1\right)$$
Albeit rather crude, this “naive” estimator is in the form of kernel density estimators that we have met at various points;

\[
f(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left[ \frac{x_i - x}{h} \right], \quad \text{where } K[z] = 1[-1/2 < z < 1/2]
\]

The naive estimator has several shortcomings. It is neither smooth nor continuous. More importantly, the shape of the histogram will be crucially dependent on the bandwidth, itself.

The crudeness of the weighting function in the estimator is easy to remedy. A number of candidates have been suggested.

1. Epanechnikov. \[.75(1 - .2z^2)/2.236 \text{ if } |z| \leq 5, \ 0 \text{ else}\]
2. Normal \( \phi(z) \) normal density
3. Logit \( \Gamma(z)[1 - \Gamma(z)] \) (logistic density)
4. Uniform .5 if \(|z| \leq 1, \ 0 \text{ else}\)
5. Beta \((1 - z)(1 + z)/24 \text{ if } |z| \leq 1, \ 0 \text{ else}\)
6. Cosine \( 1 + \cos(2\pi z) \) if \(|z| \leq .5, \ 0 \text{ else}\)
7. Triangle \( 1 - |z|, \text{ if } |z| \leq 1, \ 0 \text{ else}\)
8. Parzen \( 4/3 - 8z^2 + 8|z|^3 \text{ if } |z| \leq .5, \ 8(1 - |z|)^3/3 \text{ else}\)

It has been observed that in constructing density estimator, the choice of kernel function is rarely crucial, and is usually minor in importance compared to the more difficult problem of choosing the bandwidth.

The kernel density function is an estimator. For any specific \( x \), \( \hat{f}(x) \) is a sample statistic,

\[
\hat{f}(z) = \frac{1}{n} \sum_{i=1}^{n} g(x_i|z,h)
\]

Common applications typically use a bandwidth equal to some multiple of \( n^{-1/5} \) for this reason.

16.3.2 Nonparametric Regression

The regression function of a variable \( y \) on a single variable \( x \) is specified as

\[
y = \mu(x) + \epsilon
\]

Functional form are made at the outset; \( \mu(x) \) may be quite nonlinear. The general class may be defined by a conditional mean estimating function

\[
\hat{\mu}(x^*) = \sum_{i=1}^{n} w_i(x^*|x_1, x_2, ..., x_n)y_i = \sum_{i=1}^{n} w_i(x^*)y_i
\]

where the weights sum to 1. The locally weighted smoothed regression estimator is based on explicitly defining a neighborhood of points that is close to \( x^* \). This requires the choice
of a bandwidth, \( h \). A suitable weight is then required. Cleveland (1979) recommends the tricube weight,

\[
T_i(x^*|x, h) = \left[ 1 - \left( \frac{|x_i - x^*|}{h} \right)^3 \right]^3
\]

Combining terms,

\[
w_i(x^*|x, h) = 1(x_i \text{ in the neighborhood } \times T_i(x^*|x))
\]

As always, the bandwidth is crucial. This produces the kernel weighted regression estimator

\[
\hat{\mu}(x^*|x, h) = \frac{\sum_{i=1}^{n} \frac{1}{h} K \left[ \frac{x_i - x^*}{h} \right] y_i}{\sum_{i=1}^{n} \frac{1}{h} K \left[ \frac{x_i - x^*}{h} \right]}
\]

which has become a standard tool in nonparametric analysis.

### 16.4 Properties of Estimators

#### 16.4.1 Statistical Properties of Estimators

Properties that we have considered are as follows:

1. Unbiasedness
2. Consistency
3. Asymptotic normality
4. Asymptotic efficiency

#### 16.4.2 Extremum Estimators

An **extremum estimator** is one which is obtained as the optimizer of a **criterion function** \( q(\theta|\text{data}) \). Three that have occupied much of our effort thus far are

1. Least squares: \( \hat{\theta}_{LS} = \text{Argmax}\left[ -\frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i, \theta_{LS}))^2 \right] \)
2. Maximum likelihood: \( \hat{\theta}_{ML} = \text{Argmax}\left[ \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|x_i, \theta_{ML}) \right] \)
3. GMM: \( \hat{\theta}_{GMM} = \text{Argmax}\left[ -\bar{m}(\text{data, } \theta_{GMM})' W \bar{m}(\text{data, } \theta_{GMM}) \right] \)

#### 16.4.3 Assumptions for Asymptotic Properties of Extremum Estimators

Some broad results are needed in order to establish the asymptotic properties of the classical conventional extremum estimators noted above

1. The parameter space must be convex and the parameter vector that is the object of estimation must be a point in its interior.
2. The criterion function must be concave in the parameters.
3. Identifiability of the parameters.
4. Behavior of the data
16.4.4 Asymptotic Properties of Estimators

**THEOREM Consistency of M Estimators**
If all the assumptions above are satisfied, then the M estimator converges in probability to the true parameter vector.

**THEOREM Asymptotic Normality of M Estimators**
If

1. $\hat{\theta}$ is a consistent estimator of $\theta_0$ where $\theta_0$ is a point in the interior of the parameter space;
2. $q(\theta|\text{data})$ is concave and twice continuously differentiable in $\theta$ in a neighborhood of $\theta_0$;
3. $\sqrt{n}[\partial q(\theta_0|\text{data})/\partial \theta_0] \xrightarrow{d} N[0, \Phi]$
4. For any $\theta$ in $\Theta$, $\lim_{n \to \infty} Pr[(\partial^2 q(\theta_0|\text{data})/\partial \theta_k \partial \theta_m) - h_{km}(\theta)] > \epsilon = 0 \forall \epsilon > 0$ where $h_{km}(\theta)$ is a continuous finite valued function of $\theta$;
5. The matrix of elements $H(\theta)$ is nonsingular at $\theta_0$, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, [H^{-1}(\theta_0) \Phi H^{-1}(\theta_0)])$$
Chapter 17

Maximum Likelihood Estimation

17.1 The Likelihood Function and Identification of the Parameters

The probability density function, or pdf for a random variable $y$, conditioned on a set of parameters, $\theta$, is denoted $f(y|\theta)$. The joint density of $n$ independent and identically distributed (iid) observations from this process is the product of the individual densities;

$$f(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y)$$

This joint density is the likelihood function, defined as a function of the unknown parameter vector, $\theta$, where $y$ is used to indicate the collection of sample data.

It is usually simpler to work with the log of the likelihood function:

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

In closer detail,

$$\ln L(\theta|y, X) = \sum_{i=1}^{n} \ln f(y_i|x_i, \theta) = -\frac{1}{2} \sum_{i=1}^{n} [\ln \sigma^2 + \ln(2\pi) + (y_i - x_i'\beta)^2/\sigma^2]$$

where $X$ is the $n \times K$ matrix of data with $i$th row equal to $x_i'$.

**DEFINITION Identification**

The parameter vector $\theta$ is identified if for any other parameter vector, $\theta^* \neq \theta$, for some data $y$, $L(\theta^*|y) \neq L(\theta|y)$

17.2 Properties of Maximum Likelihood Estimators

**DEFINITION Asymptotic Efficiency**

An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed (CAN), and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.
If certain regularity conditions are met, the MLE will have these properties. We use the following notation: \( \hat{\theta} \) is the maximum likelihood estimator; \( \theta_0 \) denotes the true value of the parameter vector; \( \theta \) denotes another possible value of the parameter vector, not the MLE and not necessarily the true values. Then we have the following theorem,

**THEOREM  Properties of an MLE**
Under regularity, the maximum likelihood estimator (MLE) has the following asymptotic properties:

M1. Consistency: \( \text{plim} \hat{\theta} = \theta_0 \),

M2. Asymptotic normality: \( \hat{\theta} \overset{a}{\sim} N[\theta_0, \{I(\theta_0)\}^{-1}] \), where

\[
I(\theta_0) = -E_0[\partial^2 \ln L/\partial \theta \partial \theta']
\]

M3. Asymptotic efficiency: \( \hat{\theta} \) is asymptotically efficient and achieves the Cramer-Rao lower bound for consistent estimators,

M4. Invariance: The maximum likelihood estimator of \( \gamma_0 = c(\theta_0) \) is \( c(\hat{\theta}) \) if \( c(\theta_0) \) is a continuous and continuously differentiable function.

### 17.2.1 Regularity Conditions

**DEFINITION  Regularity Conditions**

R1. The first three derivatives of \( \ln f(y_i|\theta) \) with respect to \( \theta \) are continuous and finite for almost all \( y_i \) and for all \( \theta \). This condition ensures the existence of a certain Taylor series approximation and the finite variance of the derivatives of \( \ln L \).

R2. The conditions necessary to obtain the expectations of the first and second derivatives of \( \ln f(y_i|\theta) \) are met.

R3. For all values of \( \theta, |\partial^3 \ln f(y_i|\theta)/\partial \theta_j \partial \theta_k \partial \theta_l| \) is less than a function that has a finite expectation. This condition will allow us to truncate the Taylor series.

### 17.2.2 Properties of Regular Densities

Densities that are "regular" by definition above have three properties which are used in establishing the properties of maximum likelihood estimators:

**THEOREM  Moments of the Derivatives of the Log-Likelihood**

D1. \( \ln f(y_i|\theta), g_i = \partial \ln f(y_i|\theta)/\partial \theta, \) and \( H_i = \partial^2 \ln f(y_i|\theta)/\partial \theta_0 \partial \theta'_l, i = 1, \ldots, n \) are all random samples of random variables. This statement follows from our assumption of random sampling. The notation \( g_i(\theta_0) \) and \( H_i(\theta_0) \) indicates the derivative evaluated at \( \theta_0 \).

D2. \( E_0[g_i(\theta_0)] = 0 \)

D3. \( \text{Var}[g_i(\theta_0)] = -E[H_i(\theta_0)] \)

Condition D1 is simply a consequence of the definition of the density.
For the moment, we allow the range of $y_i$ to depend on the parameters; $A(\theta_0) \leq y_i \leq B(\theta_0)$. By definition,
\[
\int_{B(\theta_0)}^{A(\theta_0)} f(y_i|\theta_0) dy_i = 1
\]

Now, differentiate this expression with respect to $\theta$. Leibnitz’s theorem gives
\[
\frac{f_{A(\theta_0)} f(y_i|\theta_0) dy_i}{\partial \theta} = \int_{B(\theta_0)}^{A(\theta_0)} \frac{f(y_i|\theta_0)}{\partial \theta} dy_i + f(B(\theta_0)|\theta_0) \frac{\partial B(\theta_0)}{\partial \theta} - f(A(\theta_0)|\theta_0) \frac{\partial A(\theta_0)}{\partial \theta}
\]

Sufficient conditions are that the range of the observed random variable, $y_i$, does not depend on the parameters, which means that $\partial A(\theta_0)/\partial \theta_0 = \partial B(\theta_0)/\partial \theta_0 = 0$ or that the density is zero at the terminal points.
\[
\frac{\partial}{\partial \theta_0} \int f(y_i|\theta_0) dy_i = \int \frac{\partial f(y_i|\theta_0)}{\partial \theta_0} dy_i = \int \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} f(y_i|\theta_0) dy_i = E_0 \left[ \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \right] = 0
\]
This proves D2. We differentiate under the integral once again to obtain
\[
\int \left[ \frac{\partial^2 \ln f(y_i|\theta_0)}{\partial \theta_0 \partial \theta_0} f(y_i|\theta_0) + \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \frac{\partial f(y_i|\theta_0)}{\partial \theta_0} \right] dy_i = 0
\]
But
\[
\frac{\partial f(y_i|\theta_0)}{\partial \theta_0} = f(y_i|\theta_0) \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0}
\]
and the integral of a sum is the sum of integrals. Therefore,
\[
- \int \left[ \frac{\partial^2 \ln f(y_i|\theta_0)}{\partial \theta_0 \partial \theta_0} \right] f(y_i|\theta_0) dy_i = \int \left[ \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \frac{\partial \ln f(y_i|\theta_0)}{\partial \theta_0} \right] f(y_i|\theta_0) dy_i = [0]
\]
The left-hand side of the equation is the negative of the expected second derivatives matrix. The right-hand side is the expected square of the first derivative vector. But, since this vector has expected value 0, the right-hand side is the variance of the first derivative vector, which proves D3.

17.2.3 The Likelihood Equation

The log-likelihood function is
\[
\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta)
\]
The first derivative vector, or score vector, is
\[
g = \frac{\partial \ln L(\theta|y)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta)}{\partial \theta} = \sum_{i=1}^{n} g_i
\]
Since we are just adding terms, it follows from D1 and D2 that at $\theta_0$
\[
E_0 \left[ \frac{\partial \ln L(\theta_0|y)}{\partial \theta_0} \right] = E_0[g_0] = 0
\]
which is the likelihood equation mentioned earlier.
17.2.4 The Information Matrix Equality

The Hessian of the log-likelihood is

\[ H = -\frac{\partial^2 \ln L(\theta|y)}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^{N} H_i \]

Evaluating once again at \( \theta_0 \), by taking

\[ E_0[g_0g_0'] = E_0\left[ \sum_{i=1}^{n} \sum_{j=1}^{n} g_{0i}g_{0j}' \right] \]

Dropping terms with unequal subscripts we obtain

\[ E_0[g_0g_0'] = E_0\left[ \sum_{i=1}^{n} g_{0i}g_{0i}' \right] = E_0\left[ \sum_{i=1}^{n} (-H_{0i}) \right] = -E_0[H_0] \]

so that

\[ \text{Var}_0 \left[ \frac{\partial \ln L(\theta_0|y)}{\partial \theta_0} \right] = E_0 \left[ \frac{\partial \ln L(y_i|\theta_0) \partial \ln L(y_i|\theta_0)}{\partial \theta_0 \partial \theta_0'} \right] = -E_0 \left[ \frac{\partial^2 \ln L(\theta_0|y)}{\partial \theta_0 \partial \theta_0'} \right] \]

This very useful result is known as the information matrix equality.

17.2.5 Asymptotic Properties of the Maximum Likelihood Estimator

Consistency

Since \( \hat{\theta} \) is the MLE, in any finite sample, for any \( \theta \neq \hat{\theta} \) it must be true that

\[ \ln L(\hat{\theta}) \geq \ln L(\theta) \]

Since the log function is strictly concave, from Jensen’s Inequality, we have

\[ E_0 \left[ \log \frac{L(\theta)}{L(\theta_0)} \right] < \log E_0 \left[ \frac{L(\theta)}{L(\theta_0)} \right] \]

The expectation on the right hand side is exactly equal to one, as

\[ E_0 \left[ \frac{L(\theta)}{L(\theta_0)} \right] = \int \left( \frac{L(\theta)}{L(\theta_0)} \right) L(\theta_0) dy = 1 \]

is simply the integral of a joint density. Now, take logs on both sides, then divide by \( n \) to produce

\[ E_0[1/n \ln L(\theta_0)] - E_0[1/n \ln L(\theta_0)] < 0 \]

**THEOREM Likelihood Inequality**

\[ E_0[1/n \ln L(\theta_0)] > E_0[1/n \ln L(\theta)] \quad \text{for any} \ \theta \neq \theta_0 \quad \text{(including} \ \hat{\theta}) \]
In words, the expected value of the log-likelihood is maximized at the true value of the parameters. For any \( \theta \), including \( \hat{\theta} \),

\[
(1/n) \ln L(\theta) = (1/n) \sum_{i=1}^{n} \ln f(y_i | \theta)
\]

is the sample mean of \( n \) iid random variables, with expectation \( E[1/n \ln L(\theta_0)] \). We can invoke the Khinchine Theorem, the sample mean converges in probability to the population mean. Using \( \theta = \hat{\theta} \), as \( n \to \infty \), \( \lim \text{Prob}\{[1/n \ln L(\hat{\theta})] - [1/n \ln L(\theta_0)] \geq 1\} = 0 \), if \( \hat{\theta} \neq \theta_0 \). But, \( \hat{\theta} \) is the MLE, so for every \( n \), \( [1/n \ln L(\hat{\theta})] \geq [1/n \ln L(\theta_0)] \). The only way these can both be true is if \( 1/n \) times the sample log-likelihood evaluated at the MLE converges to the population expectation of \( 1/n \) times the log-likelihood evaluated at the true parameters. There remains one final step. Does \( [1/n \ln L(\hat{\theta})] \to [1/n \ln L(\theta_0)] \) imply that \( \hat{\theta} \to \theta_0 ? \) If there is a single parameter and the likelihood function is one to one, then clearly so. If the likelihood is strictly continuous and twice differentiable, which we assumed in the regularity conditions, and if the parameters of the model are identified which we assumed at the beginning of this discussion, then yes, it does, so we have the result.

### Asymptotic Normality

At the maximum likelihood estimator, the gradient of the log-likelihood equals zero

\[
g(\hat{\theta}) = 0
\]

Expand this set of equations in a second-order Taylor series around the true parameters \( \theta_0 \).

\[
g(\hat{\theta}) = g(\theta_0) + H(\hat{\theta})(\hat{\theta} - \theta_0) = 0
\]

The Hessian is evaluated at a point \( \hat{\theta} \) that is between \( \hat{\beta} \) and \( \theta_0 \) (\( \hat{\theta} = w\hat{\theta} + (1 - w)\theta_0 \) for some \( 0 < w < 1 \)). We then rearrange this function and multiply the result by \( \sqrt{n} \) to obtain

\[
\sqrt{n}(\hat{\theta} - \theta_0) = [-H(\hat{\theta})]^{-1}[\sqrt{n}g(\theta_0)]
\]

Because \( \text{plim}(\hat{\theta} - \theta_0) = 0 \), \( \text{plim}(\hat{\theta} - \theta) = 0 \) as well. If the limiting distribution exists, then

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} [-H(\theta_0)]^{-1}[\sqrt{n}g(\theta_0)]
\]

By dividing \( H(\theta_0) \) and \( g(\theta_0) \) by \( n \), we obtain

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \left[ -\frac{1}{n} H(\theta_0) \right]^{-1}[\sqrt{n}g(\theta_0)]
\]

We may apply the Lindberg-Levy central limit theorem to \( [\sqrt{n}g(\theta_0)] \)

\[
\sqrt{n}g(\theta_0) \xrightarrow{d} N\left[ 0, -E_0 \left( \frac{1}{n} H(\theta_0) \right) \right]
\]

We can combine results to obtain

\[
\left[ -\frac{1}{n} H(\theta_0) \right]^{-1}[\sqrt{n}g(\theta_0)] \xrightarrow{d} N\left[ 0, \left\{ -E_0 \left( \frac{1}{n} H(\theta_0) \right) \right\}^{-1}\left\{ -E_0 \left( \frac{1}{n} H(\theta_0) \right) \right\}^{-1}\right]
\]

or

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left[ 0, \left\{ -E_0 \left( \frac{1}{n} H(\theta_0) \right) \right\}^{-1}\right]
\]

which gives the asymptotic distribution of the MLE

\[
\hat{\theta} \xrightarrow{d} N[\theta_0, \{I(\theta_0)\}^{-1}]
\]

155
Asymptotic Efficiency

**THEOREM  Cramer-Rao Lower Bound**
Assuming that the density of \(y\) satisfies the regularity conditions R1-R3, the asymptotic variance of a consistent and asymptotically normally distributed estimator of the parameter vector \(\theta_0\) will always be at least as large as

\[
[I(\theta_0)]^{-1} = \left( -E_0 \left[ \frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta'_0} \right] \right)^{-1} = \left( E_0 \left[ \left( \frac{\partial \ln L(\theta_0)}{\partial \theta_0} \right) \left( \frac{\partial \ln L(\theta_0)}{\partial \theta'_0} \right) \right] \right)^{-1}
\]

The asymptotic variance of the MLE is, in fact equal to the Cramer-Rao Lower Bound for the variance of a consistent estimator, so this completes the argument.

Invariance

The invariance property is a mathematical result of the method of computing MLEs; it is not a statistical result as such. Suppose the normal log-likelihood is parameterized in terms of the precision parameter, \(\theta^2 = 1/\sigma^2\). The log-likelihood becomes

\[
\ln L(\mu, \theta^2) = -(n/2) \ln(2\pi) + (n/2) \ln \theta^2 - \frac{\theta^2}{2} \sum_{i=1}^{n} (y_i - \mu)^2
\]

The MLE for \(\mu\) is clearly still \(\bar{x}\). But the likelihood equation for \(\theta^2\) is now

\[
\frac{\partial \ln L(\mu, \theta^2)}{\partial \theta^2} = 2 \left[ \frac{n}{\theta^2} - \sum_{i=1}^{n} (y_i - \mu)^2 \right] = 0
\]

which has solution \(\hat{\theta}^2 = n \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = 1/\hat{\sigma}^2\), as expected.

**17.2.6 Estimating the Asymptotic Variance of the Maximum Likelihood Estimator**

If the form of the expected values of the second derivatives of the log-likelihood is known, then

\[
[I(\theta_0)]^{-1} = \left( -E_0 \left[ \frac{\partial^2 \ln L(\theta_0)}{\partial \theta_0 \partial \theta'_0} \right] \right)^{-1}
\]

can be evaluated at \(\hat{\theta}\) to estimate the covariance matrix for the MLE. This estimator will rarely be available. A second estimator is

\[
[\hat{I}(\hat{\theta})]^{-1} = \left( -E_0 \left[ \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right] \right)^{-1}
\]

This estimator is computed simply by evaluating the actual second derivatives matrix of the log-likelihood function at the maximum likelihood estimates. A third estimator based on result that the expected second derivatives matrix is the covariance matrix of the first derivatives vector is

\[
[\hat{I}(\hat{\theta})]^{-1} = \left[ \sum_{i=1}^{n} \hat{g}_i \hat{g}_i' \right]^{-1} = [\hat{G}' \hat{G}]^{-1}
\]
where
\[ \hat{g}_i = \frac{\partial \ln f(x_i, \hat{\theta})}{\partial \hat{\theta}} \]
and
\[ \hat{G} = [\hat{g}_1, \hat{g}_2, ..., \hat{g}_n]' \]
\( \hat{G} \) is an \( n \times K \) matrix with \( i \)th row equal to the transpose of the \( i \)th vector of derivatives in the terms of the log-likelihood function. This estimator is extremely convenient, in most cases, because it does not require any computations beyond those required to solve the likelihood equation. It has the added virtue that it is always nonnegative definite. This estimator is known as the BHHH estimator and the outer product of gradients, or OPG, estimator.

### 17.2.7 Conditional Likelihoods and Econometric Models

By partitioning the joint density of \( y_i \) and \( x_i \) into the product of the conditional and the marginal, the log-likelihood function may be written
\[
\ln L(\alpha|\text{data}) = \sum_{i=1}^{n} \ln f(y_i, x_i|\alpha) = \sum_{i=1}^{n} \ln f(y_i|x_i, \alpha) + \sum_{i=1}^{n} \ln g(x_i|\alpha)
\]
where any nonstochastic elements in \( x_i \) such as a time trend or dummy variable, are being carried as constants. We will assume that the process generating \( x_i \) takes place outside the model of interest. We partition \( \alpha \) into \( [\theta, \delta] \) so that the log-likelihood function may be written
\[
\ln L(\theta, \delta|\text{data}) = \sum_{i=1}^{n} \ln f(y_i, x_i|\alpha) = \sum_{i=1}^{n} \ln f(y_i|x_i, \theta) + \sum_{i=1}^{n} \ln g(x_i|\delta)
\]
As long as \( \theta \) and \( \delta \) have no elements in common and no restrictions connect them, then the two parts of the log-likelihood may be analyzed separately. For present purposes, the following minimal assumptions should suffice:

1. Parameter space. Parameter spaces that have gaps and nonconvexities in them will generally disable these procedures.
2. Identifiability. Estimation must be feasible
3. Well behaved data. Laws of large numbers apply to sample means involving the data and some form of central limit theorem can be applied to the gradient.

### 17.3 Three Asymptotically Equivalent Test Procedures

We consider maximum likelihood estimation of a parameter \( \theta \) and a test of the hypothesis \( H_0 : c(\theta) = 0 \). There are three approaches to testing the hypothesis.

1. Likelihood ratio test. If the restriction \( c(\theta) = 0 \) is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on difference, \( \ln L_U - \ln L_R \), where \( L_U \) is the value of the likelihood function at the unconstrained value of \( \theta \) and \( L_R \) is the value of the likelihood function at the restricted estimate.
2. Wald test. If the restriction is valid, then \( c(\hat{\theta}_{MLE}) \) should be close to zero since the MLE is consistent. Therefore, the test is based on \( c(\hat{\theta}_{MLE}) \). We reject the hypothesis if this value is significantly different from zero.

3. Lagrange multiplier test. If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave rather differently in a small sample.

17.3.1 The Likelihood Ratio Test

Let \( \theta \) be a vector of parameters to be estimated, and let \( H_0 \) specify some sort of restriction on these parameters. Let \( \hat{\theta}_U \) be the maximum likelihood estimator of \( \theta \) obtained without regard to the constraints, and let \( \hat{\theta}_R \) be the constrained maximum likelihood estimator. If \( \hat{L}_R \) and \( \hat{L}_U \) are the likelihood functions evaluated at these two estimates, then the likelihood ratio is

\[
\lambda = \frac{\hat{L}_R}{\hat{L}_U}
\]

This function must be between zero and one. Both likelihoods are positive, and \( \hat{L}_R \) cannot be larger than \( \hat{L}_U \). If \( \lambda \) is too small, then doubt is cast on the restrictions. The formal test procedure is based on the following result:

**THEOREM Limiting Distribution of the Likelihood Ratio Test Statistic**

Under regularity and under \( H_0 \), the large sample distribution of \(-2 \ln \lambda\) is chi-squared, with degrees of freedom equal to the number of restrictions imposed.

17.3.2 The Wald Test

**THEOREM Limiting Distribution of the Wald Test Statistic**

The Wald statistic is

\[
W = [c(\hat{\theta}) - q]'(\text{Asy.Var}[c(\hat{\theta}) - q])^{-1}[c(\hat{\theta}) - q]
\]

Under \( H_0 \), in large samples, \( W \) has a chi-squared distribution with degrees of freedom equal to the number of restrictions. [i.e. the number of equations in \( c(\hat{\theta}) - q \)].

For testing a set of linear restrictions \( R\theta = q \), the Wald test would be based on

\[
H_0 : c(\theta) - q = R\theta - q = 0
\]

\[
\hat{C} = \left[ \frac{\partial c(\hat{\theta})}{\partial \theta'} \right] = R'
\]

\[
\text{Est.Asy.Var}[c(\hat{\theta}) - q] = R\text{Est.Asy.Var}[\hat{\theta}]R
\]

and

\[
W = [\hat{R}\hat{\theta} - q]'[R\text{Est.Asy.Var}[\hat{\theta}]R]^{-1}[\hat{R}\hat{\theta} - q]
\]
The degrees of freedom is the number of rows in $R$. If the test is

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

then the earlier test is based on

$$z = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}$$

where $s(\hat{\theta})$ is the estimated asymptotic standard error. The Wald test will be based on

$$W = \left(\frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}\right)^2 \frac{\text{Asy.Var}[\hat{\theta}]}{\text{Asy.Var}[(\hat{\theta} - \theta_0) - 0]} = z^2$$

Here $W$ has a chi-squared distribution with one degree of freedom.

To summarize, the Wald test is based on measuring the extent to which the unrestricted estimates fail to satisfy the hypothesized restrictions. There are two short-comings of the Wald test. First, it is a pure significance test against the null hypothesis, not necessarily for a specific alternative hypothesis. As such, its power may be limited in some settings. The second shortcoming is not shared by either of the other test statistics discussed here. The Wald statistic is not invariant to the formulation of the restrictions.

### 17.3.3 The Lagrange Multiplier Test

The third test procedure is the Lagrange multiplier (LM) or efficient score (or just score) test. It is based on estimating the restricted model instead of the unrestricted model.

**THEOREM Limiting Distribution of Lagrange Multiplier Statistic**

The Lagrange Multiplier test statistic is

$$LM = \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta_R}\right)^\prime \left[I(\hat{\theta}_R)\right]^{-1} \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta_R}\right)$$

Under the null hypothesis, LM has a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. All terms are computed at restricted estimator.

The LM statistic has a useful form. Let $\hat{g}_{iR}$ denote the $i$th term in the gradient of the log-likelihood function. Then,

$$\hat{g}_R = \sum_{i=1}^n \hat{g}_{iR} = \hat{G}_R^i$$

where $\hat{G}_R$ is the $n \times K$ matrix with $i$th row equal to $g_{iR}^\prime$ and $i$ is a column of 1s. If we use the BHHH (outer product of gradients) estimator to estimate the Hessian then

$$[\hat{I}(\hat{\theta})]^{-1} = [\hat{G}_R^\prime \hat{G}_R]^{-1}$$

and

$$LM = i^\prime \hat{G}_R [\hat{G}_R^\prime \hat{G}_R]^{-1} \hat{G}_R^i$$

Now, since $i^\prime i$ equals $n$, $LM = n(i^\prime \hat{G}_R [\hat{G}_R^\prime \hat{G}_R]^{-1} \hat{G}_R^i / n) = nR_i^2$, which is $n$ times the uncentered squared multiple correlation coefficient in a linear regression of a column of 1s on the derivatives of the log-likelihood function computed at the restricted estimator.
17.4 Application of Maximum Likelihood Estimation

17.4.1 The Normal Linear Regression Model

The linear regression model is

\[ y_i = x_i' \beta \]

The likelihood function for a sample of \( n \) independent, identically and normally distributed disturbances is

\[ L = (2\pi \sigma^2)^{-n/2} e^{-e' e/(2\sigma^2)} \]

Making the transformation, we find that the likelihood function for the \( n \) observations on the observed random variable is

\[ L = (2\pi \sigma^2)^{-n/2} e^{-1/(2\sigma^2)(y-X\beta)'(y-X\beta)/2\sigma^2} \]

Taking logs, we obtain the log-likelihood function for the classical regression model:

\[ \ln L = -n \ln 2\pi - n \ln \sigma^2 - \frac{(y-X\beta)'(y-X\beta)}{2\sigma^2} \]

The necessary conditions for maximizing this log-likelihood are

\[ \left[ \frac{\partial \ln L}{\partial \beta} \right] = \left[ \frac{X'(y-X\beta)}{\sigma^2} + \frac{e' e}{2\sigma^2} \right] = \left[ \begin{array}{c} 0 \\ 0 \end{array} \right] \]

The values that satisfy these equations are

\[ \hat{\beta}_{ML} = (X'X)^{-1} X'y = b \quad \text{and} \quad \hat{\sigma}^2_{ML} = \frac{e' e}{n} \]

The Cramer-Rao bound for the variance of an unbiased estimator is the negative inverse of the expectation of

\[ \left[ \begin{array}{cc} \frac{\partial^2 \ln L}{\partial \beta_1 \partial \beta_2} & \frac{\partial^2 \ln L}{\partial \beta_1 \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \beta_2 \partial \sigma^2} & \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} \end{array} \right] = \left[ \begin{array}{cc} -\frac{X'X}{\sigma^4} & -\frac{X'e}{\sigma^4} \\ -\frac{e' X}{2\sigma^4} & \frac{n}{2\sigma^4} - \frac{e' e}{\sigma^4} \end{array} \right] \]

In taking expected values, the off-diagonal term vanishes leaving

\[ [I(\beta, \sigma^2)]^{-1} = \left[ \begin{array}{cc} \sigma^2 (X'X)^{-1} & 0 \\ 0' & 2\sigma^4/n \end{array} \right] \]

We showed earlier that \( s^2 = e' e/(n-K) \) is an unbiased estimator of \( \sigma^2 \). Therefore, the maximum likelihood estimator is biased toward zero:

\[ E[\hat{\sigma}^2_{ML}] = \frac{n-K}{n} \sigma^2 = \left( 1 - \frac{K}{n} \right) \sigma^2 < \sigma^2 \]

We know that

\[ \sqrt{n}(\hat{\sigma}^2_{ML} - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4/n) \]

It follows

\[ z_n = \left( 1 - \frac{K}{n} \right) \sqrt{n}(\hat{\sigma}^2_{ML} - \sigma^2) + \frac{K}{\sqrt{n}} \sigma^2 \xrightarrow{d} N(0, 2\sigma^4) + \frac{K}{\sqrt{n}} \sigma^2 \]
But $K/\sqrt{n}$ and $K/n$ vanish as $n \to \infty$, so the limiting distribution of $z_n$ is also $N[0, 2\sigma^2]$. Since $z_n = \sqrt{n}(s^2 - \sigma^2)$, we have shown that the asymptotic distribution of $s^2$ is the same as that of the maximum likelihood estimator.

The standard test statistic for assessing the validity of a set of linear restrictions in the linear model, $R\beta - q = 0$, so the $F$ ratio,

$$F[J, n - K] = \frac{\left( e' e - e' e / (n - K) \right) / J}{e' e / (n - K)} = \frac{(Rb - q)' [Rs^2(X'X)^{-1} R]'^{-1} (Rb - q)}{J}$$

With normally distributed disturbances, the $F$ test is valid in any sample size. Consider general hypotheses of the form

$$H_0 : c(\beta) = 0 \quad H_1 : c(\beta) \neq 0$$

The **Wald statistic** for testing this hypothesis and its limiting distribution under $H_0$ would be

$$W = c(b)' \{ C(b)[\hat{\sigma}^2(X'X)^{-1}] C(b)' \}^{-1} c(b) \xrightarrow{d} \chi^2[J]$$

where

$$C(b) = [\partial c(\beta) / \partial b']$$

The **likelihood ratio (LR) test** is carried out by comparing the values of the log-likelihood function with and without the restrictions imposed. The test statistic and it’s limiting distribution under $H_0$ are

$$LR = -2[\ln L - \ln L] \xrightarrow{d} \chi^2[J]$$

The **Lagrange multiplier (LM) test** is based on the gradient of the log-likelihood function. The log-likelihood function can be maximized subject to a set of restrictions by using

$$\ln L_{LM} = -\frac{n}{2} \left[ \ln 2\pi + \ln \sigma^2 + \left[ (y - X\beta)'(y - X\beta)] / n \right] + \lambda c(\beta) \right]$$

The first-order conditions for a solution are

$$\begin{bmatrix}
\frac{\partial \ln L_{LM}}{\partial \beta} \\
\frac{\partial \ln L_{LM}}{\partial \sigma^2} \\
\frac{\partial \ln L_{LM}}{\partial \lambda}
\end{bmatrix} = \begin{bmatrix}
\frac{X'(y - X\beta)}{\sigma^2} + C(\beta)' \lambda \\
-\frac{n - \sigma^2}{\sigma^2} + \frac{(y - X\beta)'(y - X\beta)}{2\sigma^4} \\
\frac{c(\beta)}{\sigma^2}
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}$$

The solutions to these equations give the restricted least squares estimator, $b_*$; the usual variance estimator, now $e'_* e / n$; and the Lagrange multipliers. For this model, the solution for $\lambda_*$ is $\lambda_* = [R(X'X)^{-1} R]'^{-1} (Rb - q)$. If we carry out a Wald test of the hypothesis that $\lambda_*$ equals 0, then the statistic will be

$$LM = \lambda_* \{ \text{Est. Var}[\lambda_*] \}^{-1} \lambda_* = (Rb - q)' [Rs^2(X'X)^{-1} R'] (Rb - q)$$

An alternative way to compute the statistic is based on the general result that under the hypothesis being tested,

$$E[\partial \ln L / \partial \ln \beta] = E[(1/\sigma^2) X' e] = 0$$
and
\[ \text{Asy.Var}[\partial \ln L/\partial \beta] = -E[\partial^2 \ln L/\partial \beta \partial \beta'] = \sigma^2 (X'X)^{-1} \]

We can test the hypothesis that at the restricted estimator, the derivatives are equal to zero. The statistic would be
\[ \text{LM} = \frac{\epsilon'X(X'X)^{-1}X'e_*}{\epsilon'e_*/n} = nR^2_* \]

In this form, the LM statistic is \( n \) times the coefficient of determination in a regression of the residuals \( e_* = (y_i - x_i'b_*) \) on the full set of regressors. With some manipulation we can show that \( W = [n/(n - K)]JF \) and LR and LM are approximately equal to this function of \( F \). All three statistics converge to \( JF \) as \( n \) increases.

### 17.4.2 Maximum Likelihood Estimation of Nonlinear Regression Models

Suppose that, in general, the model is
\[ g(y_i, \theta) = h(x_i, \beta) + \epsilon_i \]

For normally distributed disturbances, the density of \( y_i \) is
\[ f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\frac{(g(y_i, \theta) - h(x_i, \beta))^2}{(2\sigma^2)}} \]

The Jacobian of the transformation is
\[ J(y_i, \theta) = \frac{\partial g(y_i, \theta)}{\partial y_i} = J_i \]

After collecting terms, the log-likelihood function will be
\[ \ln L = \sum_{i=1}^{n} -\frac{1}{2}[\ln 2\pi + \ln \sigma^2] + \sum_{i=1}^{n} \ln J(y_i, \theta) - \frac{1}{2}\sum_{i=1}^{n} \frac{(g(y_i, \theta) - h(x_i, \beta))^2}{2\sigma^2} \]

The maximum likelihood estimator of \( \sigma^2 \) will be
\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(g(y_i, \hat{\theta}) - h(x_i, \hat{\beta}))^2}{2\sigma^2} \]

The likelihood equations for the unknown parameters are
\[ \begin{bmatrix} \frac{\partial \ln L}{\partial \sigma^2} \\
\frac{\partial \ln L}{\partial \theta} \\
\frac{\partial \ln L}{\partial \beta} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} \epsilon_i \frac{\partial h(x_i, \beta)}{\partial \beta} \\
\sum_{i=1}^{n} \frac{1}{\sigma^2} \epsilon_i \frac{\partial h(x_i, \beta)}{\partial \theta} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 \\
\sum_{i=1}^{n} \frac{1}{\sigma^2} \epsilon_i \frac{\partial g(y_i, \theta)}{\partial \theta} \end{bmatrix} = \begin{bmatrix} 0 \\
0 \\
0 \end{bmatrix} \]

These equations will usually be nonlinear, so a solution must be obtained iteratively. We obtain the concentrated log-likelihood,
\[ \ln L_c = \sum_{i=1}^{n} \ln J(y_i, \theta) - \frac{n}{2}[1 + \ln(2\pi)] - \frac{n}{2} \ln \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \right] \]
This equation is a function only of $\theta$ and $\beta$. We can maximize it with respect to $\theta$ and $\beta$ and obtain the estimate of $\sigma^2$ as a by-product.

An estimate of the asymptotic covariance matrix of the maximum likelihood estimators can be obtained by inverting the estimated information matrix. However, that the Berndt et al. (1974) estimator will be much easier to compute. The derivatives of $\ln L_i$ with respect to the unknown parameters are

$$g_i = \left[ \frac{\partial \ln L_i}{\partial \beta}, \frac{\partial \ln L_i}{\partial \sigma^2} \right] = \left[ \frac{(\epsilon_i/\sigma^2)[\partial h(x_i, \beta)]/\partial \beta}{(1/J_i)[\partial J_i/\partial \theta] - (\epsilon_i/\sigma^2)[\partial g(y_i, \theta)]/\partial \theta}, 1/(2\sigma^2)[\epsilon_i^2/\sigma^2 - 1] \right]$$

The asymptotic covariance matrix for the maximum likelihood estimators is estimated using

$$\text{Est. Asy. Var}[\text{MLE}] = \left( \sum_{i=1}^{n} \hat{g}_i \hat{g}_i' \right)^{-1} = (\hat{G}'\hat{G})^{-1}$$

### 17.4.3 Nonnormal Disturbances—The Stochastic Frontier Model

This final application will examine a regression-like model in which the disturbances do not have a normal distribution. The implication for an empirical regression model is that in a formulation such as $y = h(x, \beta) + u$, $u$ must be negative. Since the theoretical production function is an ideal—the frontier of efficient production—any nonzero disturbance must be interpreted as the result of inefficiency. An empirical frontier production model such as

$$\ln y = \beta_1 + \sum_k \beta_k \ln x_k - u, \quad u \geq 0$$

Production function could arise from two sources: (1) productive inefficiency as we have defined it above and that would necessarily be negative; and (2) idiosyncratic effects that are specific to the firm and that could enter the model with either sign.

$$\ln y = \beta_1 + \sum_k \beta_k \ln x_k - u + v, \quad u \geq 0, \quad v \sim N[0, \sigma_v^2]$$

The frontier for any particular firm is $h(x, \beta) + v$, hence the name stochastic frontier. Let $\epsilon = v - u$, $\lambda = \sigma_u/\sigma_v$, $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$, and $\Phi(z) = \text{probability to the left of } z$ in the standard normal distribution. The “half-normal” model,

$$\ln h(\epsilon_i|\beta, \lambda, \sigma) = \left[ -\ln \sigma - \left( \frac{1}{2} \right) \log \left( \frac{\epsilon_i}{\sigma} \right)^2 + \ln \Phi \left( \frac{-\epsilon_i \lambda}{\sigma} \right) \right]$$

whereas for the exponential model

$$\ln h(\epsilon_i|\theta, \sigma_v) = \left[ \ln \theta + \frac{1}{2} \theta^2 \sigma_v^2 + \theta \epsilon_i + \ln \Phi \left( \frac{-\epsilon_i}{\sigma_v} - \theta \sigma_v \right) \right]$$

Both these distributions are asymmetric.
We will work through maximum likelihood estimation of the half-normal model in detail to illustrate the technique. The log likelihood is

\[
\ln L = -n \ln \sigma - \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} \frac{(\epsilon_i)^2}{\sigma} + \sum_{i=1}^{n} \ln \Phi \left( -\frac{\epsilon_i \lambda}{\sigma} \right).
\]

If we let \( \theta = 1/\sigma \) and \( \gamma = (1/\sigma)\beta \), the log-likelihood function becomes

\[
\ln L = -n \ln \sigma - \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} \left( \theta y_i - \gamma' x_i \right)^2 + \sum_{i=1}^{n} \ln \Phi \left( -\lambda(\theta y_i - \gamma' x_i) \right)
\]

We will make repeated use of the functions

\[
\alpha_i = \frac{\epsilon_i}{\sigma} = \theta y_i - \gamma' x_i
\]

\[
\phi(y_i, x_i, \lambda, \theta, \gamma) = \frac{\phi[-\lambda \alpha_i]}{\Phi[-\lambda \alpha_i]} = \delta_i
\]

\[
\Delta_i = -\delta_i(-\lambda \alpha_i + \delta_i)
\]

If will also be convenient to define the \((K+1) \times 1\) columns vectors \( z_i = (x_i', -y_i')' \) and \( t_i = (0', 1/\theta)' \). The likelihood equations are

\[
\frac{\partial \ln L}{\partial (\gamma', \theta)} = \sum_{i=1}^{n} t_i + \sum_{i=1}^{n} \alpha_i z_i + \lambda \sum_{i=1}^{n} \delta_i z_i = 0,
\]

\[
\frac{\partial \ln L}{\partial \lambda} = -\sum_{i=1}^{n} \delta_i \alpha_i = 0
\]

and the second derivatives are

\[
H(\gamma, \theta, \lambda) = \sum_{i=1}^{n} \left\{ \left[ \frac{(\lambda^2 \Delta_i - 1) z_i z_i'}{(\delta_i - \lambda \alpha_i \Delta_i) z_i'} \right] - \left[ \begin{array}{cc}
0 & 0 \\
0 & \lambda^2 \Delta_i
\end{array} \right] \right\}
\]

The estimator of the asymptotic covariance matrix for the directly estimated parameters is

\[
\text{Est.Asy.Var}[\hat{\gamma}', \hat{\theta}, \hat{\lambda}]' = -H[\hat{\gamma}', \hat{\theta}, \hat{\lambda}]^{-1}
\]

Since these transformations are one to one, the MLEs of \( \sigma \) and \( \beta \) are \( 1/\hat{\theta} \) and \( \hat{\gamma}/\hat{\theta} \). To compute an asymptotic covariance matrix for these estimators we will use the delta method, which will use the derivative matrix.

\[
G = \begin{bmatrix}
\frac{\partial \hat{\gamma}}{\partial \gamma'} & \frac{\partial \hat{\gamma}}{\partial \theta} & \frac{\partial \hat{\gamma}}{\partial \lambda'} \\
\frac{\partial \hat{\theta}}{\partial \gamma'} & \frac{\partial \hat{\theta}}{\partial \theta} & \frac{\partial \hat{\theta}}{\partial \lambda'} \\
\frac{\partial \hat{\lambda}}{\partial \gamma'} & \frac{\partial \hat{\lambda}}{\partial \theta} & \frac{\partial \hat{\lambda}}{\partial \lambda'}
\end{bmatrix} = \begin{bmatrix}
(1/\hat{\theta})1 & -(1/\hat{\theta}^2)\hat{\gamma} & 0 \\
0' & -(1/\hat{\theta}^2) & 0 \\
0' & 0 & 1
\end{bmatrix}
\]

Then, for the recovered parameters, we

\[
\text{Est.Asy.Var}[\hat{\beta}', \hat{\sigma}, \hat{\lambda}]' = G \times \left\{ -H[\hat{\gamma}', \hat{\theta}, \hat{\lambda}] \right\}^{-1} \times G'
\]

164
### 17.4.4 Conditional Moment Tests of Specification

A spate of studies has shown how to use conditional moment restrictions for specification testing as well as estimation.

Consider a general case in which the moment condition is written in terms of variables in the model \([y_i, x_i, z_i]\) and parameters \(\hat{\theta}\). The sample moment can be written

\[
\bar{r} = \frac{1}{n} \sum_{i=1}^{n} r_i(y_i, x_i, z_i, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{r}_i
\]

The hypothesis is that based on the true \(\theta\), \(E[r_i] = 0\). Under the hypothesis that \(E[r_i] = 0\) and assuming that \(\text{plim} \hat{\theta} = \theta\) and that a central limit theorem applies to \(\sqrt{n\bar{r}(\theta)}\) so that

\[
\sqrt{n\bar{r}(\theta)} \xrightarrow{d} N[0, \Sigma]
\]

for some covariance matrix \(\Sigma\) that we have yet to estimate, it follows that the Wald statistic,

\[
n\bar{r}'\hat{\Sigma}^{-1}\bar{r} \xrightarrow{d} \chi^2(J)
\]

where the degrees of freedom \(J\) is the number of moment restrictions being tested and \(\hat{\Sigma}\) is an estimate of \(\Sigma\). Thus, the statistic can be referred to the chi-squared table. The full derivation of \(\Sigma\) is fairly complicated. But when the vector of parameter estimators is a maximum likelihood estimator, as it would be for the least squares estimator with normally distributed disturbances and for most of the other estimators we consider, a surprisingly simple estimator can be used. Suppose that the parameter vector used to compute the moments above is obtained by solving the equations

\[
\frac{1}{n} \sum_{i=1}^{n} g(y_i, x_i, z_i, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i = 0
\]

where \(\hat{\theta}\) is the estimated parameter vector. Let the matrix \(G\) be the \(n \times K\) matrix with \(i\)th row equal to \(\hat{g}_i'\). In a maximum likelihood problem, \(G\) is the matrix of derivatives of the individual terms in the log-likelihood function with respect to the parameters. Let \(R\) be the \(n \times J\) matrix whose \(i\)th row is \(\bar{r}_i'\). Pagan and Vella show that for maximum likelihood estimators, \(\Sigma\) can be estimated using

\[
S = \frac{1}{n} [R'R - R'G(G'G)^{-1}G'R]
\]

The operational version of the statistic is

\[
C = n\bar{r}'S^{-1}\bar{r} = \frac{1}{n} \bar{r}'[R'R - R'G(G'G)^{-1}G'R]^{-1}R'
\]

where \(i\) is an \(n \times 1\) column of ones, which, once again, is referred to the appropriate critical value in the chi-squared table.

For testing one of the \(L\) conditions, say the \(l\)th one, the test can be carried out by a simple \(t\) test of whether the constant term is zero in a linear regression of the \(l\)th column of \(R\) on a constant term and all the columns of \(G\).
17.5 Two-Step Maximum Likelihood Estimation

Consider an example in which we have the following:

Model 1. Expected number of children = \( E[y_1|x_1, \theta_1] \)

Model 2. Decision to enroll in job training = \( y_2 \), a function of \( (x_2, \theta_2, E[y_1|x_1, \theta_1]) \)

There are two parameter vectors, \( \theta_1 \) and \( \theta_2 \). The first appears in the second model, although not the reverse. In such a situation, there are two ways to proceed. Full information maximum likelihood (FIML) estimation would involve forming the joint distribution \( f(y_1, y_2|x_1, x_2, \theta_1, \theta_2) \) of the two random variables and then maximizing the full log-likelihood function,

\[
\ln L = \sum_{i=1}^{n} f(y_{i1}, y_{i2}|x_{i1}, x_{i2}, \theta_1, \theta_2)
\]

A second, or two-step, limited information maximum likelihood (LIML) procedure for this kind of model could be done by estimating the parameters of model 1, since it does not involve \( \theta_2 \), and then maximizing a conditional log-likelihood function using the estimates from Step 1:

\[
\ln \hat{L} = \sum_{i=1}^{n} f(y_{i2}|x_{i2}, \theta_2, (x_{i1}, \hat{\theta}_1))
\]

Suppose, then, at our model consists of the two marginal distributions, \( f_1(y_1|x_1, \theta_1) \) and \( f_2(y_2|x_1, x_2, \theta_1, \theta_2) \). Estimation proceeds in two steps.

1. Estimate \( \theta_1 \) by maximum likelihood in Model 1. Let \( (1/n)\hat{V}_1 \) be \( n \) times any of the estimators of the asymptotic covariance matrix of this estimator.

2. Estimate \( \theta_2 \) by maximum likelihood in Model 2, with \( \hat{\theta}_1 \) inserted in place of \( \theta_1 \) as if it were known. Let \( (1/n)\hat{V}_2 \) be \( n \) times any appropriate estimator of the asymptotic covariance matrix of \( \hat{\theta}_2 \).

The argument for consistency of \( \hat{\theta}_2 \) is essentially that if \( \theta_1 \) were known, then all our results for MLEs would apply for estimation of \( \theta_2 \), and since \( \text{plim} \hat{\theta} = \theta_1 \), asymptotically, this line of reasoning is correct.

**THEOREM** Asymptotic Distribution of the Two-Step MLE [Murphy and Topel (1985)]

If the standard regularity conditions are met for both log-likelihood functions, then the second-step maximum likelihood estimator of \( \theta_2 \) is consistent and asymptotically normally distributed with asymptotic covariance matrix

\[
V_2^* = \frac{1}{n} [V_2 + V_2[C\nu_{\nu'} - R\nu_{\nu'} - C\nu_{\nu'} R']V_2]
\]

where

\[
V_1 = \text{Asy.Var}[\sqrt{n}(\hat{\theta}_1 - \theta_1)] \text{ based on } \ln L_1,
\]

\[
V_2 = \text{Asy.Var}[\sqrt{n}(\hat{\theta}_2 - \theta_2)] \text{ based on } \ln L_2|\theta_1
\]

\[
C = E \left[ \frac{1}{n} \left( \frac{\partial \ln L_2}{\partial \theta_2} \right) \left( \frac{\partial \ln L_2}{\partial \theta_1'} \right) \right]
\]

\[
R = E \left[ \frac{1}{n} \left( \frac{\partial \ln L_2}{\partial \theta_2} \right) \left( \frac{\partial \ln L_1}{\partial \theta_1'} \right) \right]
\]
The correction of the asymptotic covariance matrix at the second step requires some additional computation. Matrices $V_1$ and $V_2$ are estimated by the respective uncorrected covariance matrices. Typically, the BHHH estimators,

$$\hat{V}_1 = \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial \ln f_{i1}}{\partial \theta_1} \right) \left( \frac{\partial \ln f_{i1}}{\partial \theta_1'} \right) \right]^{-1}$$

and

$$\hat{V}_2 = \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial \ln f_{i2}}{\partial \theta_2} \right) \left( \frac{\partial \ln f_{i2}}{\partial \theta_2'} \right) \right]^{-1}$$

are used. The matrices $R$ and $C$ are obtained by summing the individual observations on the cross products of the derivatives. These are estimated with

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial \ln f_{i2}}{\partial \theta_2} \right) \left( \frac{\partial \ln f_{i1}}{\partial \theta_1'} \right)$$

and

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial \ln f_{i2}}{\partial \theta_2} \right) \left( \frac{\partial \ln f_{i1}}{\partial \theta_1'} \right)$$

17.6 Maximum Simulated Likelihood Estimation

The density of $y_{it}$ when the parameter vector is $\beta_i$ is $f(y_{it}|x_{it}, \beta_i)$. The parameter vector $\beta_i$ is randomly distributed over individuals according to

$$\beta_i = \beta + \Delta z_i + v_i$$

where $\beta + \Delta z_i$ is the mean of the distribution, which depends on time invariant individual characteristics as well as parameters yet to be estimated, and the random variation comes from the individual heterogeneity, $v_i$. This random vector is assumed to have mean zero and covariance matrix, $\Sigma$. The conditional density of the parameters is denoted

$$g(\beta_i|z_i, \beta, \Delta, \Sigma) = g(v_i + \beta + \Delta z_i, \Sigma)$$

where $g(\cdot)$ is the underlying marginal density of the heterogeneity. For the $T$ observations in group $i$, the joint conditional density is

$$f(y_i|X_i, \beta_i) = \prod_{t=1}^{T} f(y_{it}|x_{it}, \beta_i)$$

The unconditional density for $y_i$ is obtained by integrating over $\beta_i$

$$f(y_i|X_i, z_i, \beta, \Delta, \Sigma) = E_{\beta_i}[f(y_i|X_i, \beta_i)] = \int_{\beta_i} f(y_i|X_i, \beta_i) g(\beta_i|z_i, \beta, \Delta, \Sigma) d\beta_i$$

Collecting terms, and making the transformation from $v_i$ to $\beta_i$, the true log-likelihood would be

$$\ln L = \sum_{i=1}^{n} \ln \left\{ \int_{v_i} \left[ \prod_{t=1}^{T} f(y_{it}|x_{it}, \beta + \Delta z_i + v_i) \right] g(v_i|\Sigma) dv_i \right\}$$

$$= \sum_{i=1}^{n} \ln \left\{ \int_{v_i} f(y_i|X_i, \beta + \Delta z_i + v_i) g(v_i|\Sigma) dv_i \right\}$$

167
Each of the $n$ terms involves an expectation over $v_i$. The end result of the integration is a function of $(\beta, \Delta, \Sigma)$ which is then maximized.

We now consider estimating the parameters of a model of this form, maximum simulated likelihood estimation. The terms in the log-likelihood are each of the form

$$\ln L_i = E_{v_i}[f(y_i|X_i, \beta + \Delta z_i + v_i)]$$

As noted, we do not have a closed form for this function, so we cannot compute it directly. Suppose we could sample randomly from the distribution of $v_i$. If an appropriate law of large numbers can be applied, then

$$\lim_{R \to \infty} \frac{1}{R} \sum_{r=1}^{R} f(y_i|X_i, \beta + \Delta z_i + V_{ir}) = E_r[f(y_i|X_i, \beta + \Delta z_i + V_i)]$$

where $V_{ir}$ is the $r$th random draw from the distribution. With sufficient random draws, the approximation can be made as close to the true function as desired. Write $\Sigma$ in the Cholesky form $\Sigma = LL'$ where $L$ is a lower triangular matrix. Now, let $u_{ir}$ be a vector of $k$ independent draws from the standard normal distribution. Then a draw from the multivariate distribution with covariance matrix $\Sigma$ is simply $v_{ir} = Lu_{ir}$. The simulated log-likelihood is

$$\ln L_S = \sum_{i=1}^{n} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \left[ \prod_{t=1}^{T} f(y_{it}|x_{it}, \hat{\beta}_{ir}) \right] \right\}$$

The resulting function is maximized with respect to $\beta, \Delta$ and $L$.

As before, we are interested in estimating person specific parameters. A prior estimate might simply use $\beta + \Delta z_i$, but this would not use all the information in the sample. A posterior estimate would compute

$$\hat{E}_{v_i}[\beta_{i}|\beta, \Delta, z_i, \Sigma] = \frac{\sum_{r=1}^{R} \hat{\beta}_{ir} f(y_i|X_i, \hat{\beta}_{ir})}{\sum_{r=1}^{R} f(y_i|X_i, \hat{\beta}_{ir})}, \quad \hat{\beta}_{ir} = \hat{\beta} + \hat{\Delta} z_i + \hat{L} u_{ir}$$

### 17.7 Pseudo-Maximum Likelihood Estimation and Robust Asymptotic Covariance Matrices

Let $f(y_i|x_i, \beta)$ be the true probability density for a random variable $y_i$ given a set of co-variates $x_i$ and parameter vector $\beta$. The log-likelihood function is $(1/n) \log L(\beta|y, X) = (1/n) \sum_{i=1}^{n} \log f(y_i|x_i, \beta)$. The MLE, $\hat{\beta}_{ML}$, is the sample statistic that maximizes this function. MLE is obtained by solving the set of empirical moment equations

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(y_i|x_i, \hat{\beta}_{ML})}{\partial \hat{\beta}_{ML}} = \frac{1}{n} \sum_{i=1}^{n} d_i(\hat{\beta}_{ML}) = d(\hat{\beta}_{ML}) = 0$$

The population counterpart to the sample moment equation is

$$E \left[ \frac{1}{n} \frac{\partial \log L}{\partial \beta} \right] = E \left[ \frac{1}{n} \sum_{i=1}^{n} d_i(\beta) \right] = E[\hat{d}(\beta)] = 0$$

168
Using what we know about GMM estimators, if $E[\tilde{d}(\beta)] = 0$, then $\hat{\beta}_{ML}$ is consistent and asymptotically normally distributed, with asymptotic covariance matrix equal to

$$V_{ML} = [G(\beta)'G(\beta)]^{-1}G(\beta)'\{\text{Var}[\tilde{d}(\beta)]\}G(\beta)[G(\beta)'G(\beta)]^{-1}$$

where $G(\beta)$ is the derivative vector, $G(\beta)$ is $1/n$ times the expected Hessian of $\log L$; that is, $(1/n)\tilde{d}(\beta) = \tilde{H}(\beta)$. As we saw earlier, $\text{Var}[\partial \log L / \partial \beta] = -E[H(\beta)]$. Collecting all seven appearances of $(1/n)E[H(\beta)]$, we obtain the familiar result $V_{ML} = \{-E[H(\beta)]\}^{-1}$.

Suppose that we obtain an estimator by maximizing some other function, $M_n(y, X, \beta)$ that, although not the log-likelihood function, also attains its unique maximum at the true $\beta$ as $n \to \infty$. The log-likelihood for a linear regression model with normally distributed disturbances with different variances, $\sigma^2_{\omega_i}$, is

$$\hat{h}_n(y, X, \beta) = \frac{1}{n} \sum_{i=1}^{n} \left\{- \frac{1}{2} \left[ \log(2\pi \sigma^2_{\omega_i}) + \frac{(y_i - x_i'\beta)^2}{\sigma^2_{\omega_i}} \right] \right\}$$

Now consider the general case. Suppose that we estimate $\beta$ by maximizing a criterion function

$$M_n(y|X, \beta) = \frac{1}{n} \sum_{i=1}^{n} \log g(y_i | x_i, \beta)$$

Suppose as well that plim$_n M_n(y, X, \beta) = E[M_n(y, X, \beta)]$ and that as $n \to \infty$, $E[M_n(y, X, \beta)]$ attains its unique maximum at $\beta$. Then, by the argument we used above for the MLE, plim$_n \partial M_n(y, X, \beta) / \partial \beta = E[\partial M_n(y, X, \beta) / \partial \beta] = 0$. Let $\beta_E$ be the estimator that maximizes $M_n(y, X, \beta)$. Then the estimator is defined by

$$\frac{\partial M_n(y, X, \beta)}{\partial \beta_E} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log g(y_i | x_i, \beta_E)}{\partial \beta_E} = \bar{m}(\beta_E) = 0$$

Thus, $\beta_E$ is a GMM estimator. Using the notation of our earlier discussion, $G(\beta_E)$ is the symmetric Hessian of $E[M_n(y, X, \beta)]$, which we will denote $(1/n)E[H_M(\beta_E)] = \tilde{H}_M(\beta_M)$. Proceeding as we did above to obtain $V_{ML}$, we find that the appropriate asymptotic covariance matrix for the extremum estimator would be

$$V_E = [\tilde{H}_M(\beta_M)]^{-1} \left( \frac{1}{n} \Phi \right) [\tilde{H}_M(\beta_M)]^{-1}$$

where $\Phi = \text{Var}[\partial \log g(y_i | x_i, \beta) / \partial \beta]$, and, as before, the asymptotic distribution is normal.

The Hessian in $V_E$ can easily be estimated by using its empirical counterpart,

$$\text{Est.}[H_M(\beta_M)] = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log g(y_i | x_i, \beta_E)}{\partial \beta_E \partial \beta'_E}$$

But, $\Phi$ remains to be specified, and it is unlikely that we would know that function to use. We can consistently estimate $\Phi$ by using the sample variance of the first derivatives,

$$\hat{\Phi} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\partial \log g(y_i | x_i, \hat{\beta})}{\partial \beta} \right] \left[ \frac{\partial \log g(y_i | x_i, \hat{\beta})}{\partial \beta'} \right]$$
If we obtain $\hat{\beta}_E$ by maximizing a criterion function that satisfies the other assumptions, then the appropriate estimator of the asymptotic covariance matrix is

$$\text{Est.} V_E = \frac{1}{n} [H(\hat{\beta}_E)]^{-1} \hat{\Phi}(\hat{\beta}_E)[H(\hat{\beta}_E)]^{-1}$$

If $\hat{\beta}_E$ is the true MLE, then $V_E$ simplifies to $\{-H(\hat{\beta}_E)\}^{-1}$. In the current literature, this estimator has been called the “sandwich” estimator.
Chapter 18

The Generalized Method of Moments

18.1 Consistent Estimation: The Method of Moments

The basis of the method of moments is as follows: In random sampling, under generally benign assumptions, a sample statistic will converge in probability to some constant.

18.1.1 Random Sampling and Estimating the Parameters of Distributions

Consider independent, identically distributed random sampling from a distribution \( f(y|\theta_1, \ldots, \theta_K) \) with finite moments up to \( E[y^{2K}] \). The sample consists of \( n \) observations, \( y_1, \ldots, y_n \). The \( k \)th “raw” or uncentered moments is

\[
\bar{m}_k' = \frac{1}{n} \sum_{i=1}^{n} y_i^k,
\]

By Theorem,

\[
E[\bar{m}_k'] = \mu_k' = E[y_i^k]
\]

and

\[
\text{Var}[\bar{m}_k'] = \frac{1}{n} \text{Var}[y_i^k] = \frac{1}{n} (\mu_{2k} - \mu_k')
\]

By convention, \( \mu_1' = E[y_i] = \mu \). By the Khinchine Theorem,

\[
\text{plim} \bar{m}_k' = \mu_k' = E[y_i^k]
\]

Finally, by the Lindberg-Levy Central Limit Theorem,

\[
\sqrt{n}(\bar{m}_k' - \mu_k') \xrightarrow{d} N[0, \mu_{2k} - \mu_k']
\]

In general, \( \mu_k' \) will be a function of the underlying parameters. By computing \( K \) raw moments and equating them to these functions, we obtain \( K \) equations that can be solved to provide estimates of the \( K \) unknown parameters.

Let \( m_k(\cdot) \) be a continuous and differentiable function not involving the sample size \( n \), and let

\[
\bar{m}_k = \frac{1}{n} \sum_{i=1}^{n} m_k(y_i), \quad k = 1, 2, \ldots, K
\]
These are also “moments” of the data.

\[ \text{plim} \bar{m}_k = E[m_k(y_i)] = \mu_k(\theta_1, ..., \theta_K) \]

We assume that \( \mu_k(\cdot) \) involves some of or all the parameters of the distribution. With \( K \) parameters to be estimated, the \( K \) \textbf{moment equations},

\[
\begin{align*}
\bar{m}_1 - \mu_1(\theta_1, ..., \theta_K) &= 0, \\
\bar{m}_2 - \mu_2(\theta_1, ..., \theta_K) &= 0, \\
\vdots \\
\bar{m}_K - \mu_K(\theta_1, ..., \theta_K) &= 0,
\end{align*}
\]

provide \( K \) equations in \( K \) unknowns, \( \theta_1, ..., \theta_K \). If the equations are continuous and functional independent, then \textbf{method of moments estimators} can be obtained by solving the system of equations for

\[ \hat{\theta}_k = \hat{\theta}_k[\bar{m}_1, ..., \bar{m}_K] \]

In most cases, method of moments estimators are not efficient. The exception is in random sampling from \textbf{exponential families} of distributions.

**DEFINITION** \textbf{Exponential Family}

An exponential (parametric) family of distributions is one whose log-likelihood is of the form

\[ \ln L(\theta|\text{data}) = a(\text{data}) + b(\theta) + \sum_{k=1}^{K} c_k(\text{data}) s_k(\theta), \]

where \( a(\cdot), b(\cdot), c(\cdot) \) and \( s(\cdot) \) are functions. The members of the “family” are distinguished by the different parameter values.

If the log-likelihood function is of this form, then the functions \( c_k(\cdot) \) are called \textbf{sufficient statistics}. When sufficient statistics exist, method of moments estimators can be functions of them. In this case, the method of moments estimators will also be the maximum likelihood estimators, so, of course, they will be efficient, at least asymptotically.

18.1.2 \textbf{Asymptotic Properties of the Method of Moments Estimator}

In general, when the moments are computed as

\[ \bar{m}_k = \frac{1}{n} \sum_{i=1}^{n} m_k(y_i), \quad k = 1, ..., K \]

where \( y_i \) is an observation on a vector of variables, an appropriate estimator of the asymptotic covariance matrix of \( [\bar{m}_1, ..., \bar{m}_K] \) can be computed using

\[
\frac{1}{n} F_{jk} = \frac{1}{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} [(m_j(y_i) - \bar{m}_j)(m_k(y_i) - \bar{m}_k)] \right\}, \quad j, k = 1, ..., K
\]

This estimator provides the asymptotic covariance matrix for the moments used in computing the estimated parameters. Under our assumption of iid random sampling from a
distribution with finite moments up to $2K$, $F$ will converge in probability to the appropriate covariance matrix of the normalized vector of moments, $\Phi = \text{Asy.Var}[\sqrt{n} \bar{m}_n(\theta)]$.

To formalized the remainder of this derivation, refer back to the moment equations, which we will now write

$$\bar{m}_{n,k}(\theta_1, \theta_2, ..., \theta_K) = 0, \quad k = 1, ..., K$$

The subscript $n$ indicates the dependence on a data set of $n$ observations. Let $\bar{G}_n(\theta)$ be the $K \times K$ matrix whose $k$th row is the vector of partial derivatives

$$\bar{G}_{n,k} = \frac{\partial \bar{m}_{n,k}}{\partial \theta}$$

Now, expand the set of solved moment equations around the true values of the parameters $\theta_0$ in a linear Taylor series. The linear approximation is

$$0 \approx [\bar{m}_n(\theta_0)] + \bar{G}_n(\theta_0)(\hat{\theta} - \theta_0)$$

Therefore

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -[\bar{G}_n(\theta_0)]^{-1}\sqrt{n}[\bar{m}_n(\theta_0)]$$

We have from (the Central Limit Theorem) that $\sqrt{n} \bar{m}_n(\theta_0)$ has a limit normal distribution with mean vector 0 and covariance matrix equal to $\Phi$. Assuming that the functions in the moment equation are continuous and functionally independent, we can expect $\bar{G}_n(\theta_0)$ to converge to a nonsingular matrix of constants, $\Gamma(\theta_0)$. Jumping to the conclusion, we expect the asymptotic distribution of $\hat{\theta}$ to be normal with mean vector $\theta_0$ and covariance matrix $(1/n)\{ -\Phi[\Gamma'(\theta_0)]^{-1}\} \Phi \{ -\Phi[\Gamma'(\theta_0)]^{-1}\}$. Thus, the asymptotic covariance matrix for the method of moments estimator may be estimated with

$$\text{Est. Asy.Var}[\hat{\theta}] = \frac{1}{n}[\bar{G}_n'(\hat{\theta})F^{-1}\bar{G}_n(\hat{\theta})]^{-1}$$

### 18.2 The Generalized Method of Moments (GMM) Estimator

#### 18.2.1 Estimation Based on Orthogonality Conditions

The model specified for the random variable $y_i$ implies certain expectations, for example

$$E[y_i] = \mu$$

where $\mu$ is the mean of the distribution of $y_i$. Estimation of $\mu$ then proceeds by forming a sample analog to the population expectation:

$$E[y_i - \mu] = 0$$

The sample counterpart to this expectation is the **empirical moment equation**,

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{\mu}) = 0$$

The estimator is the value of $\tilde{\mu}$ that satisfies the sample moment equation.
Now, consider the apparently different case of the least squares estimator of the parameters in the classical linear regression model. An important assumption of the model is

\[ E[x_i \epsilon_i] = E[x_i(y_i - x'_i \beta)] = 0 \]

The sample analog is

\[ \frac{1}{n} \sum_{i=1}^{n} x_i \hat{\epsilon}_i = \frac{1}{n} \sum_{i=1}^{n} x_i (y_i - x'_i \hat{\beta}) = 0 \]

The estimator of \( \beta \) is the one that satisfies these moment equations, which are just the normal equations for the least squares estimator. So, we see that the OLS estimator is a method of moments estimator.

For the instrumental variables estimator, we relied on a large sample analog to the moment condition

\[ \text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} z_i \epsilon_i \right) = \text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} z_i (y_i - x'_i \hat{\beta}) \right) = 0 \]

We resolved the problem of having more instruments than parameters by solving the equations

\[ \left( \frac{1}{n} X'Z \right) \left( \frac{1}{n} Z'Z \right)^{-1} \left( \frac{1}{n} Z'\hat{\epsilon} \right) \]

where the columns of \( \hat{X} \) are the fitted values in regression on all the columns of \( Z \).

The nonlinear least squares estimator was defined similarly, though in this case, the normal equations are more complicated since the estimator is only implicit. The population orthogonality condition for the nonlinear regression model is \( E[x^0_i \epsilon_i] = 0 \). The empirical moment equation is

\[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial E[y_i|x_i, \beta]}{\partial \beta} \right) (y_i - E[y_i|x_i, \beta]) = 0 \]

All the maximum likelihood estimators that we have looked at thus far and will encounter later are obtained by equating the derivatives of a log-likelihood to zero. The scaled log-likelihood function is

\[ \frac{1}{n} \ln L = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\theta, x_i) \]

where \( f(\cdot) \) is the density function and \( \theta \) is the parameter vector. For densities that satisfy the regularity conditions,

\[ E \left[ \frac{\partial \ln f(y_i|\theta, x_i)}{\partial \theta} \right] = 0, \]

The maximum likelihood estimator is obtained by equating the sample analog to zero:

\[ \frac{1}{n} \frac{\partial \ln L}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|x_i, \hat{\theta})}{\partial \theta} = 0 \]

The upshot is that nearly all the estimators we have discussed and will encounter later can be construed as method of moments estimators.
18.2.2 Generalizing the Method of Moments

Suppose that the model involves $K$ parameters, $\theta = (\theta_1, \theta_2, ..., \theta_K)$, and that the theory provides a set of $L > K$ moment conditions,

$$E[m_l(y_i, x_i, z_i, \theta)] = E[m_l(\theta)] = 0$$

where $y_i, x_i,$ and $z_i$ are variables that appear in the model and the subscript $i$ on $m_l(\theta)$ indicates the dependence on $(y_i, x_i, z_i)$. Denote the corresponding sample means as

$$\bar{m}_l(y, X, Z, \theta) = \frac{1}{n} \sum_{i=1}^{n} m_l(y_i, x_i, z_i, \theta)$$

Unless the equations are functionally dependent, the system of $L$ equations in $K$ unknown parameters,

$$\bar{m}_l(\theta) \frac{1}{n} \sum_{i=1}^{n} m_l(y_i, x_i, z_i, \theta) = 0, \quad l = 1, ..., L$$

will not have a unique solution. One possibility is to minimize a criterion function, such as the sum of squares,

$$q = \sum_{i=1}^{L} \bar{m}_l^2 = \bar{m}(\theta)'\bar{m}(\theta)$$

It can be shown that under the assumptions we have made so far, specifically that \( \text{plim} \bar{m}(\theta) = 0 \) where $W_n$ is any positive definite matrix that may depend on the data but is not a function of $\theta$, such as $I$, to produce a consistent estimator of $\theta$. We can, in fact, use as the criterion a weighted sum of squares,

$$q = \bar{m}(\theta)'W_n\bar{m}(\theta)$$

Let $W$ be a diagonal matrix whose diagonal elements are the reciprocals of the variances of the individual moments,

$$w_{ll} = \frac{1}{\text{Asy.Var}[\sqrt{n}\bar{m}_l]} = \frac{1}{\phi_{ll}}$$

Then, a weighted least squares procedure would minimize

$$q = \bar{m}(\theta)'\Phi^{-1}\bar{m}(\theta)$$

In general, the $L$ elements of $\bar{m}$ are freely correlated. To use generalized least squares, we would define the full matrix,

$$W = \left\{ \text{Asy.Var}[\sqrt{n}\bar{m}] \right\}^{-1} = \Phi^{-1}$$

The estimators defined by choosing $\theta$ to minimize

$$q = \bar{m}(\theta)'W_n\bar{m}(\theta)$$

are minimum distance estimators. The general result is that if $W_n$ is a positive definite matrix and if

$$\text{plim}\bar{m}(\theta) = 0,$$
then the minimum distance estimator of $\theta$ is consistent.

The asymptotic covariance matrix of this generalized method of moments estimator is
\[ V_{GMM} = \frac{1}{n} [\Gamma' W \Gamma]^{-1} = \frac{1}{n} [\Gamma' \Phi^{-1} \Gamma]^{-1} \]
where $\Gamma$ is the matrix of derivatives with $j$th row equal to
\[ \Gamma^j = \text{plim} \frac{\partial \bar{m}_j(\theta)}{\partial \theta} \]
and $\Phi = \text{Asy.Var}[\sqrt{n} \bar{m}]$. Finally, by virtue of the central limit theorem applied to the sample moments and the Slutsky theorem applied to this manipulation, we can expect the estimator to be asymptotically normally distributed.

18.2.3 Properties of the GMM Estimator

The GMM estimator is based on the set of population orthogonality conditions,
\[ E[m_i(\theta_0)] = 0 \]
where we denote the true parameter vector by $\theta_0$. The subscript $i$ on the term on the right hand side indicates dependence on the observed data, $y_i, x_i, z_i$. Averaging this over the sample observations produces the sample moment equation
\[ E[\bar{m}_n(\theta_0)] = 0 \]
where
\[ \bar{m}_n(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} m_i(\theta_0) \]
This moment is a set of $L$ equations involving the $K$ parameters.

We make the following assumptions about the model and these empirical moments:

**ASSUMPTION 18.1 Convergence of the Empirical Moments:** The data generating process is assumed to meet the conditions for a law of large numbers to apply, so that we may assume that the empirical moments converge in probability to their expectation. What is required for this assumption is that
\[ \bar{m}_n(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} m_i(\theta_0) \xrightarrow{P} 0 \]
For this more general case, the, we would assume that the sequence of observations $m(\theta)$ constant a jointly $(L \times 1)$ stationary and ergodic process.

The empirical moments are assumed to be continuous and continuously differentiable functions of the parameters. For our example above, this would mean that the conditional mean function, $h(x_i, \beta)$ is a continuous function of $\beta$.

With continuity and differentiability, we also will be able to assume that the derivatives of the moments
\[ \bar{G}_n(\theta_0) = \frac{\partial \bar{m}_n(\theta_0)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial m_i(\theta_0)}{\partial \theta} \]
converge to a probability limit, say $\text{plim} \bar{G}_n(\theta_0) = \bar{G}(\theta_0)$. 

176
ASSUMPTION 18.2 Identification: For any \( n \geq K \), if \( \theta_1 \) and \( \theta_2 \) are two different parameter vectors, then there exist data sets such that \( \bar{m}_n(\theta_1) \neq \bar{m}_n(\theta_2) \). Formally, identification is defined to imply that the probability limit of the GMM criterion function is uniquely minimized at the true parameters, \( \theta_0 \).

The identification has three important implications:

1. Order Condition. The number of moment conditions is at least as large as the number of parameter; \( L \geq K \). This is necessary but not sufficient for identification.

2. Rank Condition. The \( L \times K \) matrix of derivatives, \( \bar{G}_n(\theta_0) \) will have row rank equal to \( K \).

3. Uniqueness. With the continuity assumption, the identification assumption implies that the parameter vector that satisfies the population moment condition is unique. We know that at the true parameter vector, \( \text{plim} \; \bar{m}_n(\theta_0) = 0 \). If \( \theta_1 \) is any parameter vector that satisfies this condition, then \( \theta_1 \) must equal \( \theta_0 \).

ASSUMPTION 18.3 Asymptotic Distribution of Empirical Moments:

We assume that the empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix, \( (1/n)\Phi \), so that

\[
\sqrt{n} \bar{m}_n(\theta_0) \xrightarrow{d} N[0, \Phi]
\]

We have assumed that

\[ E[m_i(\theta_0)] = 0 \]

If we can go a step further and assume that the functions \( m_i(\theta_0) \) are an ergodic, stationary martingale difference series,

\[ E[m_i(\theta_0)|m_{i-1}(\theta_0), m_{i-2}(\theta_0), ...] = 0 \]

For regression models, the assumption takes the form

\[ E[z_i\epsilon_i|z_{i-1}\epsilon_{i-1}, ...] = 0 \]

which will often be part of the central structure of the model. With the assumptions in place, we have

THEOREM Asymptotic Distribution of the GMM Estimator

Under the preceding assumptions,

\[
\hat{\theta}_{GMM} \xrightarrow{P} \theta \\
\hat{\theta}_{GMM} \xrightarrow{a} N[\theta, V_{GMM}]
\]

where \( V_{GMM} \) is defined

\[ V_{GMM} = \frac{1}{n} [\Gamma'W\Gamma]^{-1} = \frac{1}{n} [\Gamma'\Phi^{-1}\Gamma]^{-1} \]
The GMM estimator is obtained by minimizing the criterion function
\[ q_n(\theta) = \bar{m}_n(\theta)' W_n \bar{m}_n(\theta) \]
where \( W_n \) is the weighting matrix used. By our assumptions of strict continuity and Assumption 18.1, \( q_n(\theta_0) \) converges to 0. Since \( W_n \) is positive definite, for any finite \( n \), we know that
\[ 0 \leq q_n(\tilde{\theta}_{GMM}) \leq q_n(\theta_0) \]
That is, in the finite sample, \( \tilde{\theta}_{GMM} \) actually minimizes the function, so the sample value of the criterion is not larger at \( \tilde{\theta}_{GMM} \) than at any other value, including the true parameters. But, at the true parameter values, \( q_n(\theta_0) \xrightarrow{p} 0 \). So, \( q_n(\tilde{\theta}_{GMM}) \xrightarrow{p} 0 \) as well because of the identification assumption, 18.2. As \( n \to \infty, q_n(\tilde{\theta}_{GMM}) \) and \( q_n(\theta) \) converge to the same limit. It must be the case, then, that as \( n \to \infty, \bar{m}_n(\tilde{\theta}_{GMM}) \to \bar{m}_n(\theta_0) \), since the function is quadratic and \( W \) is positive definite. The identification condition that we assumed earlier now assures that as \( n \to \infty, \tilde{\theta}_{GMM} \) must equal \( \theta_0 \). This establishes consistency of the estimator.

The first order conditions for the GMM estimator are
\[ \frac{\partial q_n(\tilde{\theta}_{GMM})}{\partial \tilde{\theta}_{GMM}} = 2G_n(\tilde{\theta}_{GMM})' W_n \bar{m}_n(\tilde{\theta}_{GMM}) = 0 \]
We expand the empirical moments in a linear Taylor series around the true value. \( \theta \);
\[ \bar{m}_n(\tilde{\theta}_{GMM}) = \bar{m}_n(\theta_0) + \tilde{G}_n(\bar{\theta})(\tilde{\theta}_{GMM} - \theta_0) \]
where \( \bar{\theta} \) is a point between \( \tilde{\theta}_{GMM} \) and the true parameters, \( \theta_0 \). Thus, for each element \( \bar{\theta}_k = w_k \bar{\theta}_{GMM} + (1 - w_k)\theta_0 \) for some \( w_k \) such that \( 0 < w_k < 1 \).
\[ G_n(\tilde{\theta}_{GMM})' W_n \bar{m}_n(\theta_0) + G_n(\tilde{\theta}_{GMM})' W_n \tilde{G}_n(\bar{\theta})(\tilde{\theta}_{GMM} - \theta_0) = 0 \]
Solve this equation for the estimation error and multiply by \( \sqrt{n} \). This produces
\[ \sqrt{n}(\tilde{\theta}_{GMM} - \theta_0) = -[G_n(\tilde{\theta}_{GMM})' W_n \tilde{G}_n(\bar{\theta})]^{-1} G_n(\tilde{\theta}_{GMM})' W_n \sqrt{n} \bar{m}_n(\theta_0) \]
By the consistency of \( \tilde{\theta}_{GMM} \) we know that \( \tilde{\theta}_{GMM} \) and \( \bar{\theta} \) both converge to \( \theta_0 \). By the strict continuity assumed, it must also be the case that
\[ G_n(\bar{\theta}) \xrightarrow{p} \bar{G}_n(\theta_0) \quad \text{and} \quad G_n(\tilde{\theta}_{GMM}) \xrightarrow{p} \tilde{G}_n(\theta_0) \]
We have also assumed that the weighting matrix, \( W_n \), converges to a matrix of constants, \( W \). Collecting terms,
\[ \sqrt{n}(\tilde{\theta}_{GMM} - \theta_0) \xrightarrow{p} \{ [\bar{G}(\theta_0)' W \tilde{G}(\theta_0)]^{-1} \bar{G}(\theta_0)' W \} \sqrt{n} \bar{m}_n(\theta_0) \]
We now invoke Assumption 18.3. Collecting terms, we have the result
\[ V_{GMM} = \frac{1}{n} [\Gamma' W \Gamma]^{-1} = \frac{1}{n} [\Gamma' \Phi^{-1} \Gamma]^{-1} \]
If the optimal weighting matrix, \( W = \Phi^{-1} \), is used, then the expression collapses to
\[ V_{GMM,\text{optimal}} = \frac{1}{n} [\bar{G}(\theta_0)' \Phi^{-1} \bar{G}(\theta_0)]^{-1} \]
If we use least squares or instrumental variables with \( W = I \), then
\[ V_{GMM} = \frac{1}{n} (\bar{G}' \bar{G})^{-1} \bar{G}' \Phi \bar{G}(\theta_0)' \bar{G} \]
This equation is essentially the White or **Newey-West estimator**, which returns us to our departure point and provides a neat symmetry to the GMM principle.
18.2.4 GMM Estimation of Some Specific Econometric Models

Suppose that the theory specifies a relationship

\[ y_i = h(x_i, \beta) + \epsilon_i \]

where \( \beta \) is a \( K \times 1 \) parameter vector that we wish to estimate. This may not be a regression relationship, since it is possible that

\[ \text{Cov}[\epsilon_i, h(x_i, \beta)] \neq 0, \]

or even

\[ \text{Cov}[\epsilon_i, x_j] \neq 0, \quad \text{for all } i \text{ and } j \]

For the present, we assume that

\[ E[\epsilon | X] \neq 0 \]

and

\[ E[\epsilon \epsilon' | X] = \sigma^2 \Omega = \Sigma \]

where \( \Sigma \) is symmetric and positive definite but otherwise unrestricted. Suppose that at each observation \( i \) we observe a vector of \( L \) variables, \( z_i \), such that \( z_i \) is uncorrelated with \( \epsilon_i \). You will recognize \( z_i \) as a set of instrumental variables. The assumptions thus far have implied a set of orthogonality conditions.

\[ E[z_i \epsilon_i | x_i] = 0 \]

which may be sufficient to identify (if \( L = K \)) or even overidentify (if \( L > K \)) the parameters of the model.

For convenience, define

\[ e(X, \hat{\beta}) = y_i - h(x_i, \hat{\beta}), \quad i = 1, \ldots, n \]

and

\[ Z = n \times L \text{ matrix whose ith row is } z_i' \]

The sample moments will be

\[ m_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} z_i e(x_i, \beta) = \frac{1}{n} Z' e(X, \beta) \]

The minimum distance estimator will be the \( \hat{\beta} \) that minimizes

\[ q = m_n(\hat{\beta})' W m_n(\hat{\beta}) = \left( \frac{1}{n} e(X, \hat{\beta})' Z \right) W \left( \frac{1}{n} [Z' e(X, \hat{\beta})] \right) \]

for some choice of \( W \) that we have yet to determine. The criterion given above produces the nonlinear instrumental variable estimator. If we use \( W = (Z' Z)^{-1} \), then we have exactly the estimation criterion we used where we defined the nonlinear instrumental variables estimator. Hansen (1982) showed that the optimal choice of \( W \) for this estimator is

\[ W_{GMM} = \{ \text{Asy.Var}[\sqrt{n} m_n(\beta)] \}^{-1} \]

\[ = \{ \text{Asy.Var} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \epsilon_i \right] \}^{-1} = \{ \text{Asy.Var}[\sqrt{n} m_n(\beta)] \}^{-1} \]

\[ = \{ \text{Asy.Var} \left[ \frac{1}{\sqrt{n}} Z' e(X, \beta) \right] \}^{-1} \]

179
For our model, this is
\[
W = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[z_i \epsilon_i, z_j \epsilon_j] = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} z_i z_j' = \frac{Z' \Sigma Z}{n}
\]
we obtain the criterion for the GMM estimator:
\[
q = \left[ \frac{1}{n} e(X, \hat{\beta})' Z \right] \left( \frac{Z' \Sigma Z}{n} \right)^{-1} \left[ \frac{1}{n} Z' e(X, \hat{\beta}) \right]
\]
Estimation appears to require that an estimate of \( \beta \) be in hand already, even though it is the object of estimation. Suppose for the present that one is and what is required is
\[
\frac{1}{n} Z' \Sigma Z = \frac{1}{n} \sum_{i=1}^{n} z_i z_i' \text{Var}[y_i - h(x_i, \beta)]
\]
We can use the White (1980) estimator for this case:
\[
S_0 = \frac{1}{n} \sum_{i=1}^{n} z_i z_i'(y_i - h(x_i, \hat{\beta}))^2
\]
If the disturbances are autocorrelated but the process is stationary, then Newey and West’s (1987a) estimator is available:
\[
S = \left[ S_0 + \frac{1}{n} \sum_{l=1}^{p} w(l) \sum_{i=l+1}^{n} e_i e_{i-l} (z_i z_i'_{i-l} + z_{i-l} z_i') \right] = \sum_{l=0}^{p} w(l) S_l
\]
where
\[
w(l) = 1 - \frac{l}{p + 1}
\]
The maximum lag length \( p \) must be determined in advance. We will require that observations that are far apart in time—that is, for which \(|i - l| \) is large—must have increasingly smaller covariances for us to establish the convergence results that justify OLS, GLS and now GMM estimation. The choice of \( p \) is a reflection of how far back in time one must go to consider the autocorrelation negligible for purposes of estimating \((1/n)Z' \Sigma Z\). Current practice suggests using the smallest integer greater than or equal to \( T^{1/4} \).

One possibility is to obtain an inefficient but consistent GMM estimator by using \( W = I \). The first-step estimator can then be used to construct \( W \), which in turn, can be used in the GMM estimator. Another possibility is that \( \beta \) may be consistently estimable by some straightforward procedure other than GMM.

Once the GMM estimator has been computed, its asymptotic covariance matrix and asymptotic distribution can be estimated. Recall that
\[
\bar{m}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} z_i \epsilon_i
\]
which is a sum of \( L \times 1 \) vectors. The derivative, \( \partial \bar{m}_n(\beta)/\partial \beta' \), is a sum of \( L \times K \) matrices, so
\[
\bar{G}(\beta) = \partial \bar{m}(\beta)/\partial \beta' = \frac{1}{n} \sum_{i=1}^{n} G_i(\beta) = \frac{1}{n} \sum_{i=1}^{n} z_i \left[ \frac{\partial \epsilon_i}{\partial \beta'} \right]
\]
In the model we are considering here,
\[ \frac{\partial \epsilon_i}{\partial \beta'} = -\frac{\partial h(x_i, \beta)}{\partial \beta'} \]

Using the notation defined there,
\[ \frac{\partial \epsilon_i}{\partial \beta} = -x_{i0} \]
so
\[ \bar{G}(\beta) = \frac{1}{n} \sum_{i=1}^{n} G_i(\beta) = \frac{1}{n} \sum_{i=1}^{n} -z_i'x_{i0} = - \frac{1}{n} Z'X_0 \]

With this matrix in hand, the estimated asymptotic covariance matrix for the GMM estimator is
\[ \text{Est. Asy. Var}[\hat{\beta}] = \begin{bmatrix} G(\hat{\beta})' \left( \frac{1}{n} Z' \hat{\Sigma} Z \right)^{-1} G(\hat{\beta}) \end{bmatrix}^{-1} = \left[ (X_0'Z)(Z'\hat{\Sigma}Z)^{-1}(Z'X_0) \right]^{-1} \]

### 18.3 Testing Hypotheses in the GMM Framework

#### 18.3.1 Testing the Validity of the Moment Restrictions

In the exactly identified cases, the criterion for GMM estimation
\[ q = \bar{m}(\theta)W\bar{m}(\theta) \]
would be exactly zero because we can find a set of estimates for which \( \bar{m}(\theta) \) is exactly zero. Thus in the exactly identified case when there are the same number of moment equations as there are parameters to estimate, the weighting matrix \( W \) is irrelevant to the solution. But if the parameters are overidentified by the moment equations, then these equations imply substantive restrictions. By construction, when the optimal weighting matrix is used,
\[ nq = \left\{ \text{Est. Asy. Var}[\sqrt{n}\bar{m}(\hat{\theta})] \right\}^{-1} \left[ \sqrt{n}\bar{m}(\hat{\theta})' \right] \]
so \( nq \) is a Wald statistic. Therefore, under the hypothesis of the model,
\[ nq \overset{d}{\rightarrow} \chi^2[L - K] \]

Suppose \( \theta \) is subjected to \( J \) restrictions which restrict the number of free parameters from \( K \) to \( K - J \). The constrained problem may be stated in terms of
\[ q_R = \bar{m}(\theta_R)'W\bar{m}(\theta_R) \]
It follows then that
\[ nq_R \overset{d}{\rightarrow} \chi^2[L - (K - J)] \]
The weighted sum of squares with the restrictions imposed, \( nq_R \) must be larger than the weighted sum of squares obtained without the restrictions, \( nq \). The difference is
\[ (nq_R - nq) \overset{d}{\rightarrow} \chi^2[J] \]
The test is attributed to Newey and West (1987b).
18.3.2 GMM Counterparts to the Wald, LM and LR Tests

Newey and West (1987b) have devised counterparts to these test statistics for the GMM estimator. The Wald statistic is computed identically, using the results of GMM estimation rather than maximum likelihood. That is we would use the unrestricted GMM estimator of $\theta$. The appropriate asymptotic covariance matrix is

$$V_{GMM,\text{optimal}} = \frac{1}{n}[\bar{G}(\theta_0)'\Phi^{-1}\bar{G}(\theta_0)]^{-1}$$

The computation is exactly the same. The counter part to the LR statistic is the difference in the values of $nq$ in

$$(nq_R - nq) \xrightarrow{d} \chi^2[J]$$

It is necessary to use the same weighting matrix, $W$, in both restricted and unrestricted estimators. Since the unrestricted estimator is consistent under both $H_0$ and $H_1$, a consistent, unrestricted estimator of $\theta$ is used to computed $W$. Label this $\Phi^{-1}_1 = \{\text{Asy.Var}[\sqrt{n}\bar{m}_1(c_0)]\}$. In each occurrence, the subscript 1 indicates reference to the unrestricted estimator. Then $q$ is minimized without restrictions to obtain $q_1$ and then subject to the restrictions to obtain $q_0$. Then statistic is then $(nq_0 - nq_1)$.

Finally, the counterpart to the LM statistic would be

$$LM_{GMM} = n[\bar{m}_1(c_0)'\hat{\Phi}_1^{-1}\bar{G}_1(c_0)] [\bar{G}_1(c_0)'\hat{\Phi}_1^{-1}\bar{G}_1(c_0)] [\bar{G}_1(c_0)'\hat{\Phi}_1^{-1}\bar{m}_1(c_0)]$$

The logic for this LM statistic is the same as that for the MLE. The derivatives of the minimized criterion $q$ are

$$g_1(c_0) = \frac{\partial q}{\partial c_0} = 2\bar{G}_1(c_0)'\hat{\Phi}_1^{-1}\bar{m}(c_0)$$

The LM statistic, $LM_{GMM}$, is a Wald statistic for testing the hypothesis that this vector equals zero under the restrictions of the null hypothesis. From our earlier results, we would have

$$\text{Est.Asy.Var}[g_1(c_0)] = \frac{4}{n}\bar{G}_1(c_0)'\hat{\Phi}_1^{-1} \{\text{Est.Asy.Var}[\sqrt{n}\bar{m}(c_0)]\} \bar{G}_1(c_0)$$

The estimated asymptotic variance of $\sqrt{n}\bar{m}(c_0)$ is $\hat{\Phi}_1$, so

$$\text{Est.Asy.Var}[g_1(c_0)] = \frac{4}{n}\bar{G}_1(c_0)'\hat{\Phi}_1^{-1}\bar{G}_1(c_0)$$

The Wald statistic would be

$$\text{Wald} = g_1(c_0)' \{\text{Est.Asy.Var}[g_1(c_0)]\} g_1(c_0)$$

$$= n\bar{m}_1(c_0)\hat{\Phi}_1^{-1}\bar{G}(c_0) \left\{\bar{G}(c_0)'\hat{\Phi}_1^{-1}\bar{G}(c_0)\right\}^{-1}\bar{G}(c_0)'\hat{\Phi}_1^{-1}\bar{m}_1'(c_0)$$
Chapter 19

Models with Lagged Variables

19.1 Dynamic Regression Models

In some settings, economic agents respond not only to current values of independent variables but to past values as well.

19.1.1 Lagged Effects in A Dynamic Model

The general form of a dynamic regression model is

\[ y_t = \alpha + \sum_{i=0}^{\infty} \beta_i x_{t-i} + \epsilon_t \]

In this model, a one-time change in \( x \) at any point in time will affect \( E[y_s|x_t, x_{t-1},...] \) in every period thereafter.

The response of \( y \) to a change in \( x \) is assumed to be immediate and to be complete at the end of the period of measurement. In a dynamic model, the counterpart to a marginal effect is the effect of a one-time change in \( x_t \) on the equilibrium of \( y_t \). If the level of \( x_t \) has been unchanged from, say, \( \bar{x} \) for many periods prior to time \( t \), then the equilibrium value of \( E[y_t|x_t, x_{t-1},...] \) will be

\[ \bar{y} = \alpha + \sum_{i=0}^{\infty} \beta_i \bar{x} = \alpha + \bar{x} \sum_{i=0}^{\infty} \beta_i \]

where \( \bar{x} \) is the permanent value of \( x_t \). For this value to be finite, we require that

\[ \left| \sum_{i=0}^{\infty} \beta_i \right| < \infty \]

**DEFINITION**  Impact Multiplier

\( \beta_0 \) = impact multiplier=short-run multiplier.

**DEFINITION**  Cumulated Effect

The accumulated effect \( \tau \) periods later of an impulse at time \( t \) is \( \beta_\tau = \sum_{i=0}^{\tau} \beta_0 \)
**DEFINITION**  
Equilibrium Multiplier

\[ \beta = \sum_{i=0}^{\infty} \beta_i = \text{equilibrium multiplier}=\text{long-run multiplier}. \]

Define the Lag weights:

\[ w_i = \frac{\beta_i}{\sum_{j=0}^{\infty} \beta_j} \]

so that \( \sum_{i=0}^{\infty} w_i = 1 \), and to rewrite the model as

\[ y_t = \alpha + \beta \sum_{i=0}^{\infty} w_i x_{t-i} + \epsilon_t \]

Two useful statistics, based on the lag weights, that characterize the period of adjustment to a new equilibrium are the median lag = smallest \( q^* \) such that \( \sum_{i=q^*}^{\infty} w_i \geq 0.5 \) and the mean lag = \( \sum_{i=0}^{\infty} i w_i \).

### 19.1.2 The Lag and Difference Operators

A convenient device for manipulating lagged variables is the lag operator,

\[ L x_t = x_{t-1} \]

A related operation is the first difference,

\[ \Delta x_t = x_t - x_{t-1} \]

obviously, \( \Delta x_t = (1 - L)x_t \) and \( x_t = x_{t-1} + \Delta x_t \).

The dynamic regression model can be written

\[ y_t = \alpha + \beta \sum_{i=0}^{\infty} \beta_i L^i x_t + \epsilon_t = \alpha + B(L) x_t + \epsilon_t \]

where \( B(L) \) is a polynomial in \( L \), \( B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + ... \). A polynomial in the lag operator that reappears in many contexts is

\[ A(L) = 1 + aL + (aL)^2 + (aL)^3 + ... = \sum_{i=0}^{\infty} (aL)^i \]

If \( |a| < 1 \), then

\[ A(L) = \frac{1}{1-aL} \]

A distributed lag model in the form

\[ y_t = \alpha + \beta \sum_{i=0}^{\infty} \gamma^i L^i x_t + \epsilon_t \]

can be written

\[ y_t = \alpha + \beta (1 - \gamma L)^{-1} x_t + \epsilon_t \]
if $|\gamma| < 1$. This form is called the **moving-average form** or **distributed lag form**. If we multiply through by $(1 - \gamma L)$ and collect terms, then we obtain the **autoregressive form**,

$$y_t = \alpha(1 - \gamma) + \beta x_t + \gamma y_{t-1} + (1 - \gamma L)\epsilon_t$$

In more general terms, consider the $p$th order autoregressive model,

$$y_t = \alpha + \beta x_t + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \epsilon_t$$

which may be written

$$C(L)y_t = \alpha + \beta x_t + \epsilon_t$$

where

$$C(L) = 1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p$$

Formally,

$$y_t = [C(L)]^{-1}(\alpha + \beta x_t + \epsilon_t) = A(L)(\alpha + \beta x_t + \epsilon_t)$$

By this arrangement, it follows that $C(L)A(L) = 1$ where

$$A(L) = (\alpha_0 L^0 - \alpha_1 L - \alpha_2 L^2 - \cdots)$$

By collecting like powers of $L$ in

$$(1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p)(\alpha_0 L^0 + \alpha_1 L + \alpha_2 L^2 + \cdots) = 1,$$

If the system is stable in this sense, then the polynomial $C(L)$ is said to be **invertible**.

Finally, two useful results are

$$B(1) = \beta_0 1^0 + \beta_1 1^1 + \beta_2 1^2 + \cdots = \beta = \text{long-run multiplier}$$

and

$$B'(1) = [dB(L)/dL]_{L=1} = \sum_{i=0}^{\infty} i\beta_i$$

It follows that $B'(1)/B(1) = \text{mean lag}$.

### 19.1.3 Specification Search for the Lag Length

Let us suppose that there is an appropriate, “true” value of $p > 0$ that we seek. A simple-to-general approach to finding the right lag length would depart from a model with only the current value of the independent variable in the regression, and add deeper lags until a simple $t$ test suggested that the last one added is statistically insignificant.

A general-to-simple approach would begin from a model that contains more than $p$ lagged values—it is assumed that though the precise value of $p$ is unknown, the analyst can posit a maintained value that should be larger than $p$. Least squares or instrumental variables regression of $y$ on a constant and $(p + d)$ lagged values of $x$ consistently estimates $\theta = [\alpha, \beta_0, \beta_1, \ldots, \beta_p, 0, 0, \ldots]$.

Since models with lagged values are often used for forecasting, researchers have tended to look for measures that have produced better results for assessing “out of sample” prediction properties. The adjusted $R^2$ is one possibility. Others include the Akaike (1973) information criterion, AIC($p$),

$$AIC(p) = \ln \frac{e' e}{T} + \frac{2p}{T}$$
and Schwartz’s criterion, $SC(p)$:

$$SC(p) = AIC(p) + \left( \frac{p}{T} \right) (\ln T - 2)$$

If some maximum $P$ is known, then $p < P$ can be chosen to minimize $AIC(p)$ or $SC(p)$. An alternative approach, also based on a known $P$, is to do sequential $F$ tests on the last $P > p$ coefficients, stopping when the test rejects the hypothesis that the coefficients are jointly zero.

19.2 Simple Distributed Lag Models

19.2.1 Finite Distributed Lag Models

An unrestricted finite distributed lag model would be specified as

$$y_t = \alpha + \sum_{i=0}^{p} \beta_i x_{t-i} + \epsilon_t$$

We also assume that $\epsilon_t$ is distributed with mean zero and variance $\sigma^2$. If the lag length $p$ is known, then the equation above is a classical regression model.

The polynomial model assumes that the true distribution of lag coefficients can be well approximated by a low-order polynomial,

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \cdots + \alpha_p i^q, \quad i = 0, 1, \ldots, p > q$$

After substituting and collecting terms, we obtain

$$y_t = \gamma + \alpha_0 \left( \sum_{i=0}^{p} i^0 x_{t-i} \right) + \alpha_1 \left( \sum_{i=0}^{p} i^1 x_{t-i} \right) + \cdots + \alpha_q \left( \sum_{i=0}^{p} i^q x_{t-i} \right) + \epsilon_t$$

$$= \gamma + \alpha_0 z_0 + \alpha_1 z_1 + \cdots + \alpha_q z_q + \epsilon_t$$

Each $z_{jt}$ is a linear combination of the current and $p$ lagged values of $x_t$. With the assumption of strict exogeneity of $x_t$, $\gamma$ and $(\alpha_0, \alpha_1, \ldots, \alpha_q)$ can be estimated by ordinary or generalized least squares. The parameters of the regression model, $\beta_i$ and asymptotic standard errors for the estimators can then be obtained using the delta method.

19.2.2 An Infinite Lag Model: the Geometric Lag Model

The general form of the model is

$$y_t = \alpha + \beta \sum_{i=1}^{\infty} (1 - \lambda) \lambda^i x_{t-i} + \epsilon_t, \quad 0 < \lambda < 1$$

$$= \alpha + \beta B(L)x_t + \epsilon_t$$

where

$$B(L) = (1 - \lambda)(1 + \lambda L + \lambda^2 L^2 + \lambda^3 L^3 + \cdots) = \frac{1 - \lambda}{1 - \lambda L}$$

The lag coefficients are $\beta_i = \beta (1 - \lambda) \lambda^i$. The model incorporates infinite lags, but it assigns arbitrarily small weights to the distant past. The lag weights decline geometrically;

$$w_i = (1 - \lambda) \lambda^i, \quad 0 \leq w_i < 1$$
The mean lag is
\[ \bar{w} = \frac{B'(1)}{B(1)} = \frac{\lambda}{1 - \lambda} \]

The median lag is \( p^* \) such that \( \sum_{i=0}^{p^* - 1} w_i = 0.5 \). We can solve for \( p^* \) by using the result
\[ \sum_{i=0}^{p} \lambda^i = \frac{1 - \lambda^{p+1}}{1 - \lambda} \]

Thus,
\[ p^* = \frac{\ln 0.5}{\ln \lambda} - 1 \]

The impact multiplier is \( \beta(1 - \lambda) \). The long run multiplier is \( \beta \sum_{i=0}^{\infty} (1 - \lambda)\lambda^i = \beta \).

The geometric lag model can be motivated with an economic model of expectations. We begin with a regression in an expectations variable such as an expected future price based on information available at time \( t \), \( x^*_{t+1|t} \), and perhaps a second regressor, \( w_t \),
\[ y_t = \alpha + \beta x^*_{t+1|t} + \delta w_t + \epsilon_t \]

and a mechanism for the formation of the expectation,
\[ x^*_{t+1|t} = \lambda x^*_{t|t-1} + (1 - \lambda)x_t = \lambda L x^*_{t+1|t} + (1 - \lambda)x_t \]

The currently formed expectation is a weighted average of the expectation in the previous period and the most recent observation. The expectation variable can be written as
\[ x^*_{t+1|t} = \frac{1 - \lambda}{1 - \lambda L} x_t = (1 - \lambda)[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots] \]

Inserting produces the geometric distributed lag model,
\[ y_t = \alpha + \beta(1 - \lambda)[x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots] + \delta w_t + \epsilon_t \]

The geometric lag model can be estimated by nonlinear least squares. Rewrite it as
\[ y_t = \alpha + \gamma z_t(\lambda) + \delta w_t + \epsilon_t, \quad \gamma = \beta(1 - \lambda) \]

The constructed variable \( z_t(\lambda) \) obeys the recursion \( z_t(\lambda) = x_t + \lambda z_{t-1}(\lambda) \). For the first observation, we use \( z_1(\lambda) = x^*_{1|0} = x_1/(1 - \lambda) \). One can then scan over the range of \( \lambda \) from zero to one to locate the value that minimizes the sum of squares. Once the minimum is located, an estimate of the asymptotic covariance matrix of the estimators of \( (\alpha, \gamma, \delta, \lambda) \) can be found. For the regression function \( h_t(data|\alpha, \gamma, \delta, \lambda) \), \( x^*_1 = 1 \), \( x^*_2 = z_t(\lambda) \), and \( x^*_3 = w_t \). The derivative with respect to \( \lambda \) can be computed by using the recursion
\[ d_t(\lambda) = \partial z_t(\lambda)/\partial \lambda = z_{t-1}(\lambda) + \lambda \partial z_{t-1}(\lambda)/\partial \lambda \].

If \( z_1 = x_1/(1 - \lambda) \), then \( d_1(\lambda) = z_1/(1 - \lambda) \). Then \( x^*_4 = d_t(\lambda) \). Finally, we estimate \( \beta \) from the relationship \( \beta = \gamma/(1 - \lambda) \) and use the delta method to estimate the asymptotic standard error.

The partial adjustment model describes the desired level of \( y_t \)
\[ y^*_t = \alpha + \beta x_t + \delta w_t + \epsilon_t \]

and an adjustment equation,
\[ y_t - y_{t-1} = (1 - \lambda)(y^*_t - y_{t-1}) \]
If we solve the second equation for $y_t$ and insert the first expression for $y_t^*$, then we obtain

$$y_t = \alpha(1 - \lambda) + \beta(1 - \lambda)x_t + \delta(1 - \lambda)w_t + \lambda y_{t-1} + (1 - \lambda)\epsilon_t$$

$$= \alpha' + \beta'x_t + \delta'w_t + \lambda y_{t-1} + \epsilon'_t$$

The parameters of this model can be estimated consistently and efficiently by ordinary least squares.

19.3 Autoregressive Distributed Lag Models

Autoregressive Distributed Lag (ARDL) model,

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{j=0}^{r} \beta_j x_{t-j} + \delta w_t + \epsilon_t$$

in which $\epsilon_t$ is assumed to be serially uncorrelated and homoscedastic. We can write this more compactly as

$$C(L)y_t = \mu + B(L)x_t + \delta w_t + \epsilon_t$$

by defining polynomials in the lag operator,

$$C(L) = 1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p$$

and

$$B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_r L^r$$

The model in this form is denoted ARDL($p, r$) to indicate the orders of the two polynomials in $L$.

19.3.1 Estimation of the ARDL Model

The ARDL is a linear model with a classical disturbance. As such, ordinary least squares is the efficient estimator. For testing restrictions, the Wald statistic can be used, although the $F$ statistic is generally preferable in finite samples because of its more conservative critical values.

19.3.2 Computation of the Lag Weights in the ARDL Model

The distributed lag form of the ARDL model is

$$y_t = \frac{\mu}{C(L)} + \frac{B(L)}{C(L)} + \frac{1}{C(L)}\delta w_t + \frac{1}{C(L)}\epsilon_t$$

$$= \frac{\mu}{1 - \gamma_1 - \cdots - \gamma_p} + \sum_{j=0}^{\infty} \alpha_j x_{t-j} + \delta \sum_{l=0}^{\infty} \theta_{l} w_{t-l} + \sum_{l=0}^{\infty} \theta_{l} \epsilon_{t-l}$$

The lag coefficients on $x_t, x_{t-1}, \ldots$ in the ARDL model are the individual terms in the ratio of polynomials that appear in the distributed lag form. We denote these as coefficients

$$\alpha_0, \alpha_1, \alpha_2, \ldots = \text{the coefficient on } 1, L, L^2, \ldots \text{ in } \frac{B(L)}{C(L)}.$$
A convenient way to compute these coefficients is to write the equation above as \( A(L)C(L) = B(L) \). Then we can just equate coefficients on the powers of \( L \).

The long-run effect in a rational lag model is \( \sum_{i=0}^{\infty} \alpha_i \). This result is easy to compute since it is simply

\[
\sum_{i=0}^{\infty} \alpha_i = \frac{B(1)}{C(1)}
\]

A standard error for the long-run effect can be computed using the delta method.

### 19.3.3 Stability of a Dynamic Equation

Suppose that \( x_t \) is fixed at some value \( \bar{x} \), \( w_t \) is fixed at zero, and the disturbances \( \epsilon_t \) are fixed at their expectation of zero. Would \( y_t \) converge to an equilibrium?

The relevant dynamic equation is

\[
y_t = \bar{\alpha} + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \epsilon_t
\]

where \( \bar{\alpha} = \mu + B(1)\bar{x} \). If \( y_t \) converges to an equilibrium, then, that equilibrium is

\[
\bar{y} = \frac{\mu + B(1)\bar{x}}{C(1)} = \frac{\bar{\alpha}}{C(1)}
\]

Stability of a dynamic equation hinges on the characteristic equation for the autoregressive part of the model. The roots of the characteristic equation

\[
C(z) = 1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_p z^p = 0
\]

must be greater than one in absolute value for the model to be stable.

The univariate autoregression,

\[
y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \epsilon_t
\]

can be augmented with the \( p-1 \) equations

\[
y_{t-1} = y_{t-1} \\
y_{t-2} = y_{t-2}
\]

and so on to give a vector autoregression, VAR

\[
y_t = \mu + Cy_{t-1} + \epsilon_t
\]

where \( y_t \) has \( p \) elements \( \epsilon_t = (\epsilon_t, 0, \ldots)' \) and \( \mu = (\mu, 0, 0, \ldots)' \). Now, by successive substitution, we obtain

\[
y_t = \mu + C\mu + C^2\mu + \cdots,
\]

which may or may not converge. Write \( C \) in the spectral form \( C = \Lambda P \Lambda Q \), where \( PQ = I \) and \( \Lambda \) is a diagonal matrix of the characteristic roots. We then obtain

\[
y_t = \sum_{i=0}^{\infty} P \Lambda^i Q \mu
\]

If all the roots of \( C \) are less than one in absolute value, then this vector will converge to the equilibrium

\[
y_\infty = (I - C)^{-1} \mu
\]

Nonexplosion of the powers of the roots of \( C \) is equivalent to \( |\lambda_p| < 1 \), or \( |1/\lambda_p| > 1 \), which was our original requirement. At equilibrium, therefore, we must have \( y_t = y_{t-1} = \cdots = y_\infty \).
19.3.4 Forecasting

Consider, first, a one-period-ahead forecast of $y_t$ in the ARDL($p,r$) model. It will be convenient to collect the terms in $\mu, x_t, w_t$, and so on in a single term,

$$
\mu_t = \mu + \sum_{j=0}^{r} \beta_j x_{t-j} + \delta w_t
$$

Now, the ARDL model is just

$$
y_t = \mu_t + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + \epsilon_t
$$

Conditioned on the full set of information available up to time $T$ and on forecasts of the exogenous variables, the one-period-ahead forecast of $y_t$ would be

$$
\hat{y}_{T+1} = \hat{\mu}_{T+1} + \gamma_1 \hat{y}_T + \cdots + \gamma_p \hat{y}_{T-p+1} + \hat{\epsilon}_{T+1}
$$

To form a prediction interval, we will be interested in the variance of the forecast error,

$$
\sigma^2 = \text{Var}[\epsilon_{T+1} | \hat{\mu}_{T+1} + 1]
$$

Ignoring for the moment the variation in $\hat{\mu}_{T+1}$. The variance of the forecast error will be simply

$$
\text{Var}[\epsilon_{T+1} | x_{T+1}, w_{T+1}, \mu, \beta, \delta, y_T, \ldots] = \text{Var}[\epsilon_{T+1}] = \sigma^2
$$

Let $z_{T+1} = [1, x_{T+1}, x_T, \ldots, x_{T-r+1}, w_T, y_T, y_{T-1}, \ldots, y_{T-p+1}]$, and let $\hat{\theta}$ denote the full estimated parameter vector. Then we would use

$$
\text{Est. Var}[\epsilon_{T+1} | z_{T+1}] = s^2 + z_{T+1}' \left( \text{Est. Asy.Var}[\hat{\theta}] \right) z_{T+1}
$$

Now, consider forecasting further out beyond the sample period:

$$
\hat{y}_{T+2} = \hat{\mu}_{T+2} + \gamma_1 \hat{y}_{T+1} + \cdots + \gamma_p \hat{y}_{T-p+2} + \hat{\epsilon}_{T+2}
$$

Note that for period $T+1$, the forecasted $y_{T+1}$ is used. Making the substitution for $\hat{y}_{T+1}$, we have

$$
\hat{y}_{T+2} = \hat{\mu}_{T+2} + \gamma_1 (\hat{\mu}_{T+1} + \gamma_1 y_T + \cdots + \gamma_p y_{T-p+1} + \hat{\epsilon}_{T+1}) + \cdots + \gamma_p y_{T-p+2} + \hat{\epsilon}_{T+2}
$$

and, likewise, for subsequent periods. For the first forecast period, write the forecast with the previous $p$ lagged values as

$$
\begin{bmatrix}
\hat{y}_{T+1} \\
y_T \\
y_{T-1} \\
\vdots
\end{bmatrix} =
\begin{bmatrix}
\hat{\mu}_{T+1} \\
0 \\
0 \\
\vdots
\end{bmatrix} +
\begin{bmatrix}
\gamma_1 & \gamma_2 & \cdots & \gamma_p \\
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots
\end{bmatrix}
\begin{bmatrix}
y_T \\
y_{T-1} \\
y_{T-2} \\
\vdots
\end{bmatrix} +
\begin{bmatrix}
\hat{\epsilon}_{T+1} \\
0 \\
0 \\
\vdots
\end{bmatrix}
$$

The coefficient matrix on the right-hand side is $C$. With this modification, then, our forecast is the top element of the vector of forecasts,

$$
\hat{y}_{T+1} = \hat{\mu}_{T+1} + Cy_T + \hat{\epsilon}_{T+1}
$$
Now, extend this notation to forecasting out to periods $T + 2, T + 3$ and so on:

$$\hat{y}_{T+2|T} = \hat{\mu}_{T+2|T} + C\hat{y}_{T+1|T} + \hat{\epsilon}_{T+2|T}$$
$$\hat{y}_{T+3|T} = \hat{\mu}_{T+3|T} + C\hat{\mu}_{T+2|T} + C^2y_{T} + \hat{\epsilon}_{T+2|T} + C\hat{\epsilon}_{T+1|T}$$

Once again, the only unknowns are the disturbances, so the forecast variance for this two-period-ahead forecasted vector is

$$\text{Var}[\hat{\epsilon}_{T+2|T} + C\hat{\epsilon}_{T+1|T}] = \begin{bmatrix} \sigma^2 & 0 & \ldots \\ 0 & 0 & \vdots \\ \vdots & \ldots & \vdots \end{bmatrix} + C\begin{bmatrix} \sigma^2 & 0 & \ldots \\ 0 & 0 & \vdots \\ \vdots & \ldots & \vdots \end{bmatrix} C'$$

Thus, the forecast variance for the two-step-ahead forecast is $\sigma^2[1 + \Phi(1)_{11}]$, where $\Phi(1)_{11}$ is the 1, 1 element of $\Phi(1) = Cjj' C'$, where $j' = [\sigma, 0, \ldots, 0]$. By extending this device to a forecast $F$ periods beyond the sample period, we obtain

$$\hat{y}_{T+F|T} = \sum_{f=1}^{F} C^{f-1}\hat{\mu}_{T+F-(f-1)|T} + C^Fy_T + \sum_{f=1}^{F} C^{f-1}\hat{\epsilon}_{T+F-(f-1)|T}$$

We also obtain our expression for the conditional forecast variance,

$$\text{Conditional Var}[\hat{y}_{T+F|T}] = \sigma^2[1 + \Phi(1)_{11} + \Phi(2)_{11} + \cdots + \Phi(F-1)_{11}]$$

where $\Phi(i) = C^{i}j''j'C''$.

### 19.4 Methodological Issues in the Analysis of Dynamic Models

#### 19.4.1 An Error Correction Model

Consider the ARDL(1,1) model, which has become a workhorse of the modern literature on time-series analysis. By defining the first differences $\Delta y_t = y_t - y_{t-1}$ and $\Delta x_t = x_t - x_{t-1}$ we can rearrange

$$y_t = \mu + \gamma_1y_{t-1} + \beta_0x_t + \beta_1x_{t-1} + \epsilon_t$$

to obtain

$$\Delta y_t = \mu + \beta_0\Delta x_t + (\gamma_1 - 1)(y_{t-1} - \theta x_{t-1}) + \epsilon_t$$

where $\theta = -(\beta_0 + \beta_1)/(\gamma_1 - 1)$. This form of the model is in the error correction form. In this form, we have an equilibrium relationship, $\Delta y_t = \mu + \beta_0\Delta x_t + \epsilon_t$, and the equilibrium error, $(\gamma_1 - 1)(y_{t-1} - \theta x_{t-1})$, which account for the deviation of the pair of variables from that equilibrium.

#### 19.4.2 Autocorrelation

Consider once again the simplest model of autocorrelation form Chapter 12.

$$y_t = \beta x_t + v_t, \quad v_t = \rho v_{t-1} + \epsilon_t$$
where $\epsilon_t$ is nonautocorrelated. As we found earlier, this model can be written as

$$y_t - \rho y_{t-1} = \beta(x_t - \rho x_{t-1}) + \epsilon_t$$

or

$$y_t = \rho y_{t-1} + \beta x_t - \beta x_{t-1}) + \epsilon_t$$

This model is an ARDL(1,1) model in which $\beta_1 = -\gamma_1\beta_0$.

Now, let us take this argument to its logical conclusion. The AR(p) disturbance model,

$$v_t = \rho_1 v_{t-1} + \cdots + \rho_p v_{t-p} + \epsilon_t$$

or $R(L)v_t = \epsilon_t$, can be written in its moving average form as

$$v_t = \frac{\epsilon_t}{R(L)}$$

The regression model with this AR(p) disturbance is, therefore,

$$y_t = \beta x_t + \frac{\epsilon_t}{R(L)}$$

But consider instead the ARDL(p,p) model

$$C(L)y_t = \beta B(L)x_t + \epsilon_t$$

These coefficients are the same model if $B(L) = C(L)$. The implication is that any model with an $AR(p)$ disturbance can be interpreted as a nonlinearly restricted version of an ARDL(p,p) model.

### 19.4.3 Common Factor Restrictions

In general, we can formulated it as a test of

$$H_0 : y_t = x_t'\beta + \rho y_{t-1} - \rho(x_{t-1}'\beta) + \epsilon_t$$

versus

$$H_1 : y_t = x_t'\beta + \rho y_{t-1} + x_{t-1}'\gamma + \epsilon_t$$

Since the models are both classical regression models, the test can be carried out by referring the $F$ statistic

$$F[J, T - K] = \frac{(\epsilon_0'\epsilon_0 - \epsilon_1'\epsilon_1)/J}{\epsilon_1'\epsilon_1/(T - K)}$$

to the appropriate critical value from the $F$ distribution.

In an AR(1) model, this common factor restriction takes the form

$$(1 - \gamma L)y_t = (\beta_0 + \beta_1 L)x_1 + \epsilon_t, \quad \beta_1 = -\gamma\beta_0$$

Consider, instead, an AR(2) model. The “restricted” and unrestricted models would appear as

$$H_0 : (1 - \rho_1 L - \rho_2 L^2)y_t = (1 - \rho_1 L - \rho_2 L^2)x_t'\beta + \epsilon_t$$

$$H_1 : y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + x_t'\beta_0 + x_{t-1}'\beta_1 + x_{t-2}'\beta_2 + \epsilon_t$$
so the full set of restrictions is $\beta_1 = -\gamma_1 \beta_0$ and $\beta_2 = -\gamma_2 \beta_0$. This expanded model can be handled analogously to the AR(1) model. Once again, an $F$ test of the nonlinear restrictions can be used.

The polynomials in the lag operator, $C(L)$ and $B(L)$, can be factored into products of linear, primitive terms. A quadratic equation in $L$, for example may always be written as
\[
C(L) = (1 - \gamma_1 L - \gamma_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L)
\]
where the $\lambda$'s are the roots of the characteristic polynomials $C(z) = 0$. Here, $B(L)$ may be factored likewise, say into $(1 - \tau_1 L)(1 - \tau_2 L)$. With these results in hand, rewrite the basic model $C(L)y_t = B(L)x_t + \epsilon_t$ in the form
\[
(1 - \lambda_1 L)(1 - \lambda_2 L)y_t = (1 - \lambda_1 L)(1 - \lambda_2 L)x_t \beta + \epsilon_t
\]
Now suppose that $\lambda_1 = \tau_1 = \rho$. Dividing through both sides of the equation by $(1 - \rho L)$ produces the restricted model
\[
(1 - \lambda_2 L)y_t = (1 - \tau_2 L)x_t \beta + \frac{\epsilon_t}{1 - \rho L}
\]
Its disturbance is an AR(1) process in $\rho$. The restricted model is appropriate only if the two polynomials have a common factor, $(1 - \lambda_2) = (1 - \tau_2)$, hence the name for the procedure.

Write the distributed lag part, $B(L)$, as $\beta_0(1 - \beta_1 L - \beta_2 L^2)$. Multiplying out the factors, we see that the unrestricted model,
\[
y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \beta_0(1 - \beta_1 - \beta_2 L^2)x_t + \epsilon_t
\]
can be written as
\[
y_t = \mu + (\lambda_1 + \lambda_2)y_{t-1} - (\lambda_1 \lambda_2)y_{t-2} + \beta_0 x_t - \beta_0(\tau_1 + \tau_2)x_{t-1} + \beta_0(\tau_1 \tau_2)x_{t-2} + \epsilon_t
\]
this equation is intrinsically linear. Now impose the common factor restriction $(1 - \lambda_1) = (1 - \tau_1)$, or $\lambda_1 = \tau_1$. The now very nonlinear regression model
\[
y_t = \mu + (\tau_1 + \lambda_2)y_{t-2} - (\lambda_1 \lambda_2)y_{t-2} + \beta_0 x_t - \beta_0(\tau_1 + \tau_2)x_{t-1} + \beta_0(\tau_1 \tau_2)x_{t-2} + \epsilon_t
\]
has six terms on the right-hand side but only five parameters and is overidentified. This model can be fit as is by nonlinear least squares. The $F$ test of one restriction suggested earlier can now be carried out.

### 19.5 Vector Autoregressions

The resulting autoregressive model is
\[
y_t = \mu + \mu_1 y_{t-1} + \cdots + \mu_p y_{t-p} + \epsilon_t
\]
where $\epsilon_t$ is a vector of nonautocorrelated disturbances with zero means and contemporaneous covariance matrix $E[\epsilon_t \epsilon_t'] = \Omega$. This equation system is a vector autoregression, or VAR. The equation above can also be written as
\[
\Gamma(L)y_t = \mu + \epsilon_t
\]
where $\Gamma(L)$ is a matrix of polynomials in the lag operator. The individual equations are
\[
y_{mt} = \mu_m + \sum_{j=1}^{p} (\Gamma_j)_{m1} y_{1,t-j} + \sum_{j=1}^{p} (\Gamma_j)_{m2} y_{2,t-j} + \cdots + \sum_{j=1}^{p} (\Gamma_j)_{mm} y_{M,t-j} + \epsilon_{mt}
\]
where $(\Gamma_j)_{lm}$ indicates the $(l, m)$ element of $\Gamma_j$. 193
19.5.1 Model Forms

To simplify things for the present, we note that the \( p \)th order VAR can be written as a first-order VAR as follows:

\[
\begin{bmatrix}
    y_t \\
    y_{t-1} \\
    \vdots \\
    y_{t-p+1}
\end{bmatrix} =
\begin{bmatrix}
    \mu \\
    0 \\
    \vdots \\
    0
\end{bmatrix} +
\begin{bmatrix}
    \Gamma_1 & \Gamma_2 & \cdots & \Gamma_p \\
    1 & 0 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & \cdots & I & 0
\end{bmatrix}
\begin{bmatrix}
    y_{t-1} \\
    y_{t-2} \\
    \vdots \\
    y_{t-p}
\end{bmatrix} +
\begin{bmatrix}
    \epsilon_t
\end{bmatrix}
\]

The VAR may also be viewed as the reduced form of a simultaneous equations model; the corresponding structure would then be

\[
\Theta y_t = \alpha + \Psi y_{t-1} + \omega_t
\]

where \( \Theta \) is a nonsingular matrix and \( \text{Var}[\omega] = \Sigma \).

19.5.2 Estimation

Without autocorrelation of the disturbances–Vars are particularly simple to estimate. Although the equation system can be exceedingly large, the equations should be estimated separately by OLS. The disturbance covariance matrix can then be estimated with average sums of squares or cross-products of the least squares residuals. If the disturbances are normally distributed, then these least squares estimators are also maximum likelihood.

19.5.3 Testing Procedures

We begin by assuming that the disturbances have a joint normal distribution. Let \( W \) be the \( M \times M \) residual covariance matrix based on a restricted model, and let \( W^* \) be its counterpart when the model is unrestricted. Then the likelihood ratio statistic,

\[
\lambda = T (\ln |W| - \ln |W^*|)
\]

can be used to test the hypothesis. The statistic would have a limiting chi-squared distribution with degrees of freedom equal to the number of restrictions. Formally, suppose the appropriate lag length is \( p \) but the model is fit with \( q \geq p + 1 \) lagged terms. Then, under the null hypothesis,

\[
\lambda_q = T [\ln |W(\mu, \Gamma_1, ..., \Gamma_{q-1})| - \ln |W^*(\mu, \Gamma_1, ..., \Gamma_q)|] \xrightarrow{d} \chi^2[M^2]
\]

For specification searches for the right lag. Lutkepohl (1993, pp.128-135) suggests an alternative approach based on the minimizing functions of the information criteria we have considered earlier;

\[
\lambda^* = \ln(|W|) + (pM^2 + M)IC(T)/T
\]

where \( T \) is the sample size, \( p \) is the number of lags, \( M \) is the number of equations and \( IC(T) = 2 \) for the Akaike information criterion and \( \ln T \) for the Schwartz (Bayesian) information criterion. We should note, this is not a test statistic; it is a diagnostic tool that we are using to conduct a specification search.
Let $c$ be the sample estimator of $\gamma$ and let $V$ denote the estimated asymptotic covariance matrix. Then, the hypothesis in equation can be cast in the form $R\gamma - q = 0$. The Wald statistic for testing the null hypothesis is

$$W = (Rc - q)'[RV R']^{-1}(Rc - q)$$

Under the null hypothesis, this statistic has a limiting chi-squared distribution with degrees of freedom equal to $J$, the number of restrictions (rows in $R$). For the specification search for the appropriate lag length, the null hypothesis will be that a certain subvector of $\gamma$, say $\gamma_0$, equals zero. In this case, the statistic will be

$$W_0 = c_0'V_{00}^{-1}c_0$$

where $V_{00}$ denotes the corresponding submatrix of $V$.

### 19.5.4 Exogeneity

Assuming that the regression relationship is linear

$$y_t = E[y_t|x_t] + (y_E[y_t|x_t])$$

where the familiar moment condition $E[x_tE_t] = 0$ follows by construction. But, this form of the model is no more the “correct” equation than would be

$$x_t = \delta_1 + \delta_2 y_t + \omega_t$$

which is

$$x_t = E[x_t|y_t] + (x_t - E[x_t|y_t])$$

and now, $E[y_t\omega_t] = 0$. Since both equations are correctly specified in the context of the bivariate distribution, there is nothing to define one variable or the other as “exogenous”.

A second form of exogeneity we will consider **strong exogeneity**, which is sometimes called **Granger noncausality**. Granger noncausality can be superficially defined by the assumption

$$E[y_t|y_{t-1}, x_{t-1}, x_{t-2}, ...] = E[y_t|y_{t-1}]$$

That is, lagged values of $x_t$ do not provide information about the conditional mean of $y_t$ once lagged values of $y_t$, itself, are accounted for. For the present, we note that most of the models we will examine will explicitly fail this assumption.

To put this back in the context of our model, we will be assuming that in the model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \gamma y_{t-1} + \epsilon_t$$

and the extensions that we will consider, $x_t$ is weakly exogenous—we can meaningfully estimate the parameters of the regression equation independently of the marginal distribution of $x_t$, but we will allow for Granger causality between $x_t$ and $y_t$, thus generally not assuming strong exogeneity.
19.5.5 Testing for Granger Causality

Causality in the sense defined by Granger (1969) and Sims (1972) is inferred when lagged values of a variable, say $x_t$, have explanatory power in a regression of a variable $y_t$ on lagged values of $y_t$ and $x_t$. The VAR can be used to test the hypothesis. Tests of the restrictions can be based on simple $F$ tests in the single equations of the VAR model.

Partition the two data vectors $y_t$ and $x_t$ into $[y_{1t}, y_{2t}]$ and $[x_{1t}, x_{2t}]$. Consistent with our earlier discussion, $x_1$ is lagged values of $y_1$ and $x_2$ is lagged values of $y_2$. The VAR with this partitioning would be

$$
\begin{bmatrix}
  y_1 \\
  y_2
\end{bmatrix}
= 
\begin{bmatrix}
  \Gamma_{11} & \Gamma_{12} \\
  \Gamma_{21} & \Gamma_{22}
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2
\end{bmatrix}
+ 
\begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2
\end{bmatrix},
\text{Var}
\begin{bmatrix}
  \epsilon_1 t \\
  \epsilon_2 t
\end{bmatrix}
= 
\begin{bmatrix}
  \Sigma_{11} & \Sigma_{12} \\
  \Sigma_{21} & \Sigma_{22}
\end{bmatrix}
$$

We would still obtain the unrestricted MLE by least squares regressions. For testing Granger causality, the hypothesis $\Gamma_{12} = 0$ is of interest. For testing the hypothesis of interest, $\Gamma_{12} = 0$, the second set of equations is irrelevant. For testing for Granger causality in the VAR model, only the restricted equations are relevant. The hypothesis can be tested using the likelihood ratio statistic. For the present application, testing means computing

$S_{11} = $ residual covariance matrix when current values of $y_1$ are regressed on values of both $x_1$ and $x_2$

$S_{11}(0) = $ residual covariance matrix when current values of $y_1$ are regressed only on values of both $x_1$

The likelihood ratio statistic is then

$$
\lambda = T(\ln |S_{11}(0)| - \ln |S_{11}|)
$$

The number of degrees of freedom is the number of zero restrictions. The Wald test or its transformation to an approximate $F$ statistic is an alternative that should be more generally applicable. Another possibility is to use the GMM counterpart to the likelihood ratio statistic based on the GMM criterion functions.

19.5.6 Impulse Response Functions

As we analyzed earlier, in the model

$$
y_t = \mu + \Gamma y_{t-1} + v_t
$$

dynamic stability is achieved if the characteristic roots of $\Gamma$ have modulus less than one. Using the lag operator to write

$$
y_t = \mu + \Gamma(L)y_t + v_t
$$
or

$$
[I - \Gamma(L)]y_t = \mu + v_t
$$

With the stability condition, we have

$$
y_t = [I - \Gamma(L)]^{-1}(\mu + v_t)
= (I - \Gamma)^{-1}\mu + \sum_{i=0}^{\infty} \Gamma^i v_{t-i}
= \bar{y} + \sum_{i=0}^{\infty} \Gamma^i v_{t-i}
= \bar{y} + v_t + \Gamma v_{t-1} + \Gamma^2 v_{t-2} + \cdots
$$
The coefficients in the powers of $\Gamma$ are the multipliers in the system. Suppose that $v$ has equaled 0 for long enough that $y$ has reached equilibrium, $\bar{y}$. Now we consider injecting a shock to the system by changing one of the $v$’s, for one period, and then returning it to zero thereafter. As we saw earlier, $y_{mt}$ will move away from, then return to, its equilibrium. The path whereby the variables return to the equilibrium is called the **impulse response** of the VAR.

In the autoregressive form of the model, we can identify each innovation, $v_{mt}$, with a particular variable in $y_t$, say $y_{mt}$. Consider then the effect of a one-time shock to the system, $dv_m$. As compared with the equilibrium, we will have, in the current period,

$$y_{mt} - \bar{y}_m = dv_{mt} = \phi_{mm}(0) dv_t$$

One period later, we will have

$$y_{m,t+1} - \bar{y}_m = (\Gamma)_{mm} dv_{mt} = \phi_{mm}(1) dv_t$$

Two periods later,

$$y_{m,t+2} - \bar{y}_m = (\Gamma^2)_{mm} dv_{mt} = \phi_{mm}(2) dv_t$$

and so on. The function, $\phi_{mm}(i)$ gives the impulse response characteristics of variable $y_m$ to innovations in $v_m$.

The impulse response function would be

$$\phi_{ml}(i) = \text{element } (m, l) \text{ in } \Gamma^i$$

Point estimation of $\phi_{ml}(i)$ using the estimated model parameters is straightforward.

### 19.5.7 Structural VARs

The VAR approach to modeling dynamic behavior of economic variables has provided some interesting insights and appears to bring some real benefits for forecasting.

A VAR model $y_t = \mu + \Gamma y_{t-1} + v_t$ could, in principle, be viewed as the reduced form of the dynamic **structural model**

$$\Theta y_t = \alpha + \Phi y_{t-1} + \epsilon_t$$

where we have embedded any exogenous variables $x_t$ in the vector of constants $\alpha$. Thus, $\Delta = \Theta^{-1} \Phi, \mu = \Theta^{-1} \alpha, v = \Theta^{-1} \epsilon,$ and $\Omega = \Theta^{-1} \Sigma (\Theta^{-1})'$. For particular configurations of $\Theta$, such as a triangular matrix, we can meaningfully interpret innovations, $\epsilon$. However, there is not sufficient information contained in the reduced form as just stated to deduce the structural parameters. A possibly large number of restrictions must be imposed on $\Theta, \Phi$ and $\Sigma$ to enable us to deduce structural forms from reduced-form estimates, which are always obtainable. The recent work on “structural VARs” centers on the types of restrictions and forms of the theory that can be brought to bear to allow this analysis to proceed.