Dynamic Econometric Models

Time Series Econometrics for Microeconometricians

Walter Beckert
Department of Economics
Birkbeck College, University of London
Institute for Fiscal Studies

26 - 27 May 2011, DIW Berlin

1 Introduction

1.1 Overview

This course provides an introduction to dynamic econometric models and methods.

The course surveys linear and nonlinear econometric models and estimation techniques, presenting them in a method of moments framework. While emphasizing their applicability under general assumptions on the data generating process, the emphasis will be on applications in time series analysis.

The first part of the course treats single equation models, while the second part is devoted to systems of equations. Starting from a review of the linear regression model (OLS, GLS, FGLS), the course revisits basic properties of stochastic processes and their implications for time-series regressions, cast in the form of general autore-

gressive distributed lag (ARDL) and error correction model (ECM) representations.

The second part of the course is devoted to estimation of systems of equations that describe the joint evolution of several time series. The primary focus is on vector autoregressive models (VARs). The concept of co-integration of time series is introduced, and its implications for VARSs is explored in the context of the vector error correction model (VECM) representation of VARs, as a multivariate generalization of ECMs for AR(DL)s.

An supplementary section focusses on second moment properties of stochastic processes. Specifically, it is devoted to time series models of heteroskedasticity which play a prominent role in the analysis of the volatility of financial time series.

The course is designed as a two-day sequence of alternating lectures and practical computer exercises. The applications in the computer practicals will use time series data from microeconomic contexts, rather than macroeconomic series.

1.2 About these Notes

These notes are intended as a reference guide to the material covered in the course. The lectures will follow the notes closely, but will focus on the main principles and results, omitting much of the intermittent algebra. The presentation of the course material rests on the kind of mathematical and statistical tools and the styles of argument that microeconometricians are typically familiar with. The primary objective is to provide an approach to econometric concept in time series analysis that appeals to the intuitive understanding of microeconometricians, not a fully rigorous delineation of results.

2 Generalized Method of Moments Estimation

2.1 General Setup and Method of Moments Estimation

This section provides a basic review of Method of Moments estimation in the familiar context of the linear regression model. It sets up the general framework and notation in which the remainder of the course and these notes will proceed.

Consider data $\{(y_t, \mathbf{x}_t'), t = 1, \cdots, T\}$, where y_t denotes a scalar dependent (or response) variable, while $\mathbf{x}_t$ denotes a $k \times 1$ vector of independent, exogenous covariates. Possible assumptions about the data generating process are:

1. distributional assumptions about the joint or conditional cumulative distribution function (CDF) $F(\mathbf{y}, \mathbf{X})$, or $F(\mathbf{y}|\mathbf{X})$ respectively, where $\mathbf{y} = (y_1, \cdots, y_T)'$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)'$; this setup gives rise to maximum likelihood estimation (MLE);

2. conditional population moment assumptions:

   (i) $E_{\mathbf{Y}|\mathbf{X}}[y_t|\mathbf{x}_t] = g(\mathbf{x}_t; \theta_0)$ a.s. for all t, where $\theta_0 \in \Theta \subset \mathbb{R}^k$ is an unknown parameter vector, and the function g is possibly nonlinear in θ_0; in the special case of linearity, $E_{\mathbf{Y}|\mathbf{X}}[y_t|\mathbf{x}_t] = \mathbf{x}_t'\theta_0$ a.s., the linear regression model; in the latter case, this is equivalent to $E_{\mathbf{Y}|\mathbf{X}}[y_t - \mathbf{x}_t'\theta_0|\mathbf{x}_t] = 0$ a.s. for all t;

   (ii) continuing with the linear model, $E_{\mathbf{Y}|\mathbf{X}}[(y_t - \mathbf{x}_t\theta_0)^2|\mathbf{x}_t] = \sigma_0^2 > 0$ a.s. for all t, which is referred to as conditional homoskedasticity.

Note: (i) by itself does not identify θ_0, unless $k = 1$; (ii) identifies σ_0^2. Based on (i), unconditional moment conditions can be derived by iterated expectations:

$$\begin{aligned} E_{\mathbf{Y}|\mathbf{X}}[y_t - \mathbf{x}_t'\theta_0|\mathbf{x}_t] &= 0 \text{ a.s.} \\ \Rightarrow \mathbf{x}_t E_{\mathbf{Y}|\mathbf{X}}[y_t - \mathbf{x}_t'\theta_0|\mathbf{x}_t] &= \mathbf{0} \text{ a.s.} \\ \Rightarrow E_{\mathbf{X}}\left[\mathbf{x}_t E_{\mathbf{Y}|\mathbf{X}}[y_t - \mathbf{x}_t'\theta_0|\mathbf{x}_t]\right] &= \mathbf{0} \\ \text{(i')} \quad \Rightarrow E_{\mathbf{YX}}\left[\mathbf{x}_t(y_t - \mathbf{x}_t'\theta_0)\right] &= \mathbf{0} = m(y_t, \mathbf{x}_t; \theta_0), \end{aligned}$$

i.e. k unconditional moments, which can identify θ_0. Note also that (ii) holds unconditionally as well: $E_{\mathbf{YX}}[(y_t - \mathbf{x}_t'\theta_0)^2] = \sigma_0^2$ for all t.

The idea behind Method of Moments (MOM) estimation of θ_0 and σ_0^2 is to replace population moments by sample analogues (empirical moments, sample averages): For any $\theta \in \Theta$,

$$\begin{aligned} \text{moments in (i')}: \; E_T[\mathbf{x}_t(y_t - \mathbf{x}_t'\theta)] &= \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{x}_t'\theta) = m_T(\mathbf{y}, \mathbf{X}; \theta) \\ \text{moments in (ii)}: \; E_T[(y_t - \mathbf{x}_t'\theta)^2)] &= \frac{1}{T}\sum_{t=1}^{T} (y_t - \mathbf{x}_t'\theta)^2. \end{aligned}$$

The MOM estimators $\hat{\theta}_T$ and $\hat{\sigma}_T^2$ solve the empirical analogues to (i') and (ii):

$$\begin{aligned} \text{(iii)} \quad E_T[\mathbf{x}_t(y_t - \mathbf{x}_t'\hat{\theta}_T)] &= \mathbf{0} \\ \text{(iv)} \quad E_T[(y_t - \mathbf{x}_t'\hat{\theta}_T)^2)] &= \hat{\sigma}_T^2. \end{aligned}$$

In this linear model, the MOM estimator for θ_0 is equivalent to the familiar OLS estimator: (iii) implies

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{x}_t'\hat{\theta}_T) &= \mathbf{0} \\ \left(\frac{1}{T}\sum_t \mathbf{x}_t\mathbf{x}_t'\right)\hat{\theta}_T &= \frac{1}{T}\sum_t \mathbf{x}_t y_t \\ E_T\left[\mathbf{x}_t\mathbf{x}_t'\right]\hat{\theta}_T &= E_T[\mathbf{x}_t y_t] \\ \hat{\theta}_T &= \left[E_T\left[\mathbf{x}_t\mathbf{x}_t'\right]\right]^{-1} E_T[\mathbf{x}_t y_t] \\ &= \left[\sum_t \mathbf{x}_t\mathbf{x}_t'\right]^{-1} \sum_t \mathbf{x}_t y_t \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \hat{\theta}_{\text{OLS}}, \end{aligned}$$

provided $\text{rk}(\mathbf{X}'\mathbf{X}) = k$. Hence, $\hat{\theta}_T$ is conditionally unbiased: $E\left[\hat{\theta}_T|\mathbf{X}\right] = \theta_0$. Its conditional variance is $\text{var}(\hat{\theta}_T|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{var}(\mathbf{y}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$; provided that the y_t are conditionally independent across t, i.e. that $\text{var}(\mathbf{y}|\mathbf{X}) = \sigma_0^2\mathbf{I}_T$, the conditional variance of the MOM estimator reduces to $\text{var}(\hat{\theta}_T|\mathbf{X}) = \sigma_0^2(\mathbf{X}'\mathbf{X})^{-1}$. In this case, the MOM estimator enjoys all the properties of the OLS estimator, a direct consequence of the Gauss-Markov Theorem which rests entirely on conditional moment

assumptions: *Suppose* $E_{\mathbf{Y}|\mathbf{X}}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\theta_0 = [\mathbf{x}_t'\theta_0]_{t=1,\cdots,T}$, *and* $\text{var}(\mathbf{y}|\mathbf{X}) = \sigma_0^2\mathbf{I}_T$, $\sigma_0^2 > 0$; *then* $\hat{\theta}_T$ *is the best linear unbiased estimator (BLUE), i.e. it is efficient (in the sense of having minimum variance among all linear, unbiased estimators of* θ_0*).*

The moment conditions (iv), involving second moments, yield

$$\hat{\sigma}_T^2 = \frac{1}{T}\sum_t (y_t - \mathbf{x}_t'\hat{\theta}_T)^2 = \frac{T-k}{T}s_T^2,$$

where s_T^2 is the OLS estimator of σ_0^2. This implies that the MOM estimator $\hat{\sigma}_T^2$ is biased in small samples (finite T).

2.2 Variants and Generalizations

2.2.1 IV and 2SLS

Suppose that the previous population orthogonality conditions between $\mathbf{x}_t$ and residuals $y_y - \mathbf{x}_t'\theta_0$ do not hold, but that for some vector of instruments $\mathbf{z}_t$, for any t,

$$\begin{aligned} E_{\mathbf{YX}|\mathbf{Z}}[y_t - \mathbf{x}_t'\theta_0|\mathbf{z}_t] &= 0 \text{ a.s.} \\ E_{\mathbf{YXZ}}[\mathbf{z}_t(y_t - \mathbf{x}_y'\theta_0)] &= \mathbf{0} \qquad (1) \\ \text{var}(\mathbf{y}|\mathbf{X},\mathbf{Z}) &= \sigma_0^2\mathbf{I}_T, \end{aligned}$$

where $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_T)'$.

Suppose (i) $\dim(\mathbf{z}_t) = \dim(\mathbf{x}_t) = \dim(\theta_0) = k$, in which case θ_0 is just identified by (1). Then, the sample analogue to (1) is

$$E_T[\mathbf{z}_t(y_t - \mathbf{x}_t\hat{\theta}_{IV})] = \mathbf{0} \quad \Rightarrow \quad \hat{\theta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y},$$

provided $\text{rk}(\mathbf{Z}'\mathbf{X}) = k$. Under the conditional homoskedasticity assumption above, the instrumental variable (IV) estimator $\hat{\theta}_{IV}$ has conditional variance $\text{var}(\hat{\theta}_{IV}|\mathbf{X},\mathbf{Z}) = \sigma_0^2(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{X})^{-1}$, which reduces to the conditional variance of the OLS estimator when $\mathbf{X}$ is a valid array of instruments, i.e. when the orthogonality conditions (1) hold with $\mathbf{z}_t = \mathbf{x}_t$.

Suppose (ii) $\dim(\mathbf{z}_t) = m > k$. In this case, θ_0 is over-identified by (1), which is now a system of m (rather than k) equations, and $\mathbf{Z}'\mathbf{X}$ is an $m \times k$ matrix, i.e. it is

not square. Let $P_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, the orthogonal projector onto the column space of $\mathbf{Z}$, col($\mathbf{Z}$).; recall that orthogonal projectors are idempotent and symmetric. Then, orthogonality of $\mathbf{y} - \mathbf{X}\theta_0$ and $\mathbf{Z}$ according to (1) implies orthogonality of $\mathbf{y} - \mathbf{X}\theta_0$ and $\mathbf{X}'P_{\mathbf{Z}}$, so that

$$E_{\mathbf{YXZ}}[\mathbf{X}'P_{\mathbf{Z}}(\mathbf{y} - \mathbf{X}\theta_0)] = \mathbf{0} \qquad (1')$$

is a system of k unconditional moment conditions. The sample analogue to (1') is

$$\begin{aligned} \frac{1}{T}\mathbf{X}'P_{\mathbf{Z}}(\mathbf{y} - \mathbf{X}\hat{\theta}_{2SLS}) &= \mathbf{0} \\ \hat{\theta}_{2SLS} &= (\mathbf{X}'P_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'P_{\mathbf{Z}}\mathbf{y} \qquad \text{(provided rk}(\mathbf{X}'P_{\mathbf{Z}}\mathbf{X}) = k) \\ &= (\mathbf{X}'P_{\mathbf{Z}}P_{\mathbf{Z}}'\mathbf{X})^{-1}\mathbf{X}'P_{\mathbf{Z}}\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \end{aligned}$$

where $\hat{\mathbf{X}} = P_{\mathbf{Z}}\mathbf{X}$ are the fitted values from the regression of the columns of $\mathbf{X}$ onto $\mathbf{Z}$, i.e. only those components of $\mathbf{X}$ which are orthogonal to $\mathbf{y} - \mathbf{X}\theta_0$ according to (1). The conditional variance of the 2SLS estimator is $\text{var}(\hat{\theta}_{2SLS}|\mathbf{X}, \mathbf{Z}) = \sigma_0^2(\mathbf{X}'P_{\mathbf{Z}}\mathbf{X})^{-1}$. Note that, if $\dim(\mathbf{z}_t) = k$ and $\text{rk}(\mathbf{Z}'\mathbf{X}) = k$, then $(\mathbf{X}'P_{\mathbf{Z}}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}((\mathbf{Z}'\mathbf{X})^{-1})'$, i.e. the conditional variance collapses to the one of the IV estimator. Note also, for future reference, that the conditional variances of the IV moment functions are

$$\text{var}(\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta_0)|\mathbf{X}, \mathbf{Z}) = \sigma_0^2(\mathbf{Z}'\mathbf{Z}) = \text{var}(\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta_0)|\mathbf{Z}).$$

2.2.2 Non-scalar Variance-Covariance Matrix

Suppose, as at the outset, that $E_{\mathbf{YX}}[\mathbf{x}_t(y - \mathbf{x}_t'\theta_0)] = 0$ for all t, but $\text{var}(\mathbf{y}|\mathbf{X}) = \Omega$, a positive definite, symmetric $T \times T$ matrix. This change in the second moment assumptions can be expected to affect the second moment properties of the OLS/MOM estimator $\hat{\theta}_T$, i.e. its conditional variance-covariance matrix and, thereby, its efficiency.

As before, the moment conditions involving the first moments yield the OLS/MOM estimator for θ_0, $\hat{\theta}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The second moment assumptions, however, now

imply

$$\begin{aligned}\text{var}(\hat{\theta}_T|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{var}(\mathbf{y}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

The Gauss-Markov Theorem implies that, while $\hat{\theta}_T$ is still conditionally unbiased, it is no longer efficient. Note also: The conditional variance of the moment functions is now $\text{var}(\mathbf{X}'(\mathbf{y} - \mathbf{X}\theta_0)|\mathbf{X}) = \mathbf{X}'\Omega\mathbf{X}$.

Consider the special case of conditional heteroskedasticity in which $\text{var}(\mathbf{y}|\mathbf{X}) = \text{diag}(\sigma_1^2, \cdots, \sigma_T^2)$. In this case, the conditional moment functions $h(y_t, \mathbf{x}_t; \theta_0) = y_t - \mathbf{x}_t'\theta_0$ have conditional variances $\text{var}(y_t - \mathbf{x}_t'\theta_0|\mathbf{x}_t) = \sigma_t^2$ for any t. In contrast to the homoskedastic case, this suggests to weight the conditional moment functions inversely proportional to their respective conditional variances, so that more informative (precise) moment conditions receive higher weight. Following this logic, the weighted conditional moment functions are $\frac{1}{\sigma_t^2}h(y_t, \mathbf{x}_t; \theta_0) = \frac{1}{\sigma_t^2}(y_t - \mathbf{x}_t'\theta_0)$, so that the weighted conditional moments are

$$E_{\mathbf{Y}|\mathbf{X}}\left[\frac{1}{\sigma_t^2}(y_t - \mathbf{x}_t'\theta_0)\middle|\mathbf{x}_t\right] = 0 \text{ a.s.},$$

and the weighted unconditional moments are

$$E_{\mathbf{YX}}\left[\mathbf{x}_t\frac{1}{\sigma_t^2}(y_t - \mathbf{x}_t'\theta_0)\right] = \mathbf{0}.$$

Their sample analogues are

$$\begin{aligned}E_T\left[\mathbf{x}_t\frac{1}{\sigma_t^2}(y_t - \mathbf{x}_t'\hat{\theta}_{GLS})\right] &= \frac{1}{T}\sum_t \mathbf{x}_t\frac{1}{\sigma_t^2}(y_t - \mathbf{x}_t'\hat{\theta}_{GLS}) = \mathbf{0} \\ \Leftrightarrow \frac{1}{T}\mathbf{X}'\Omega^{-1}(\mathbf{y} - \mathbf{X}\hat{\theta}_{GLS}) &= 0, \text{ where } \Omega^{-1} = \text{diag}(\frac{1}{\sigma_1^2}, \cdots, \frac{1}{\sigma_T^2}) \\ \Rightarrow \hat{\theta}_{GLS} &= (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y},\end{aligned}$$

provided $\text{rk}(\mathbf{X}'\Omega^{-1}\mathbf{X}) = k$. The conditional variance of the GLS estimator is $\text{var}(\hat{\theta}_T|\mathbf{X}) = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}$. Note that homoskedasticity is a special case in which $\Omega = \sigma_0^2\mathbf{I}_T$, $\hat{\theta}_{GLS} = \hat{\theta}_{OLS}$ and $\text{var}(\hat{\theta}_T|\mathbf{X}) = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1} = \text{var}(\hat{\theta}_{OLS}|\mathbf{X}) = \sigma_0^2(\mathbf{X}'\mathbf{X})^{-1}$. The GLS estimator is efficient among linear, unbiased estimators under the above assumptions. In this framework, $\hat{\theta}_{GLS}$ has the interpretation of efficient MOM estimator.

The GLS estimator above is only feasible if Ω is known. If it is not known, it needs to be estimated, based on first-stage residuals obtained from consistent, but inefficient OLS estimation of θ_0. Once a consistent estimator $\hat{\Omega}_T$ is obtained, θ_0 can be re-estimated in a second step, using $\hat{\Omega}_T$ in lieu of Ω:

$$\hat{\theta}_{FGLS} = (\mathbf{X}'\hat{\Omega}_T^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}_T^{-1}\mathbf{y}.$$

This feasible GLS (FGLS) estimator is asymptotically equivalent to the infeasible GLS estimator, i.e. it is asymptotically efficient.

2.3 Generalized Method of Moments

This line of reasoning suggests that it is generally beneficial (in the sense of efficiency) to weight moment functions by their conditional variances. The Generalized Method of Moments (GMM) proceeds in this fashion.[1]

To illustrate this, re-consider the instrumental variable set-up above, with $\dim(\mathbf{z}_t) = m \geq k$ and $\text{var}(\mathbf{y}|\mathbf{X}, \mathbf{Z}) = \sigma_0^2\mathbf{I}_T$. In this case, as shown above, the moment functions $\mathbf{z}_t(y_t - \mathbf{x}_t'\theta_0)$ have conditional variance $\text{var}(\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta_0)|\mathbf{X}, \mathbf{Z}) = \sigma_0^2(\mathbf{Z}'\mathbf{Z})$.

Starting with an arbitrary positive definite, symmetric weighting matrix Σ of dimension $m \times m$, the GMM estimator minimizes the generalized distance from zero of the empirical moments, relative to the metric defined by Σ:

$$\hat{\theta}_{GMM} = \arg\min_{\theta\in\Theta} E_T\left[\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta)\right]' \Sigma E_T\left[\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta)\right].$$

The first-order conditions of the minimization problem define the GMM estimator $\hat{\theta}_{GMM}$; in this case:

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\hat{\theta}_{GMM})'\mathbf{Z}\Sigma\mathbf{Z}'\mathbf{X} &= \mathbf{0} \\ \hat{\theta}_{GMM} &= (\mathbf{X}'\mathbf{Z}\Sigma\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\Sigma\mathbf{Z}'\mathbf{y},\end{aligned}$$

[1] Hansen, L.P. (1982): "Large Sample Properties of Generalized Methods of Moments Estimators", *Econometrica*, **50(4)**, 1029-1054; and Hansen, L.P. and K.J. Singleton (1982): "Generalized Instrumental Variables Estimators of Nonlinear Rational Expectations Models", *Econometrica*, **50(5)**, 1269-1286.

with conditional variance

$$\text{var}(\hat{\theta}_{GMM}|\mathbf{X},\mathbf{Z};\Sigma) = (\mathbf{X}'\mathbf{Z}\Sigma\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\Sigma\mathbf{Z}'\sigma_0^2\mathbf{I}_T\mathbf{Z}\Sigma\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{Z}\Sigma\mathbf{Z}'\mathbf{X})^{-1}.$$

This variance-covariance matrix is minimized with respect to Σ by choosing $\Sigma^\star = \frac{1}{\sigma_0^2}(\mathbf{Z}'\mathbf{Z})^{-1} = [\text{var}(\mathbf{Z}'(\mathbf{y}-\mathbf{X}\theta_0)|\mathbf{X},\mathbf{Z})]^{-1}$, i.e. by weighting the moment functions inversely proportional to their conditional variances. With this optimal weighting matrix in place, the optimal GMM estimator is seen to be equivalent to the 2SLS estimator:

$$\text{var}(\hat{\theta}_{GMM}|\mathbf{X},\mathbf{Z};\Sigma^\star) = \sigma_0^2(\mathbf{X}'P_\mathbf{Z}\mathbf{X})^{-1}.$$

This generalizes to general nonlinear moment functions with sufficient smoothness.

2.4 Testing of Moment Conditions

2.4.1 Hausman Test

The Hausman test examines the consistency of MOM estimators in the face of possible failures of moment conditions.[2]

Suppose $\tilde{\theta}_T$ and $\hat{\theta}_T$ are two estimators of θ_0, obtained on the basis of different assumptions about valid moment restrictions; e.g. $\tilde{\theta}_T$ uses moments beyond those used by $\hat{\theta}_T$. The null hypothesis H_0 is that both $\hat{\theta}_T$ and $\tilde{\theta}_T$ are $\sqrt{T}$ consistent; i.e., in the example, that the additional moments are valid, so that $\tilde{\theta}_T$ is relatively more efficient than $\hat{\theta}_T$. Under H_0,

$$\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) \xrightarrow{d} N(\mathbf{0}, V_D),$$

for some asymptotic variance-covariance matric V_D, which may be singular. The alternative hypothesis H_A implies that $\lim_{T\to\infty}\Pr\left(|\hat{\theta}_T - \tilde{\theta}_T| > \epsilon\right) > 0$ for any $\epsilon > 0$. The Hausman test statistic takes the usual quadratic form

$$\mathcal{H}_T = T\left(\tilde{\theta}_T - \hat{\theta}_T\right)'\hat{V}_D^-\left(\tilde{\theta}_T - \hat{\theta}_T\right),$$

[2]Hausman, J.A. (1978): "Specification Tests in Econometrics", *Econometrica*, **46(5)**, 1251-1271.

where $\hat{V}_D^-$ is a consistent estimator of the (generalized) inverse of V_D. Under the null hypothesis, its asymptotic distribution is χ^2 with degrees of freedom equal to the number of restrictions imposed by the null hypothesis.

Example: In the context of endogenous regressors, the OLS estimator $\hat{\beta}_{\text{OLS}}$ is best linear unbiased if $\mathbb{E}[\mathbf{X}'\mathbf{u}] = \mathbf{0}$, but biased otherwise. If $\mathbf{Z}$ is an array of valid instruments, then the IV/2SLS estimator $\hat{\beta}_{\text{IV/2SLS}}$ is unbiased, regardless of whether $\mathbb{E}[\mathbf{X}'\mathbf{u}] = \mathbf{0}$ holds or not, but if this moment condition holds, then it is inefficient, relative to the OLS estimator. Then, this moment condition can be tested using the Hausman testing framework. Since $\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV/2SLS}} | \mathbf{X}, \mathbf{Z} \sim N(\mathbf{0}, \mathbf{V})$, where $\mathbf{V} = \text{var}\left(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV/2SLS}}\right)$, the Hausman test statistic is

$$\mathcal{H}_T = \left(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV/2SLS}}\right)' \mathbf{V}^- \left(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV/2SLS}}\right),$$

and the null hypothesis of the validity of the k moment conditions is rejected at the α-level if $\mathcal{H}_T > \chi^2_k(1-\alpha)$. Note that it follows from the orthogonality of relatively efficient estimators that, under the null hypothesis,

$$\begin{aligned} \text{cov}\left(\hat{\beta}_{\text{OLS}}, \hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV/2SLS}}\right) &= 0 \\ \Rightarrow \text{var}\left(\hat{\beta}_{\text{OLS}}\right) &= \text{cov}\left(\hat{\beta}_{\text{OLS}}, \hat{\beta}_{\text{IV/2SLS}}\right). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbf{V} &= \text{var}\left(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV/2SLS}}\right) \\ &= \text{var}\left(\hat{\beta}_{\text{IV/2SLS}}\right) - \text{var}\left(\hat{\beta}_{\text{OLS}}\right) \\ &= \sigma_0^2 \left[(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\right]. \end{aligned}$$

Often, however, the latter matrix is singular.

A convenient fact often facilitates the computation of the Hausman test statistic $\mathcal{H}_T$. A consequence of the (conditional) orthogonality of a relative efficient estimator and its difference to other consistent, but inefficient estimators is that the (conditional) covariances between such estimators is equal to the variance of the efficient estimator. Hence, if $\tilde{\theta}_T$ is efficient relative to $\hat{\theta}_T$, then $V_D = \text{avar}(\sqrt{T}(\tilde{\theta}_T - \hat{\theta}_T)) = \text{avar}(\sqrt{T}(\hat{\theta}_T - \theta_0)) - \text{avar}(\sqrt{T}(\tilde{\theta}_T - \theta_0))$.

As an example, consider the linear, homoskedastic model and let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ consists of exogenous covariates, while $\mathbf{X}_2$ is suspected of lack of exogeneity. In other words, the validity of the set of unconditional moment conditions $E[\mathbf{X}_2(\mathbf{y} - \mathbf{X}_1\theta_1 - \mathbf{X}_2\theta_2)] = \mathbf{0}$ is in doubt. Let $\mathbf{W}$ be an array of instruments for $\mathbf{X}_2$ in case $\mathbf{X}_2$ is endogenous, and let $\mathbf{Z} = [\mathbf{X}_1, \mathbf{W}]$ denote the array of all instruments (i.e. the columns of $\mathbf{X}_1$ act as instruments for themselves). Then, the null hypothesis H_0 is that $\mathbf{X}_2$ is exogenous, while the alternative hypothesis H_A is that $\mathbf{X}_2$ is not exogenous. Under H_0, the Gauss-Markov Theorem implies that the OLS estimator for $\theta_0 = (\theta_1', \theta_2')'$, $\hat{\theta}_{OLS}$, is efficient; its asymptotic distribution is $\sqrt{T}(\hat{\theta}_{OLS} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \sigma_0^2(\mathbf{X}'\mathbf{X})^{-1})$, conditional on $\mathbf{X}$. Under H_A, $\hat{\theta}_{OLS}$ is inconsistent, but the 2SLS estimator $\hat{\theta}_{2SLS}$ is consistent; its asymptotic distribution is $\sqrt{T}(\hat{\theta}_{2SLS} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \sigma_0^2(\mathbf{X}'P_{\mathbf{Z}}\mathbf{X})^{-1})$, conditional on $\mathbf{X}$ and $\mathbf{Z}$. Since the OLS and 2SLS estimators are (conditionally) orthogonal under the null hypothesis, their conditional covariance matrix is zero under H_0. Hence, conditional on $\mathbf{X}$ and $\mathbf{Z}$,

$$\sqrt{T}(\hat{\theta}_{OLS} - \hat{\theta}_{2SLS}) \xrightarrow{d} \; N\left(0, \sigma_0^2((\mathbf{X}'P_{\mathbf{Z}}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1})\right),$$

so that the test statistic becomes

$$\mathcal{H}_T = T\left(\hat{\theta}_{OLS} - \hat{\theta}_{2SLS}\right)' \frac{[(\mathbf{X}'P_{\mathbf{Z}}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^-}{\hat{\sigma}_T^2}\left(\hat{\theta}_{OLS} - \hat{\theta}_{2SLS}\right) \xrightarrow{d} \chi^2_{\dim(\mathrm{col}(\mathbf{X}_2))},$$

where $\hat{\sigma}_T^2$ is an estimator of σ_0^2 based on either the OLS or 2SLS regression residuals. This test is referred to as Hausman-Wu Exogeneity Test.

The computation of the Hausman-Wu test statistics is complicated by the requirement of a generalized inverse. A convenient representation of the test in an easily implementable form was suggested by Wu (1973)[3]. The null hypothesis of exogeneity of $\mathbf{X}$ is equivalent to the null hypothesis that $\gamma = 0$ in the augmented regression

$$\mathbf{y} = \mathbf{X}\theta + \hat{\mathbf{X}}_2\gamma + \epsilon, \quad \text{where } \hat{\mathbf{X}}_2 = P_{\mathbf{Z}}\mathbf{X}_2.$$

This hypothesis can be tested using an F-test with $\dim(\mathrm{col}(\mathbf{X}_2))$ numerator and $\dim(\mathrm{col}(\mathbf{X})) - \dim(\mathrm{col}(\mathbf{X}_2))$ denominator degrees of freedom. The null hypothesis of exogeneity is also equivalent to the hypothesis that $\delta = 0$ in the regression

$$\mathbf{y} = \mathbf{X}\theta + \hat{\mathbf{u}}\delta + \nu, \quad \text{where } \hat{\mathbf{u}} = (\mathbf{I} - P_{\mathbf{Z}})\mathbf{X}_2,$$

[3]Wu, D. (1973): "Alternative Tests of Independence between Stochastic Regressors and Disturbances", *Econometrica*, **41(4)**, 733-775.

where $\hat{\mathbf{u}}$ is a set of the vectors of fitted residuals from the reduced form regressions of the hypothesized endogenous RHS variables onto all exogenous variables. This hypothesis can be tested using a t-test with $\dim(\text{col}(\mathbf{X}_2)) = 1$ degrees of freedom if $\mathbf{X}_2 \in \mathbb{R}$, and an F-test as above otherwise.

2.4.2 Sargan-Hansen J Test

Another test of the validity of moment conditions can be based on the GMM criterion function. When the parameter vector of interest θ_0 is exactly identified under the alternative hypothesis and over-identified under the null hypothesis, then GMM moment tests are called test of over-identifying restrictions. Let $E_{\mathbf{YX}}[m(y_t, \mathbf{x}_t; \theta_0)] = \mathbf{0}$ denote the r population moment conditions under the null hypothesis, where $\dim(\theta_0) = k$ and $r > k$, i.e. there are $r - k$ over-identifying restrictions. The empirical analogues to the population moment functions are $E_T[m(y_t, \mathbf{x}_t; \theta_0)]$. Let $\hat{\Sigma}_T^\star$ be (a consistent estimator of) the (optimal) GMM weighting matrix $\Sigma^\star$, and let $\hat{\theta}_{GMM}$ be the GMM estimator of θ_0. The minimized, second round GMM criterion function

$$\mathcal{J}_T = TE_T\left[m(y_t, \mathbf{x}_t; \hat{\theta}_{GMM})\right]' \hat{\Sigma}_T^\star E_T\left[m(y_t, \mathbf{x}_t; \hat{\theta}_{GMM})\right]$$

then serves as a test statistic for the validity of the over-identifying moment conditions. This particular test statistic is referred to as the Sargan-Hansen (1982) J-test[4]. Its asymptotic distribution, as $T \to \infty$, is χ^2_{r-k}, and the test rejects the null hypothesis when the statistic exceeds the critical values of a χ^2_{r-k} random variable for the appropriate test size. This does not permit any inference about which of the moment conditions is invalid, however.

In the case of the example in the preceding subsection, the Sargan-Hansen J test statistic of the null hypothesis that $\mathbf{Z}$ is a valid array of instruments is

$$\mathcal{J}_T = \frac{(\mathbf{y} - \mathbf{X}\hat{\theta}_{2SLS})'\mathbf{Z}' \left[\mathbf{Z}'(\mathbf{I} - P_{\mathbf{X}})\mathbf{Z}\right]^{-1} \mathbf{Z}(\mathbf{y} - \mathbf{X}\hat{\theta}_{2SLS})}{\hat{\sigma}_T^2},$$

and its asymptotic distribution under the null hypothesis is also χ^2_{r-k}; see Appendix for details. Note that, in general, the Hausman-Wu test requires estimation under

[4]Hansen, L.P. (1982): "Large Sample Properties of Generalized Methods of Moments Estimators", *Econometrica*, **50(4)**, 1029-1054

both the null and the alternative hypothesis, while the Sargan-Hansen J test only requires estimation under the null hypothesis.

This test for over-identifying restrictions can also be implemented in terms of a regression of the 2SLS residuals $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{2SLS}$ on all the instruments $\mathbf{Z}$ and testing whether all regression coefficients are jointly equal to zero. The test statistic is asymptotically distributed χ^2 with degrees of freedom equal to the number of over-identifying restrictions.

2.4.3 Weak Instruments

Broadly speaking, the case of weak instruments refers to a situation in which the correlation between the endogenous variable and its instrument(s) is low. The treatment of situations with weak instruments is an area of active current research.[5] In the case of a single endogenous variable x_2, a test for the weakness of instruments, due to Bound et al. (1995)[6], is a partial R^2, denoted by R_p^2, that isolates the impact of the instruments on the endogenous variable, after eliminating the effect of the other exogenous variables on the latter. The statistic R_p^2 is given by the R^2 of the regression

$$x_2 - \hat{x}_2 = (\mathbf{z} - \hat{\mathbf{z}})'\delta + \nu,$$

where $\hat{x}_2 = P_{\mathbf{X}_1} x_2$ and $\hat{\mathbf{z}} = P_{\mathbf{X}_1}\mathbf{z}$. When R_p^2 is low, then z is considered an array of weak instruments for x_2.

2.4.4 Model Selection, Specification Testing and Diagnostic Tests

Various tests for model selection have been proposed in the literature, but none is entirely satisfactory. In regression models, the regression $R^2 = 1 - \hat{\mathbf{u}}'\hat{\mathbf{u}}/\mathbf{y}'\mathbf{y}$ is often considered, where $\hat{\mathbf{u}}$ is the vector of fitted residuals. Superior models exhibit smaller R^2. This measure does not require distributional assumptions and, hence,

[5] For a recent survey, see Stock and Yogo (2002), NBER Technical Working Paper 284.

[6] Bound, J., Jaeger, D.A., and R.M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation between Instruments and Endogenous Explanatory Variables Is Weak", *Journal of the American Statistical Association*, **90**, No. 430, 443-350.

is embedded in the method of moments framework. Alternatively, under distributional assumptions, measures based on the log-likelihood are available and have some information theoretic interpretation. The Akaike information criterion (AIC) adjusts the sample log-likelihood at the MLE $\hat{\theta}$ for model j, $l_T(\hat{\theta}_{(j)})$, for the number of estimated parameters, $k_j = \dim(\theta_{(j)})$, so that $AIC_j = -2l_T(\hat{\theta}_{(j)}) + 2k_j$. Under normality assumptions, the AIC reduces to $AIC_j = 2k_j + T\ln\left(\hat{\mathbf{u}}_j'\hat{\mathbf{u}}_j/T\right)$, where $\hat{\mathbf{u}}_j$ is the vector of fitted residuals of model j. The Schwarz Bayesian information or posterior odds criterion (SBC), in addition, accounts for sample size T and is defined as $SBC = -2l_T(\hat{\theta}_{(j)}) + k_j\ln(T)$. The SBC is a closely related variant of the Bayes Information Criterion (BIC), which is defined as $BIC = SBC/T$. Under normality assumptions, the BIC reduces to $BIC = \ln\left(\hat{\mathbf{u}}_j'\hat{\mathbf{u}}_j/T\right) + k_j\ln(T)/T$. Models with higher information criteria are deemed superior. In comparison to AIC, the SBC/BIC criterion tends to choose more parsimonious models. Many practitioners also test the goodness-of-fit in terms of the accuracy of out-of-sample prediction.

Diagnostic tests examine various assumptions underlying estimation. In the context of the linear regression model, this section surveys the tests which are used most often and typically provided by standard statistical software.

1. Structural Stability: Tests for structural stability examine whether the parameters to be estimated are constant over the sampling period, the null hypothesis. Considering a simple linear regression model, under the alternative hypothesis,

$$\begin{aligned} y_t &= \mathbf{x}_t'\theta_1 + \epsilon_t \;\; t = 1, \cdots, T_1, \\ y_t &= \mathbf{x}_t\theta_2 + \epsilon_t, \;\; t = T_1 + 1, \cdots, T, \end{aligned}$$

where $\theta_1 \neq \theta_2$ and ϵ_t are assumed i.i.d. and homoskedastic. Under H_0, $\theta = \theta_1 = \theta_2$, i.e. $k = \dim(\mathbf{x}_t)$ restrictions are imposed. The restricted OLS estimator for θ yields the restricted sum of squares $\hat{\epsilon}'\hat{\epsilon}$ with $T - k$ degrees of freedom, while the unrestricted OLS estimators for θ_1 and θ_2 yield the unrestricted sum of squares $\hat{\epsilon}_1'\hat{\epsilon}_1 + \hat{\epsilon}_2'\hat{\epsilon}_2$ with $T - 2k$ degrees of freedom, where $\hat{\epsilon}_{it} = y_t - \mathbf{x}_t\hat{\beta}_i$ for $i = 1$ while $t \leq T_1$, and $i = 2$ while $t > T_1$. Chow's first

breakpoint test statistic is

$$\mathcal{C}_T = \frac{\left(\hat{\epsilon}'\hat{\epsilon} - \left(\hat{\epsilon}_1'\hat{\epsilon}_1 + \hat{\epsilon}_2'\hat{\epsilon}_2\right)\right)/k}{\left(\hat{\epsilon}_1'\hat{\epsilon}_1 + \hat{\epsilon}_2'\hat{\epsilon}_2\right)/(T-k)} \quad \sim \quad F_{k,T-k}.$$

This test requires that the variances of the residuals ϵ_t are the same in both subperiods. This can be tested using the Goldfeld-Quandt test

$$\mathcal{GQ}_T = \frac{s_1^2}{s_2^2} = \frac{\hat{\epsilon}_1'\hat{\epsilon}_1/(T_1-k)}{\hat{\epsilon}_2'\hat{\epsilon}_2/(T_2-k)} \quad \sim \quad F_{T_1-k,T_2-k},$$

where the larger variance estimate should form the numerator so that the statistic is greater than unity. Chow also suggested a test for predictive failure for the case when $T_2 < k$,

$$\tilde{\mathcal{C}}_T = \frac{\left(\hat{\epsilon}'\hat{\epsilon} - \hat{\epsilon}_1'\hat{\epsilon}_1\right)/T_2}{\hat{\epsilon}_1'\hat{\epsilon}_1/(T_1-k)} \quad \sim \quad F_{T_2,T_1-k}.$$

2. Non-linearity in covariates: Using the estimated residuals from the estimation of the linear model,

$$\hat{\epsilon}_t = y_t - \mathbf{x}_t'\hat{\theta},$$

the Ramsey RESET test amounts to running a second stage regression of $\hat{\epsilon}_t$ on $\mathbf{x}_t$ and the squared predicted dependent variable $\hat{y}_t^2$ and to testing whether the coefficient on $\hat{y}_t^2$ is zero, using a t-test. A numerically equivalent implementation of the test uses y_t in lieu of $\hat{\epsilon}_t$ in the second-stage regression. Higher powers of $\hat{y}_t$ can be included to test for further degrees of curvature, using F-tests.

3. Serial Correlation in Residuals: Suppose the data generating process is

$$\begin{aligned} y_t &= \mathbf{x}_t'\theta_0 + \epsilon_t \\ \epsilon_t &= \rho\epsilon_{t-1} + \nu_t, \end{aligned}$$

where ν_t is white noise, i.e. serially uncorrelated and has mean zero and constant variance. If θ_0 were estimated by OLS, then the estimated residuals would be

$$\hat{\epsilon}_t = y_t - \mathbf{x}_t'\hat{\theta} = \mathbf{x}_t'(\theta_0 - \hat{\theta}) + \rho\epsilon_{t-1} + \nu_t.$$

This suggests to test the null hypothesis of no serial correlation, i.e. $\rho = 0$, by regressing $\hat{\epsilon}_t$ onto $\mathbf{x}_t$ and $\hat{\epsilon}_{t-1}$ and test whether the coefficient on $\hat{\epsilon}_{t-1}$ is zero,

using a t-test. Testing against the alternative hypothesis of higher-order serial correlation on the process for ϵ_t can be done analogously by including further lags of $\hat{\epsilon}_t$ and testing that their coefficients are jointly equal to zero, using an F-test.

4. Heteroskedasticity: Suppose that the residual variance is suspected to be some function $\sigma^2(\cdot)$ of a vector of variables $\mathbf{z}_t$, $\text{var}(\epsilon_t) = \sigma^2(\mathbf{z}_t)$. Then, a test for heteroskedasticity against the null hypothesis of homoskedasticity amounts to regressing OLS residuals $\hat{\epsilon}_t$ on a constant and $\mathbf{z}_t$ and testing whether the coefficient vector on $\mathbf{z}_t$ is zero, using an F-test. Candidates for $\mathbf{z}_t$ are (i) elements of the regressors $\mathbf{x}_t$, (ii) squares and cross-products of the regressors (White test), (iii) the squared predicted dependent variable $\hat{y}_t^2$ (RESET version), (iv) lagged squared estimated residuals (autoregressive conditional heteroskedasticity, ARCH), and others.

   It should be noted that cases (3.) and (4.) do not impede the usual first-moment properties of the OLS estimator for θ_0 (unbiasedness, consistency), because they pertain to second-moment assumptions. But the conditional variance-covariance matrix of $\hat{\theta}$ is no longer $\sigma_0^2(\mathbf{X}'\mathbf{X})^{-1}$, but

$$\text{var}(\hat{\theta}_{OLS}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

   where the structure of Ω is as in subsection 2.2.2 in case (4.), and as in subsection 3.1.1 below in case (3.). Corrected variance-covariance matrix estimators are available, for example the Eicker-White estimator

$$\widehat{\text{var}(\hat{\theta}_{OLS}|\mathbf{X})} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

   where $\hat{\Omega}$ is a suitable consistent estimator of Ω; see subsection 3.1.2 for an example. A popular estimator in the presence of residual serial correlation is suggested by Newey and West (1987).

5. Influential Observations

   An influential observation is a data point that is crucial to inferences drawn from the data. While the various approaches described here provide quantitative measures of the statistical influence of an observation, it is important

to keep in kind, however, that only knowledge of the subject matter and the data itself can determine whether this influence is substantively informative or merely due to data reporting error.

Consider the linear regression model in which the $k \times k$ matrix $\mathbf{X}'\mathbf{X}$ has full rank case. Define the orthogonal projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then,

$$\hat{Y}_t = \mathbf{H}_{tt}Y_t + \sum_{s \neq t} \mathbf{H}_{ts}Y_s,$$

where $\mathbf{H}_{ts}$ is the row t, column s element of $\mathbf{H}$, $t, s = 1, \cdots, T$. The value $\mathbf{H}_{tt}$ is a measure of the *leverage* or *influence* of Y_t on its linear conditional prediction, $\hat{Y}_t$. It can be shown that $0 \leq \mathbf{H}_{tt} \leq 1$, and that $\mathbf{H}_{tt} = 1$ implies that $\hat{Y}_t = Y_t$, while $\mathbf{H}_{tt} = 0$ implies that $\hat{Y}_t = 0$. This suggests that, in the case of high values of $\mathbf{H}_{tt}$, the model requires a separate parameter to fit Y_t, while in the case of low values of $\mathbf{H}_{tt}$ the prediction $\hat{Y}_t = 0$ is fixed by design, i.e. by the choice of $\mathbf{X}$. Furthermore, it can be shown that $\bar{H} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}_{tt} = \frac{k}{T}$. One definition of *leverage point* is an observation indexed t for which $\mathbf{H}_{tt} > 2\frac{k}{T}$.

Define $\hat{\theta}_{-t}$ and $s^2_{-t} = \hat{\sigma}^2_{-t}$ as the estimates of θ_0 and σ^2_0 based on the $T-1$ observations, excluding the tth data point. These are referred to as Jackknife or cross-validatory estimates.[7] These can be used to obtain Jackknife residuals

$$\epsilon^{\star}_t = (y_t - \mathbf{x}'_t\hat{\theta}_{-t})/(s_{-t}\sqrt{1 - \mathbf{H}_{tt}}),$$

which can be highly effective for picking up single outliers or influential observations. Another Jackknife measure of the influence of an observation on the joint inference regarding θ_0 are Cook's distances

$$CD_t = (\hat{\theta}_{-t} - \hat{\theta})'\frac{\mathbf{X}'\mathbf{X}}{ks^2}(\hat{\theta}_{-t} - \hat{\theta}), \;\; t = 1, \cdots, T,$$

which can be compared to an F-distribution to estimate the percentage influence of Y_t on $\hat{\theta}$.

[7]The idea of the Jackknife is due to Tukey. Based on the "leave one out" estimates $\hat{\theta}_{-t}$, $t = 1, \cdots, T$, the random variables $T\hat{\theta} - (T-1)\hat{\theta}_{-t}$ may be treated as i.i.d. estimates of θ_0. They provide an effective way to get a sampling distribution of $\hat{\theta}$ without recourse to asymptotic arguments and as an alternative to the bootstrap.

Best econometric practice usually derives an estimable statistical model from an underlying economic model or theory that rationalizes the data generating process. It is important to recognize that, while the various goodness-of-fit measures and diagnostic tests may be generally useful statistical tools for specification testing and model selection, when they fail to support the estimated model they do not provide any guidance as to how to adjust the model because they are not linked to the economic model. Failures of these tests, therefore, may be indicative of a misspecified economic model and suggest a re-examination at that level of the econometric analysis.

3 Univariate Stochastic Processes

3.1 Moving Averages

3.1.1 Stochastic Properties

Let $E[y_t - \mathbf{x}_t'\theta_0|\mathbf{x}_t] = E[\epsilon_t|\mathbf{x}_t] = 0$ for all t, but suppose that

$$\epsilon_t = y_t - \mathbf{x}_t'\theta_0 = u_t + \alpha u_{t-1},$$

where

$$u_{t-s}|\mathbf{x}_t \ \sim \ \text{i.i.d. } E[u_{t-s}|\mathbf{x}_t] = 0, E[u_{t-s}^2|\mathbf{x}_t] = \sigma_0^2, s = 0, 1, \cdots$$

This assumption about the intertemporal correlation of residuals ϵ_t induces the following additional conditional second moments

$$\begin{aligned} \text{var}(y_t - \mathbf{x}_t'\theta_0|\mathbf{x}_t) &= \sigma_0^2(1+\alpha^2) \\ \text{cov}(y_t - \mathbf{x}_t'\theta_0, y_{t-s} - \mathbf{x}_{t-s}'\theta_0|\mathbf{x}_t, \mathbf{x}_{t-s})) &= \alpha\sigma_0^2 1_{\{|s|=1\}}, \end{aligned}$$

for any t, where $1_{\{A\}}$ is an indicator function taking value 1 if the event A occurs, and zero otherwise. Hence, the conditional second moment matrix of the residuals

is bidiagonal,

$$\mathrm{var}(\mathbf{y} - \mathbf{X}\theta_0|\mathbf{X}) = \sigma_0^2 \begin{bmatrix} 1+\alpha^2 & \alpha & 0 & \cdots \\ \alpha & 1+\alpha^2 & \alpha & \cdots \\ 0 & \alpha & 1+\alpha^2 & \ddots \\ \vdots & \cdots & \ddots & \ddots \end{bmatrix} =: \Omega.$$

This model is a moving average process of order 1, MA(1). It can be generalized in a straightforward manner to higher orders, MA(q) for positive integers q. In the case of an MA(q), the conditional covariances vanish for observations more than q periods apart.

3.1.2 Estimation

This model is another instance of a non-scalar conditional variance covariance matrix (next to the heteroskedastic case discussed above), and it permits estimation of θ_0 by FGLS. In order to implement the FGLS estimator for the MA(1), the bidiagonal parameters of Ω need to be estimated from first-stage OLS residuals $\{\hat{\epsilon}_t, t = 1, \cdots, T\}$. A consistent estimator of $\sigma_0^2(1+\alpha^2)$ is $s_T^2 = E_T[\hat{\epsilon}_t^2]$, while a consistent estimator of $\alpha\sigma_0^2$ is $c_T = E_{T-1}[\hat{\epsilon}_t\hat{\epsilon}_{t-1}]$. As an aside, $\frac{s_T^2}{c_T} = \frac{1+\hat{\alpha}_T^2}{\hat{\alpha}_T}$ can be solved to obtain an estimator of α (with $\mathrm{sgn}(\hat{\alpha}_T) = \mathrm{sgn}(c_T)$), and this can be used in conjunction with c_T to obtain an estimator of σ_0^2.

3.2 Autoregressive Processes

3.2.1 Stochastic Properties

In the case of the simplest autoregressive process of order 1, AR(1), $\mathbf{x}_t$ is a scalar and takes $x_t = y_{t-1}$, so that the residuals are $\epsilon_t = y_t - \rho_0 y_{t-1}$, where ρ_0 takes the rôle of the parameter of interest θ_0. The residuals ϵ_t are assumed i.i.d. with moments $E[\epsilon_t|y_{t-1}] = E[\epsilon_t] = 0$ and $\mathrm{var}(y_t - \rho_0 y_{t-1}|y_{t-1}) = \mathrm{var}(\epsilon_t|y_{t-1}) = \mathrm{var}(\epsilon_t) = \sigma_0^2$; independence implies here, in particular, independence of past realizations of the process $\{y_t, t \geq 0\}$.

Let $\mathbf{X} = \mathbf{y}_- = (y_0, \cdots, y_{T-1})'$. Note that $\mathbf{y} - \mathbf{X}\theta_0 = \mathbf{y} - \mathbf{y}_-\rho_0 = [y_t - \rho_0 y_{t-1}]_{t=1,\cdots,T}$. While $\mathbf{y} - \mathbf{X}\theta_0|\mathbf{X}$ involved T random variables with non-degenerate distribution, its analogue in the AR(1) model is the vector $\mathbf{y} - \mathbf{y}_-\rho_0|\mathbf{y}_-$, but this involves $T-1$ constants (since it is conditioned on $\mathbf{y}_-$), and only $y_T - \rho_0 y_{T-1}|\mathbf{y}_{-1} \overset{d}{=} y_T - \rho_0 y_{T-1}|y_{T-1}$ has a non-degenerate distribution. Therefore, in the case of autoregressive processes, the joint distribution of the vector $\mathbf{y}$ conditional on initial conditions (i.e. y_0 in the case of an AR(1); on $(y_0, \cdots, y_{-p+1})$ in the case of an AR(p), for integer p) needs to be determined.

By recursive substitution,

$$\begin{aligned} y_t &= \rho_0 y_{t-1} + \epsilon_t \\ &= \rho_0(\rho_0 y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \rho_0^t y_0 + \sum_{s=0}^{t-1} \rho_0^s \epsilon_{t-s}. \end{aligned}$$

This implies the conditional moments

$$\begin{aligned} E[y_t|y_0] &= \rho_0^t y_0 \\ \text{var}(y_t|y_0) &= \sigma_0^2 \sum_{\tau=0}^{t-1} \rho_0^{2\tau} = \frac{\sigma_0^2(1-\rho_0^{2t})}{1-\rho_0^2} \\ \text{cov}(y_t, y_s|y_0) &= \sigma_0^2 \sum_{\tau=0}^{\min\{t,s\}-1} \rho_0^{2\tau} = \frac{\sigma_0^2(1-\rho_0^{2\max\{t,s\}})}{1-\rho_0^2}. \end{aligned}$$

Note that both first and second conditional moments depend on t. Without further restrictions, this would imply that any MOM estimator of ρ_0 (OLS, FGLS) would depend on t as well, which is inconsistent with the notion of ρ_0 being a time-invariant population parameter. This problem could only be overcome if the *un*conditional moments did not depend on t. Regarding the first unconditional moments, by iterated expectations

$$E[y_t] = E[E[y_t|y_0]] = \rho_0^t E[y_0],$$

which is independent of t if, and only if, $E[y_0] = E[y_t] = 0$ for all t. Regarding the

second unconditional moments,

$$\begin{aligned}
\text{var}(y_t) &= \text{var}(E[y_t|y_0]) + E[\text{var}(y_t|y_0)] \\
&= \rho_0^{2t}\text{var}(y_0) + \sigma_0^2 \frac{1}{1-\rho_0^2}(1-\rho_0^{2t}) \\
&= \frac{\sigma_0^2}{1-\rho_0^2} + \rho_0^{2t}\left[\text{var}(y_0) - \frac{\sigma_0^2}{1-\rho_0^2}\right],
\end{aligned}$$

which is independent of t if, and only if, $\text{var}(y_0) = \text{var}(y_t) = \frac{\sigma_0^2}{1-\rho_0^2}$ for all t. Note that this is only valid if $|\rho_0| < 1$. If this condition holds, then the AR(1) process $\{y_t, t \geq 0\}$ is said to be covariance stationary. Covariance stationarity will be assumed for the remainder of this section.

Assuming (covariance) stationarity, i.e. $|\rho_0| < 1$, the above results on the moments of the stationary distribution can now be obtained more easily: Discarding the trivial case $\rho_0 = 0$, for the first moments, for any t,

$$E[y_t] = \rho_0 E[y_{t-1}] = \rho_0 E[y_t] = 0,$$

and for the second moments, for any t,

$$\begin{aligned}
\text{var}(y_t) &= \text{var}(\rho_0 y_{t-1} + \epsilon_t) \\
&= \rho_0^2 \text{var}(y_t) + \sigma_0^2 \quad (\text{because } \text{cov}(y_{t-1}, \epsilon_t) = 0) \\
&= \frac{\sigma_0^2}{1-\rho_0^2} \\
\text{cov}(y_t, y_{t-s}) &= \text{cov}\left(\rho_0^s y_{t-s} + \sum_{\tau=0}^{s-1} \rho_0^\tau \epsilon_{t-\tau}, y_{t-s}\right) \\
&= \rho_0^s \text{var}(y_{t-s}) \\
&= \rho_0^s \frac{\sigma_0^2}{1-\rho_0^2}.
\end{aligned}$$

Notice that, in the case of autoregressive processes, the autocovariance function $c(s) = \text{cov}(y_t, y_{t-s})$, $s = 0, \pm 1, \pm 2, \cdots$, dies off gradually, unlike in the case of moving average processes.

3.2.2 Estimation

Estimation of ρ_0 can proceed by OLS,

$$\begin{aligned}\hat{\rho}_T &= \left(\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \sum_{t=1}^{T} y_t y_{t-1} \\ &= \rho_0 + \left(\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \sum_{t=1}^{T} y_{t-1}\epsilon_t.\end{aligned}$$

Serial independence of ϵ_t and $E[\epsilon_t] = 0$ for all t yields unbiasedness, $E[\hat{\rho}_T] = \rho_0$. Assuming covariance stationarity, the asymptotic variance of $\hat{\rho}_T$ is

$$\text{avar}(\sqrt{T}(\hat{\rho}_T - \rho_0)) = \sigma_0^2 \left(E[y_t^2]\right)^{-1} = 1 - \rho_0^2.$$

Note that $|\rho_0| \to 1$ implies $\text{avar}(\sqrt{T}(\hat{\rho}_T - \rho_0) \to 0$. This feature will be discussed at length below. Note also that, surprisingly, the asymptotic variance does not depend on the data noise σ_0^2.

3.2.3 Unit Roots

Denote the characteristic polynomial in the lag operator L of the AR(1) process by $\Phi(L) = 1 - \rho_0 L$, so that $\Phi(L)y_t = \epsilon_t$.[8] It is necessary and sufficient for the AR(1) to be stationary that the roots z of the characteristic equation $|\Phi(z)| = 0$ lie outside the unit circle, i.e. that $|z| = \frac{1}{|\rho_0|} > 1$, which is equivalent to the previous condition for covariance stationarity. If ϵ_t is also i.i.d., then this is a random walk.

Suppose that $\rho_0 = 1$, so that the characteristic equation $\Phi(z) = 0$ has a unit root. The process takes then the form

$$y_t = y_{t-1} + \epsilon_t.$$

Notice that its first difference, $y_t - y_{t-1} = \epsilon_t$ is stationary. Hence, in the case of $\rho_0 = 1$, the process $\{y_t, t \geq 0\}$ is said to be difference-stationary, or integrated of order 1, denoted by I(1). In this notation, the covariance stationary case is denoted by I(0).

[8] The lag operator L is defined by $Ly_t = y_{t-1}$.

W.l.o.g. let $y_0 = 0$ a.s. for the remainder of this section. Then,

$$\hat{\rho} - 1 = \left(\sum_{t=1}^{T} y_{t-1}^2 \right)^{-1} \sum_{t=1}^{T} y_{t-1} \epsilon_t,$$

and the conditional moments are

$$\begin{aligned} E\left[y_{t-1}^2 | y_0 = 0 \right] &= E\left[\left(\sum_{s=1}^{t-1} \epsilon_s \right)^2 \right] \\ &= \sum_{s=1}^{t-1} E[\epsilon_s^2] \\ &= (t-1)\sigma_0^2 \quad \text{a.s.}, \end{aligned}$$

so that a.s.[9]

$$\begin{aligned} E\left[\sum_{t=1}^{T} y_{t-1}^2 \middle| y_0 = 0 \right] &= \sum_{t=1}^{T} (t-1)\sigma_0^2 \\ &\approx \sigma_0^2 \int_1^T (t-1) dt \\ &\propto \sigma_0^2 T^2 = O_p(T^2). \end{aligned}$$

Similarly, $E[y_{t-1}\epsilon_t] = E\left[y_{t-1} E[\epsilon_t | y_{t-1}] \right] = 0$, and, since $E[y_{t-1}\epsilon_t] = E[\frac{1}{2}(y_t^2 - y_{t-1}^2 - \epsilon_t^2)]$,

$$\begin{aligned} E\left[\sum_{t=1}^{T} y_{t-1}\epsilon_t \middle| y_0 = 0 \right] &= E\left[\frac{1}{2}(y_T^2 - y_0^2) - \frac{1}{2} \sum_{t=1}^{T} \epsilon_t^2 \middle| y_0 = 0 \right] \\ &= E\left[\frac{1}{2} \left(\left(\sum_{t=1}^{T} \epsilon_t \right)^2 - \sum_{t=1}^{T} \epsilon_t^2 \right) \middle| y_0 = 0 \right] \\ &= E\left[\sum_{s \neq t} \epsilon_s \epsilon_t \middle| y_0 = 0 \right] = 0 \ \text{a.s.}, \end{aligned}$$

so that a.s.

$$\begin{aligned} E\left[\left(\sum_{t=1}^{T} y_{t-1}\epsilon_t \right)^2 \middle| y_0 = 0 \right] &= E\left[\left(\sum_{s \neq t} \epsilon_s \epsilon_t \right)^2 \middle| y_0 = 0 \right] \\ &= T(T-1)\sigma_0^4 \\ &= O_p(T^2). \end{aligned}$$

[9] This section uses the Mann-Wald notation: A random variable $w_T = O_p(T^\alpha)$ if, for any $\delta > 0$, there exists $M > 0$ such that $\Pr(|T^{-\alpha} w_T| > M) < \delta$ for all T; $w_T = o_p(T^\alpha)$ if $\Pr(|T^{-\alpha} w_T| > \delta) \to 0$ for every $\delta > 0$, as $T \to \infty$.

This suggests that

$$\text{avar}(\hat{\rho}_T - 1) \sim (\sigma_0^2 T^2)^{-1} T(T-1)\sigma_0^4 (\sigma_0^2 T^2)^{-1} = O(T^{-2}),$$

i.e. that, in the unit root case, $T(\hat{\rho} - 1) = O_p(1)$. This is to be compared to the stationary case, in which $\sqrt{T}(\hat{\rho}_T - \rho_0) = O_p(1)$, with asymptotic distribution $N(0, 1 - \rho_0^2)$. The preceding argument makes clear why the asymptotic variance of this distribution collapses in the unit root case when $\rho \to 1$: In the unit root case, $\hat{\rho}_T$ converges to $\rho_0 = 1$ at rate T^{-1}, i.e. faster than $T^{-\frac{1}{2}}$, the reason being that $\sum_t y_{t-1}^2 = O_p(T^2)$ in the non-stationary case, while $\sum_t y_{t-1}^2 = O_p(T)$ in the stationary case. In the stationary case, $|\rho_0| < 1$,

$$\hat{\rho} - \rho_0 = \left(\frac{1}{T} \sum_{t=1}^{T} y_{t-1}^2 \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} y_{t-1} \epsilon_t,$$

it follows from a LLN that $\frac{1}{T} \sum_{t=1}^{T} y_{t-1}^2 \to E[y_t^2] = \frac{\sigma_0^2}{1-\rho_0^2}$, while $\text{var}(\frac{1}{T} \sum_{t=1}^{T} y_{t-1}\epsilon_t) = \frac{1}{T} \frac{\sigma_0^2}{1-\rho_0^2}$. Therefore, by a CLT,

$$\begin{aligned} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1}\epsilon_t &\xrightarrow{d} N\left(0, \frac{\sigma_0^2}{1-\rho_0^2}\right) \\ \sqrt{T}(\hat{\rho}_T - \rho_0) &\xrightarrow{d} N\left(0, 1-\rho_0^2\right) \\ \text{i.e. } \sqrt{T}(\hat{\rho}_T - \rho_0) &= O_p(1). \end{aligned}$$

An alternative representation of the AR(1) model is

$$\Delta y_t = (\rho_0 - 1) y_{t-1} + \epsilon_t,$$

where the difference operator $\Delta = 1 - L$. Define $\beta_0 = \rho_0 - 1$. Then, (covariance) stationarity implies $\beta_0 < 0$, while non-stationarity (unit root) implies $\beta_0 = 0$. The OLS estimator of β_0 in the re-parameterized linear regression model

$$\Delta y_t = \beta_0 y_{t-1} + \epsilon_t$$

is

$$\hat{\beta}_T = \beta_0 + \left(\sum_t y_{t-1}^2 \right)^{-1} \sum_t y_{t-1}\epsilon_t.$$

The preceding discussion shows that this estimator converges to $\beta_0 < 0$ at rate $\sqrt{T}$ if the process $\{y_t, t \geq 0\}$ is I(0), and it converges to $\beta_0 = 0$ ar rate T if $\{y_t, t \geq 0\}$ is I(1). This is the basis for the Dickey-Fuller unit root test, which tests the null hypothesis of a unit root (equivalent to $\beta_0 = 0$) against the alternative hypothesis of stationarity (equivalent to $\beta_0 < 0$). The Dickey-Fuller test statistics[10] is

$$\mathcal{DF}_T = \frac{\hat{\beta}_T}{\text{se}(\hat{\beta}_T)}.$$

The Dickey-Fuller test statistic has a non-standard (Dickey-Fuller) distribution under the null hypothesis. This distribution depends both on the estimated model and the true data generating process; e.g. the critical value of this test in this model with 5 percent probability of rejecting a true null hypothesis is approximately -2.9, while it would be around -2 for a standard one-sided t-test.

3.2.4 Extensions

(a) Deterministic Trends

Suppose a deterministic trend is included in the previous AR(1) model,

$$y_t = \alpha_0 + \rho_0 y_{t-1} + \gamma_0 t + \epsilon_t,$$

where ϵ_t is white noise, i.e. i.i.d. across t with mean zero and constant variance. The model can be re-parameterized as before, for $\beta_0 = \rho_0 - 1$,

$$\Delta y_t = \alpha_0 + \beta_0(y_{t-1} - \delta_0 t) + \epsilon_t,$$

where $\delta_0 = \gamma_0/(1 - \rho_0)$. If $\beta_0 = 0$, then $\Delta y_t = \alpha_0 + \epsilon_t$, i.e. y_t is I(1), a random walk with drift α_0:

$$y_t = y_0 + \alpha_0 t + \sum_{s=0}^{t-1} \epsilon_{t-s},$$

where y_0 is the initial condition, $\alpha_0 t$ is a deterministic trend component, while $\sum_{s=0}^{t-1} \epsilon_{t-s}$ is a stochastic trend. In this model for the true data generating process, the Dickey-Fuller test is based on the OLS estimator for β in the regression

$$\Delta y_t = \alpha + \beta y_{t-1} + \gamma t + \epsilon_t,$$

[10] Dickey, D.A. and W.A. Fuller (1979): "Distribution of the Estimators for Autoregressive Time Series with a Unit Root", *Journal of the American Statistical Association*, **74**, 427-431.

and the Dickey-Fuller test statistics is, as before, $\mathcal{DF}_T = \frac{\hat{\beta}_T}{\text{se}(\hat{\beta}_T)}$, but the distribution of this test statistic differs from above, because a deterministic trend is included in the regression.

(b) AR(p) with trend

Just as moving average models can be expanded by including further lags, so can autoregressive models. Consider the AR(2) model with trend,

$$y_t = \alpha_0 + \rho_{01} y_{t-1} + \rho_{02} y_{t-2} + \gamma_0 t + \epsilon_t.$$

In this model, the characteristic polynomial in the lag operator is $\Phi(L) = 1 - \rho_{10} L - \rho_{20} L^2$, and stationarity requires that the roots of $|\Phi(z)| = |1 - \rho_{01} z - \rho_{02} z^2| = 0$ lie outside the unit circle. Conversely, the process has a unit root if the characteristic equation permits $z = 1$ as a solution, i.e. if $1 - \rho_{01} - \rho_{02} = 0$. In this case, a re-parametrization suitable for testing the hypothesis of a unit root is

$$\begin{aligned} \Delta y_t &= \alpha_0 + (\rho_{01} + \rho_{02} - 1) y_{t-1} + \rho_{02}(y_{t-2} - y_{t-1}) + \gamma_0 t + \epsilon_t \\ &= \alpha_0 + \beta_0 (y_{t-1} - \delta_0 t) - \rho_{02} \Delta y_{t-1} + \epsilon_t, \end{aligned}$$

where $\beta_0 = \rho_{01} + \rho_{02} - 1$ is zero under the null hypothesis, and $\delta_0 = \gamma_0/(1 - \rho_{01} - \rho_{02})$. Running this regression and testing $H_0 : \ \beta_0 = 0$ yields an Augmented Dickey-Fuller test. Again, the Dickey-Fuller test statistic has a different distribution under the null hypothesis, because of the presence of the lagged differences Δy_{t-1}. Notice that, if the AR(2) process is the true data generating process, but Δy_{t-1} were omitted in the Dickey-Fuller regression, then this omission would induce serial correlation in the estimated residuals: The regression residuals in the mis-specified regression estimate $\rho_{02} \Delta y_{t-1} + \epsilon_t$, and these terms are correlated, because the y_ts are correlated.

All of this generalizes to AR(p) processes, with and without deterministic trend, where p is a positive integer. The relevant re-parametrization of an AR(p), without deterministic trend, becomes

$$\Delta y_t = \alpha_0 + \beta_0 y_{t-1} + \sum_{s=1}^{p-1} \delta_{0s} \Delta y_{t-s} + \epsilon_t,$$

where

$$\begin{aligned} \beta_0 &= \rho_{01} + \cdots + \rho_{0p} - 1 \\ \delta_{0s} &= -(\rho_{0,s+1} + \cdots + \rho_{0p}), \quad \text{for } s = 1, \cdots, p-1. \end{aligned}$$

To see this, define $\rho(L) = \rho_{01} + \cdots + \rho_{0p} - 1$, $\delta(L) = \delta_{01}L + \cdots + \delta_{0,p-1}L^{p-1}$, and notice that

$$\begin{aligned}\Delta y_t &= \alpha_0 + \rho(L)y_t + \epsilon_t \\ &= \alpha_0 + (\beta_0 L + \delta(L)(1-L))y_t + \epsilon_t,\end{aligned}$$

because

$$\begin{aligned}\beta_0 L + \delta(L)(1-L) &= \beta_0 + (\delta_{01}L + \cdots + \delta_{0,p-1}L^{p-1})(1-L) \\ &= \beta_0 L + \delta_{01}L - \delta_{01}L^2 + \delta_{02}L^2 - \delta_{02}L^3 + \cdots + \delta_{0,p-1}L^{p-1} - \delta_{0,p-1}L^p \\ &= (\beta_0 + \delta_{01})L + (\delta_{02} - \delta_{01})L^2 + \cdots + (\delta_{0,p-1} - \delta_{0,p-2})L^{p-1} - \delta_{0,p-1}L^p \\ &= (\rho_{01} - 1)L + \rho_{02}L^2 + \cdots + \rho_{0,p-1}L^{p-1} + \rho_{0p}L^p \\ &= \rho(L).\end{aligned}$$

The Augmented Dickey-Fuller test, as before, examines $H_0: \ \beta_0 = 0$ (unit root) vs. $H_A: \ \beta_0 < 0$ (stationarity).

3.3 Autoregressive Distributed Lag Models

The issues discussed above remain essentially the same when contemporaneous and lagged $\mathbf{x}_t$s are re-introduced. Such models are called autoregressive distributed lag (ARDL) models. The easiest version is an ARDL(1,1), in which $\mathbf{x}_t$ is a scalar covariate which appears next to lagged y_t (the AR(1) part) contemporaneously and with one lag (the DL(1) part),

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t.$$

The implicit assumption in this model is that the process $\{x_t, t \geq 0\}$ is weakly exogenous, i.e. the parameters of its marginal distribution are not linked with the parameters of the conditional distribution of y_t, given x_t and the past.

If $\alpha_1 = 1$, then y_t is I(1). Re-parameterizing,

$$\Delta y_t = \alpha_0 + (\alpha_1 - 1)y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t,$$

this balances LHS and RHS in terms of order of integration if x_t is I(0) and $\alpha_1 = 1$.

If x_t itself also is I(1),

$$\Delta y_t = \alpha_0 + (\alpha_1 - 1)y_{t-1} + \beta_0 \Delta x_t + (\beta_0 + \beta_1)x_{t-1} + \epsilon_t,$$

and in order to balance LHS and RHS in terms of order of integration, either

(i) $\beta_0 + \beta_1 = 0$ and $\alpha_1 = 1$, or

(ii) $\alpha_0 + (\alpha_1 - 1)y_{t-1} + (\beta_0 + \beta_1)x_{t-1}$ is I(0).[11]

Case (i) yields a model in first differences, $\Delta y_t = \alpha_0 + \beta_0 \Delta x_t + \epsilon_t$. Case (ii) is equivalent to

$$\begin{aligned} y_t &= \frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0 + \beta_1}{1-\alpha_1} x_t + \nu_t \\ y^\star &= E[y_t | x_t] = \frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0 + \beta_1}{1-\alpha_1} x_t, \end{aligned}$$

where ν_t is white noise (I(0)). In this case, with both y_t and x_t being I(1) processes (so that Δy_t and Δx_t are I(0)), but a particular linear combination of y_t and x_t, $y_t - \frac{\alpha_0}{1-\alpha_1} - \frac{\beta_0+\beta_1}{1-\alpha_1} x_t$, being I(0), the two stochastic processes are said to be co-integrated. Note that this co-integration relationship has the interpretation of a stable long-run equilibrium relationship between y_t and x_t, i.e. it is implied by the original model if $y_t = y_{t-1}$ and $x_t = x_{t-1}$. This permits the model to be re-cast in its error correction model (ECM) representation

$$\begin{aligned} \Delta y_t &= \alpha_0 + (\alpha_1 - 1)y_{t-1} + \beta_0 \Delta x_t + (\beta_0 + \beta_1)x_{t-1} + \epsilon_t \\ &= (\alpha_1 - 1)\left[y_{t-1} - \frac{\alpha_0}{1-\alpha_1} - \frac{\beta_0 + \beta_1}{1-\alpha_1} x_{t-1} \right] + \beta_0 \Delta x_t + \epsilon_t \\ &= (\alpha_1 - 1)\left[y_{t-1} - y^\star_{t-1} \right] + \beta_0 \Delta x_t + \epsilon_t \\ &= (\alpha_1 - 1)\left[y_{t-1} - y^\star_{t-1} \right] + \frac{\beta_0}{\beta_0 + \beta_1}(1 - \alpha_1)\Delta y^\star_t + \epsilon_t. \end{aligned}$$

This model provides consistent equilibrium dynamics. Note that $\alpha_1 - 1 < 0$ implies that y_t adjusts downwards (upwards) if its previous level y_{t-1} was above (below)

[11] In this case, it must be that $\alpha_1 \neq 1$ ($\beta_0 + \beta_1 \neq 0$), because otherwise this would contradict the hypothesized non-stationarity of x_t (y_t). It also must be the case that $|\alpha_1| < 1$, because otherwise the process is explosive.

its long-run equilibrium level $y^{\star}_{t-1}$, and that it adjusts upwards (downwards) if the long-run equilibrium level increases (decreases).

These re-parameterizations nest other models of interest, by imposing various restrictions on the parameters. The purely static model has $\alpha_1 = \beta_1 = 0$. A model of only partial adjustment has $\beta_1 = 0$. A model in which the two processes y_t and x_t have a so-called common factor (mathematically speaking: share a polynomial in the lag operator, say $1 - \rho L$) takes the form

$$\begin{aligned}(1-\rho L)y_t &= (1-\rho L)(\alpha + \beta x_t) + \epsilon_t \\ \Rightarrow y_t &= \alpha(1-\rho) + \rho y_{t-1} + \beta x_t - \beta\rho x_{t-1} + \epsilon_t,\end{aligned}$$

i.e. $\alpha_0 = \alpha(1-\rho)$, $\alpha_1 = \rho$, $\beta_0 = \beta$, $\beta_1 = -\beta\rho$. Note that this is equivalent to a linear model with AR(1) errors,

$$\begin{aligned}y_t &= \alpha + \beta x_t + \nu_t \\ \nu_t = \rho\nu_{t-1} + \epsilon_t.&\end{aligned}$$

A model with unit long-run coefficient would impose the restriction $\frac{\beta_0+\beta_1}{1-\alpha_1} = 1$. A random walk with drift requires $\alpha_1 = 1$ and $\beta_0 = \beta_1 = 0$.

Different re-parameterizations are of interest because they permit various interpretations of the dynamics of the processes being modelled, e.g. in terms of long-run and short-run dynamics. Moreover, they have important implications for estimation. They determine whether a model that is linear or nonlinear (e.g. ECM) in the parameters is to be estimated. And they ensure that I(0) series are balanced on the LHS and RHS of a regression equation, so that estimators enjoy standard $\sqrt{T}$ convergence properties and conventional regression output retains its validity. To appreciate this latter point, the next section illustrates a case in which failure to recognize the order of integration leads to invalid inference.

3.4 Spurious Regression

Granger and Newbold (1974)[12] and Phillips (1986)[13] were the first to identify the issue of spurious regressions. An example common in applied work, and used here to illustrate the issues involved, might consider the monthly price of a good or service provided by a firm (y_t) as a function of monthly trading volume or sales (x_t).[14] The question of interest is whether a change in industry structure, such as for example the merger of the firm with another firm in the same industry at time T_0, translated into latent synergies that were passed on to consumers in the form of lower prices. Let $\delta_t = 1_{\{t \geq T_0\}}$ denote a binary variable that takes on value 1 after the merger was completed. The proposed model is

$$y_t = \alpha_0 \delta_t + \beta_0 x_t + u_t,$$

and the hypothesis of interest is that $\alpha_0 < 0$, vs. $\alpha_0 = 0$.

Suppose that both y_t and x_t are I(1), satisfying

$$\begin{aligned} E[y_t|y_0] &= y_0, \ \text{var}(y_t|y_0) = t\sigma_y^2 \\ E[x_t|x_0] &= x_0, \ \text{var}(x_t|x_0) = t\sigma_x^2 \\ y_t &\perp x_t, \ \forall t. \end{aligned}$$

The last property, independence of y_t and x_t, implies that $\beta_0 = 0$, and in this case, if the merger has no effect on prices, then $u_t = y_t = y_{t-1} + \epsilon_t$, where ϵ_t is white noise.

Examine the OLS estimator of α_0. By the partitioned regression formula,

$$\begin{aligned} \hat{\alpha}_T &= \alpha_0 + \left(\sum_t \delta_t \left(1 - \frac{x_t^2}{\sum_t x_t^2} \right) \right)^{-1} \left(\sum_t \delta_t \left(1 - \frac{x_t^2}{\sum_t x_t^2} \right) u_t \right) \\ &= \alpha_0 + \left(\sum_t \delta_t - \sum_t \delta_t \frac{x_t^2}{\sum_t x_t^2} \right)^{-1} \left(\sum_t \delta_t u_t - \sum_t \delta_t \frac{x_t^2 u_t}{\sum_t x_t^2} \right) \\ &= \alpha_0 + \left((T - T_0) + \sum_{t \geq T_0} \frac{x_t^2}{\sum_t x_t^2} \right)^{-1} \left(\sum_{t \geq T_0} u_t + \sum_{t \geq T_0} \frac{x_t^2 u_t}{\sum_t x_t^2} \right). \end{aligned}$$

[12] Granger, C.W. and P. Newbold (1974): "Spurious Regression in Econometrics, *Journal of Econometrics*, **2**, 111-120

[13] Phillips, P.C.B. (1986): "Understanding Spurious Regressions in Econometrics", *Journal of Econometrics*, **33**, 311-340

[14] The additional issue of endogeneity of x_t is ignored in the discussion of this section.

The individual components of this expression can be expected to have the following asymptotic properties: With probability one,

$$\begin{aligned}
E[x_t^2|x_0] &= t\sigma_x^2 + x_0^2 \\
E\left[\sum_t x_t^2 \middle| x_0\right] &= \sigma_x^2 \sum_t t + Tx_0^2 \approx \frac{\sigma_x^2}{2}T^2 + Tx_0^2 = O_p(T^2) \\
E\left[\sum_{t\geq T_0} x_t^2 \middle| x_0\right] &\approx \frac{\sigma_x^2}{2}(T^2 - T_0^2) + (T - T_0)x_0^2 = O_p(T^2) \\
E\left[\sum_t u_t \middle| y_0\right] &= E\left[\sum_{t\geq T_0}(y_{t-1} + \epsilon_t) \middle| y_0\right] = (T - T_0 + 1)y_0 = O_p(T) \\
E\left[\sum_{t\geq T_0} x_t^2 u_t \middle| x_0, y_0\right] &= \sum_{t\geq T_0} E[x_t^2|x_0]E[u_t|y_0] \\
&= \sum_{t\geq T_0} E[x_t^2|x_0]E[y_{t-1} + \epsilon_t|y_0] \\
&= y_0\left(\frac{\sigma_x^2}{2}(T^2 - T_0^2) + (T - T_0)x_0^2\right) \\
&= O_p(T^2).
\end{aligned}$$

Hence,

$$\begin{aligned}
\hat{\alpha}_T &= \alpha_0 + \left(O_p(T) - \frac{O_p(T^2)}{O_p(T^2)}\right)^{-1}\left(O_p(T) - \frac{O_p(T^2)}{O_p(T^2)}\right) \\
&= \alpha_0 + O_p(1),
\end{aligned}$$

i.e. $\lim_{T\to\infty} \Pr(|\hat{\alpha}_T - \alpha_0| > \epsilon) > 0$ for any $\epsilon > 0$. In other words, if $\alpha_0 = 0$, then a conventional t-test will erroneously reject this hypothesis with positive probability.

There are two features to note about this. First, non-stationarity of a regressor (x_t) can spill over, in the sense of having an impact on statistical properties of coefficient estimates of other regressors, not just on its own coefficient. Second, if Case (ii) in the preceding section were true, i.e. y_t and x_t were co-integrated, then $\sqrt{T}$ consistency would be preserved; in this case, a linear combination of I(1) variables is stationary (I(0)), and this renders the regression residuals I(0). This also suggests one (single equation based) test for co-integration: First, the individual variables are tested for unit roots; second, if unit roots are not rejected, a linear regression model of one variable onto the others is estimated, and the estimated

regression residuals are tested for a unit root, using an ADF test (again with different critical values). This is the original Engle-Granger procedure[15]. It suffers from inherent problems, however: The assignment of the variables to LHS and RHS is arbitrary, and it implicity assumes weak exogeneity of the RHS variables. The conclusion from this is that all variables should be treated equally and symmetrically, in some sense, i.e. in a system based, multivariate, rather a single equation based, univariate approach.

4 Multivariate Stochastic Processes

4.1 Vector Auto-Regressive Processes

Let $\mathbf{y}_t = (y_{1t}, \cdots, y_{mt})'$ be an $m \times 1$ vector, and consider the vector auto-regression of order p, VAR(p) for positive integers p,

$$\mathbf{y}_t = \mathbf{A}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \epsilon_t,$$

where $\mathbf{A}_i$, $i = 0, \cdots, p$, are $m \times m$ coefficient matrices, and ϵ_t is multivariate white noise, i.e. a vector of serially uncorrelated, mean zero random variables with constant variance-covariance matrix, $E[\epsilon_t] = \mathbf{0}$ and $E[\epsilon_t \epsilon_s'] = \Sigma 1_{\{s=t\}}$ p.d.s. for all s, t. Define the underlying characteristic polynomial in the lag operator by

$$\mathbf{A}(L) = \mathbf{A}_1 L + \cdots + \mathbf{A}_p L^p.$$

Then, the VAR(p) can be written as $(\mathbf{I} - \mathbf{A}(L))\mathbf{y}_t = \mathbf{A}_0 + \epsilon_t$. It is covariance stationary if all the roots of $|\mathbf{I} - \mathbf{A}(z)| = 0$ lie outside the unit circle. Conversely, it is non-stationary (each series has a unit root) if $|\mathbf{I} - \mathbf{A}(1)| = 0$, which is equivalent to

$$\mathbf{I} - \mathbf{A}(1) = \mathbf{I} - \mathbf{A}_1 - \cdots - \mathbf{A}_p = \mathbf{0}.$$

If the process is stationary, its coefficient matrices can be estimated (with $\sqrt{T}$ consistency) using OLS for each equation. The parameters of Σ can be estimated from

[15] Engle, R.F. and C.W.J. Granger (1987): "Co-integration and Error Correction: Representation, Estimation, and Testing", *Econometrica*, **55(2)**, 251-76

the regression residuals $\{\hat{\epsilon}_t, t = p+1, \cdots, T\}$ in the usual way, i.e. the (i, j) element $\hat{\Sigma}_{ij} = \frac{1}{T-p} \sum_{t=p+1}^{T} \hat{\epsilon}_{it}\hat{\epsilon}_{jt}$, for $i, j = 1, \cdots, m$.

A component variable y_j is said to Granger cause another component variable y_i if lagged values of y_j help predicting y_i, i.e. if any of the matrix elements $\mathbf{A}_s^{i,j}$, $s = 1, \cdots, p$, are non-zero[16]. Note that Granger causality does not mean economic causality, only statistical validity as a predictor variable. Often, Granger causality and economic causality run in opposite ways. An example, borrowed from Hamilton (1994)[17], illustrates this: Dividends do not Granger cause stock prices, even though stock prices are the present discounted value of expected future dividends and capital gains; stock prices do Granger cause dividends, however, because they aggregate all the relevant information regarding expected future dividends.

Stationary VAR(p)s have an equivalent MA(∞) representation. Formally,

$$\begin{aligned} \mathbf{y}_t &= (\mathbf{I} - \mathbf{A}(L))^{-1}(\mathbf{A}_0 + \epsilon_t) \\ &= (\mathbf{I} - \mathbf{A}(1))^{-1}\mathbf{A}_0 + \sum_{i=1}^{\infty} \psi_i \epsilon_{t-i}, \end{aligned}$$

where the convention is adopted that $\psi_0 = \mathbf{I}$. The leading constant follows from

$$E[\mathbf{y}_t] = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i E[\mathbf{y}_t] = \mathbf{A}_0 + \mathbf{A}(1)E[\mathbf{y}_t] = (\mathbf{I} - \mathbf{A}(1))^{-1}\mathbf{A}_0.$$

The coefficients $\{\psi_s, s \geq 0\}$ can be determined by matching the polynomials in the lag operator

$$(\mathbf{I} - \mathbf{A}_1 L - \cdots - \mathbf{A}_p L^p)^{-1} = \mathbf{I} + \psi_1 L + \psi_2 L^2 + \cdots,$$

which is equivalent to

$$(\mathbf{I} - \mathbf{A}_1 L - \cdots - \mathbf{A}_p L^p)(\mathbf{I} + \psi_1 L + \psi_2 L^2 + \cdots) = \mathbf{I}.$$

[16]Granger, C.W.J. (1969): "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods", *Econometrica*, **37(3)**, 424-348; also, Sims, C.A. (1972): "Money, Income and Causality", *American Economic Review*, **62(4)**, 540-552

[17]Hamilton, J.D. (1994): *Time Series Analysis*, Princeton: Princeton University Press

Hence, matching coefficients on $L, L^2, \cdots$,

$$\begin{aligned}
-\mathbf{A}_1 + \psi_1 &= \mathbf{0} \;\Rightarrow\; \psi_1 = \mathbf{A}_1 \\
-\mathbf{A}_2 + \psi_2 - \mathbf{A}_1\psi_1 &= 0 \;\Rightarrow\; \psi_2 = \mathbf{A}_1\psi_1 + \mathbf{A}_2 = \mathbf{A}_1^2 + \mathbf{A}_2 \\
&\vdots \\
\text{general result: } \psi_s &= \mathbf{A}_1\psi_{s-1} + \mathbf{A}_2\psi_{s-2} + \cdots + \mathbf{A}_p\psi_{s-p}, \;\; s = 1, 2, \cdots
\end{aligned}$$

An alternative route to determine the sequence of $\{\psi_s, \geq 0\}$ is by recursive substitution. The coefficients in the MA(∞) representation can be interpreted as impulse response function, i.e. as marginal impacts of past innovations, e.g. the (i, j) element of ψ_k is the marginal impact of the innovation $\epsilon_{j,t-k}$ on $y_{i,t}$, $i, j = 1, \cdots, m$, $k = 0, 1, 2, \cdots$. Note, however, that for this interpretation to be meaningful, the components of ϵ_t must be orthogonal to each other.

4.2 Vector Error Correction Representation

In the context of modelling multivariate series and estimation of such models, essentially the same issues arise as in the univariate setting, as discussed above. Hence, in a multivariate context, error correction representations of VAR(p)s, called Vector ECMs (VECMs), are useful for the same reasons given before.

A VAR(p) can be represented as

$$\mathbf{y}_t = \mathbf{A}_0 + \Phi\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta\mathbf{y}_{t-1} + \epsilon_t,$$

which is equivalent to

$$\left[(\mathbf{I} - \Phi L) - \left(\sum_{i=1}^{p-1} \Gamma_i L^i\right)(\mathbf{I} - L)\right]\mathbf{y}_t = \mathbf{A}_0 + \epsilon_t,$$

where

$$\begin{aligned}
\Phi &= \mathbf{A}(1) = \mathbf{A}_1 + \cdots \mathbf{A}_p \\
\Gamma_i &= -\left[\mathbf{A}_{i+1} + \cdots + \mathbf{A}_p\right], \;\; i = 1, 2, \cdots, p-1.
\end{aligned}$$

To see this, note that

$$\begin{aligned}
& (\mathbf{I} - \Phi L) - \left(\sum_{i=1}^{p-1} \Gamma_i L^i \right) (\mathbf{I} - L) \\
= \; & \mathbf{I} - \Phi L - \Gamma_1 L + \Gamma_1 L^2 - \Gamma_2 L^2 + \Gamma_2 L^3 - \cdots - \Gamma_{p-1} L^{p-1} + \Gamma_{p-1} L^p \\
= \; & \mathbf{I} - (\Phi + \Gamma_1) L - (\Gamma_2 - \Gamma_1) L^2 - \cdots - (\Gamma_{p-1} - \Gamma_{p-2}) L^{p-1} + \Gamma_{p-1} L^p \\
= \; & \mathbf{I} - \mathbf{A}_1 L - \cdots \mathbf{A}_p L^p.
\end{aligned}$$

An equivalent representation is the VECM

$$\Delta \mathbf{y}_t = \mathbf{A}_0 + (\Phi - \mathbf{I}) \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta \mathbf{y}_{t-1} + \epsilon_t.$$

This is referred to as the Sims, Stock and Watson (1990) canonical representation, originally due to Fuller (1976)[18]. Notice that this is yet again simply a reparametrization, and there exists a one-to-one mapping between the coefficient matrices of the VAR and the VECM. The VECM can be estimated by OLS, and VAR coefficients can be determined via the above formulae.

Just as the coefficient on lagged y_t in a univariate ECM plays a critical rôle in determining the integration properties of the series being modelled, so does the coefficient matrix $\Pi = \Phi - \mathbf{I} = \mathbf{A}(1) - \mathbf{I}$ in the multivariate case. From the definition of covariance stationary in the multivariate context, it follows that the VAR(p) is non-stationary if, and only if, $z = 1$ is a solution to the determinatal equation $|\mathbf{I} - \mathbf{A}(z)| = 0$.

It is straightforward to see that Π having full rank m corresponds to the other extreme case of $\mathbf{y}_t$ being I(0). Suppose, to the contrary, that $\text{rk}(\Pi) < m$. Then, there exists a vector $\alpha \in \mathbb{R}^m$ such that $\alpha' \Pi = \mathbf{0}$. Consider, for simplicity, a VAR(1) for which $\Pi = \mathbf{A}_1 - \mathbf{I}$. Then,

$$\begin{aligned}
\Delta \mathbf{y}_t &= \mathbf{A}_0 + \Pi \mathbf{y}_{t-1} + \epsilon_t \\
\Rightarrow \alpha' \Delta \mathbf{y}_t &= \alpha' \mathbf{A}_0 + \alpha' \Pi \mathbf{y}_{t-1} + \alpha' \epsilon_t \\
\Rightarrow \alpha' \Delta \mathbf{y}_t &= \alpha' \mathbf{A}_0 + \alpha' \epsilon_t,
\end{aligned}$$

[18] Sims, C.A., Stock, J.H. and M.W. Watson (1990): "Inference in Linear Time Series Models with Some Unit Roots", **Econometrica**, **58(1)**, 123-144; Fuller, W.A. (1976): *Introduction to Statistical Time Series*, New York: Wiley

i.e. $\alpha'\mathbf{y}_t$ is I(1), a contradiction to the hypothesis that $\mathbf{y}_t$ is I(0). Hence, full rank of Π is equivalent to all components of $\mathbf{y}_t$ being covariance stationary.

Noting that each equation in a VECM looks just like an univariate ARDL model in which x_t represents another component of the vector $\mathbf{y}_t$, one might expect the matrix Π to be informative about co-integrating relationships as well, because $\Pi\mathbf{y}_{t-1}$ is just a collection of m linear combinations of the elements of $\mathbf{y}_{t-1}$. In order to then balance the order of integration of the LHS and RHS, it must be the case that Π, in a sense that will be made precise below, contains all coefficients of co-integrating relationships among the elements of $\mathbf{y}_t$, i.e. all co-integrating vectors that induce linear combinations of the elements of $\mathbf{y}_t$ which are I(0). It follows from the preceding two paragraphs that the case of co-integration among the component series of $\mathbf{y}_t$ corresponds to $0 < \text{rk}(\Pi) = r < m$. In this case, it is said that there exist r distinct co-integrating relationships between the m elements of $\mathbf{y}_t$, each corresponding to a co-integrating vector β_j so that $\beta_j'\mathbf{y}_t$ is I(0), $j = 1, \cdots, r$. In terms of the solutions to the determinantal equation, the case of r co-integration relationships between the m elements of $\mathbf{y}_t$ is equivalent to $m-r$ solutions (out of mp solutions of $|\mathbf{I}-\mathbf{A}(z)| = 0$) that lie on the unit circle, with a real part equal to unity, while all other solutions lie outside the unit circle and correspond to the co-integrating relationships and higher-order dynamics.

The foregoing discussion is summarized in the Granger Representation Theorem: *Consider the vector-valued process* $\{\mathbf{y}_t, t \geq 0\}$ *of dimension* m, *satisfying*

$$\mathbf{y}_t = \mathbf{A}(L)\mathbf{y}_t + \epsilon_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-1} + \epsilon_t,$$

where ϵ_t *is multivariate white noise. Suppose there exist* r *co-integrating relationships among the* m *elements of* $\mathbf{y}_t$. *Then,*

(1) *there exists an* $m \times r$ *matrix* β *with* $\text{rk}(\beta) = r$, *such that* $\mathbf{z}_t = \beta'\mathbf{y}_t$ *is a system of* r *I(0) series;*

(2) $\Delta\boldsymbol{y}_t$ *has an MA(*∞*) representation:* $\Delta\mathbf{y}_t = \psi(L)\epsilon_t$, *where* $\psi(L) = \mathbf{I} + \sum_{i=1}^{\infty} \psi_i L^i$, *and* $\beta'\psi(1) = \mathbf{0}$;

(3) $\Pi = \mathbf{A}(1) - \mathbf{I}$ *has* $\mathrm{rk}(\Pi) = r$, *and there exists an* $m \times r$ *matrix* α, *such that* $\Pi = \alpha\beta'$;

(4) *there exists a VECM:* $\Delta\mathbf{y}_t = \alpha\mathbf{z}_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta\mathbf{y}_{t-i} + \epsilon_t$.

The last assertion of part (2) is not critical for the understanding of the further development; its proof is given in an appendix.

If some of the series in the VAR are subject to a deterministic time trend - which, if present, in the case of economic series is typically linear - then it can be included into the co-integrated relationship, in analogy to Section 3.2.4 above.[19] Formally, in terms of the formalism of the preceding Theorem, if the original VAR(p) is of the form

$$\mathbf{y}_t = \mathbf{A}(L)\mathbf{y}_t + \gamma t + \epsilon_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-1} + \gamma t + \epsilon_t,$$

where γ is a $m \times 1$ vector of coefficients on the time variable t, then the associated VECM is

$$\Delta\mathbf{y}_t = \alpha(\mathbf{z}_{t-1} + \delta t) + \sum_{i=1}^{p-1} \gamma_i \Delta\mathbf{y}_{t-i} + \epsilon_t,$$

where the $r \times 1$ vector δ satisfies $\alpha\delta = \gamma$.

It is important to note that α and β are not uniquely determined, since for any non-singular $r \times r$ matrix Q, $\Pi = \alpha\beta' = \alpha Q Q^{-1}\beta' = \tilde{\alpha}\tilde{\beta}'$, where $\tilde{\alpha} = \alpha Q$ and $\tilde{\beta} = \beta(Q^{-1})'$. The same argument applies to δ. The appropriate choice of Q is usually guided by economic theory and equivalent to imposing r^2 restrictions on the elements of Q.

[19] If it were included without being restricted to be part of the co-integrating relationship, then this might imply a quadratic trend in the respective original series.

4.3 Johansen Co-integration Tests

Johansen co-integration tests[20] present a formal statistical framework to test hypotheses about the rank of the matrix Π in the VECM representation of a VAR(p), which, as shown above, relate to the integration properties of the multivariate stochastic process $\mathbf{y}_t$. Testing can thereby take various forms. For instance,

(I) $H_0: \ \mathrm{rk}(\Pi) = 0$, vs. $H_A: \ \mathrm{rk}(\Pi) > 0$.

(II) $H_0: \ \mathrm{rk}(\Pi) = 0$, vs. $H_A: \ \mathrm{rk}(\Pi) = 1$.

(III) $H_0: \ \mathrm{rk}(\Pi) = r$, vs. $H_A: \ \mathrm{rk}(\Pi) > r$.

(IV) $H_0: \ \mathrm{rk}(\Pi) = r$, vs. $H_A: \ \mathrm{rk}(\Pi) = r + 1$.

Cases (I) and (II) are considered here in turn.[21] As in the case of testing for unit roots in the case of univariate stochastic processes, there are further test variants when deterministic trends are included in the model.

4.3.1 Case (I)

Consider the model $\Delta\mathbf{y}_t = \Pi\mathbf{y}_{t-1} + \mathbf{v}_t$; here, the intercept vector and the lagged differences $\Delta\mathbf{y}_{t-s}$, $s = 1, \cdots, p_1$ are omitted, as they are irrelevant to the understanding of the underlying principles of the test procedure. Stack up the T systems $\Delta\mathbf{y}_t' = \mathbf{y}_{t-1}'\Pi' + \mathbf{v}_t'$, to form

$$\Delta\mathbf{Y} = \mathbf{Y}_{-1}\Pi' + \mathbf{v},$$

where $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_T)'$, $\mathbf{Y}_{-1} = (\mathbf{y}_0, \cdots, \mathbf{y}_{T-1})'$ and $\mathbf{v} = (\mathbf{v}_1, \cdots, \mathbf{v}_T)'$ are $T \times m$ matrices. If the rows of $\mathbf{v}$ are assumed to be normally distributed, with mean zero,

[20] Johansen, S. (1988): "Statistical Analysis of Co-integration Vectors", *Journal of Economic Dynamics and Control*, **12**, 231-254; and Johansen, S. (1991): "Estimation and Hypothesis Testing of Co-integration Vectors in Gaussian Vector Autoregressive Models", *Econometrica*, **59(6)**, 1551-1580

[21] The organization and presentation of Johansen tests provided in this section builds on Tom Rothenberg's exposition of this material in a graduate time-series course at U.C. Berkeley. I am indebted to him for his lucid introduction to this topic. All errors are mine.

contemporaneous variance-covariance matrix Ω and serially independent, then the joint probability density of this model, or the likelihood function of the parameters Π and Ω, given the data, is

$$\begin{aligned}
\prod_{t=1}^{T} f(\mathbf{v}_t;\Pi,\Omega) &\propto |\Omega|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_t \mathbf{v}_t'\Omega^{-1}\mathbf{v}_t\right) \\
&= |\Omega|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\sum_t \mathbf{v}_t'\Omega^{-1}\mathbf{v}_t\right)\right) \\
&= |\Omega|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_t \mathrm{tr}\left(\mathbf{v}_t'\Omega^{-1}\mathbf{v}_t\right)\right) \\
&= |\Omega|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_t \mathrm{tr}\left(\Omega^{-1}\mathbf{v}_t\mathbf{v}_t'\right)\right) \\
&= |\Omega|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Omega^{-1}\sum_t \mathbf{v}_t\mathbf{v}_t'\right)\right) \\
&= |\Omega|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Omega^{-1}\mathbf{V}'\mathbf{V}\right)\right),
\end{aligned}$$

where $\mathbf{V} = \Delta\mathbf{Y} - \mathbf{Y}_{-1}\Pi'$.[22] Given Π, Ω can be concentrated out in the usual way, i.e. by choosing $\Omega = \frac{1}{T}\mathbf{V}'\mathbf{V}$.[23] Then, the concentrated likelihood function is

$$\begin{aligned}
\prod_{t=1}^{T} f(\mathbf{v}_t;\Pi) &\propto \left|\frac{1}{T}\mathbf{V}'\mathbf{V}\right|^{-\frac{T}{2}} \exp\left(\frac{1}{2}\mathrm{tr}\left(\left(\frac{\mathbf{V}'\mathbf{V}}{T}\right)^{-1}\mathbf{V}'\mathbf{V}\right)\right) \\
&\propto \left|\frac{1}{T}\mathbf{V}'\mathbf{V}\right|^{-\frac{T}{2}} \\
&= \left|\frac{1}{T}\left(\mathbf{Y}-\mathbf{Y}_{-1}\Pi'\right)'\left(\mathbf{Y}-\mathbf{Y}_{-1}\Pi'\right)\right|^{-\frac{T}{2}} \\
&\to \max_{\Pi} \\
\Leftrightarrow \qquad & \min_{\Pi} \frac{T}{2}\ln\left(\left|\left(\mathbf{Y}-\mathbf{Y}_{-1}\Pi'\right)'\left(\mathbf{Y}-\mathbf{Y}_{-1}\Pi'\right)\right|\right).
\end{aligned}$$

Imposing the null hypothesis of Case (I), i.e. the m^2 restrictions $\Pi = \mathbf{0}$, yields $\frac{T}{2}\ln\left(|\Delta\mathbf{Y}'\Delta\mathbf{Y}|\right)$, which is proportional to the log-likelihood function under the null hypothesis.

[22] Strictly speaking, the preceding expression is the conditional density of $\mathbf{Y}$, given $\mathbf{y}_0$.

[23] Appendix B.3 is a brief review of concentrating out parameters from a likelihood function.

Under the alternative hypothesis, the unrestricted estimator of Π is the OLS estimator (on each equation), and the log-likelihood function, evaluated at the estimator, is proportional to the logarithm of the residual sum of squares, i.e. proportional to $\frac{T}{2}\ln\left(\left|\Delta\mathbf{Y}'M_{\mathbf{Y}_{-1}}\Delta\mathbf{Y}\right|\right)$, where the $T\times T$ matrix $M_{\mathbf{Y}_{-1}}=\mathbf{I}-\mathbf{Y}_{-1}(\mathbf{Y}'_{-1}\mathbf{Y}_{-1})^{-1}\mathbf{Y}'_{-1}$ is the orthogonal projector onto the space orthogonal to the column space of $\mathbf{Y}_{-1}$.

The likelihood ratio test statistic for Case (I) is then, as usual, twice the difference between the log-likelihood of the unrestricted and restricted model, i.e.

$$\mathcal{LR}_T=-T\ln\left(\left|\frac{\Delta\mathbf{Y}'M_{\mathbf{Y}_{-1}}\Delta\mathbf{Y}}{\Delta\mathbf{Y}'\Delta\mathbf{Y}}\right|\right)\quad\sim\quad\chi^2_{m^2},$$

and the null hypothesis is rejected when this statistic exceeds the critical value of a $\chi^2_{m^2}$ distribution for the appropriate size of the test.

The usual representation of the Johansen test is in terms of certain eigenvalues. To deduce this, notice that

$$\begin{aligned}
\mathcal{LR}_T &= -T\ln\left(\left|\Delta\mathbf{Y}'\Delta\mathbf{Y}\right|^{-1}\left|\Delta\mathbf{Y}'M_{\mathbf{Y}_{-1}}\Delta\mathbf{Y}\right|\right)\\
&= -T\ln\left(\left|\Delta\mathbf{Y}'\Delta\mathbf{Y}\right|^{-1}\left|\Delta\mathbf{Y}'\Delta\mathbf{Y}-\Delta\mathbf{Y}'\mathbf{Y}_{-1}(\mathbf{Y}'_{-1}\mathbf{Y}_{-1})^{-1}\mathbf{Y}'_{-1}\Delta\mathbf{Y}\right|\right)\\
&= -T\ln\left(\left|\mathbf{I}-(\Delta\mathbf{Y}'\Delta\mathbf{Y})^{-\frac{1}{2}}\Delta\mathbf{Y}'\mathbf{Y}_{-1}(\mathbf{Y}'_{-1}\mathbf{Y}_{-1})^{-1}\mathbf{Y}'_{-1}\Delta\mathbf{Y}\,(\Delta\mathbf{Y}'\Delta\mathbf{Y})^{-\frac{1}{2}}\right|\right)\\
&= -T\ln\left(\prod_{i=1}^{m}\mu_i\right)=-T\sum_{i=1}^{m}\ln(\mu_i),
\end{aligned}$$

where the third (fourth) equality follows from a linear algebra result provided in Appendix B.1.1 (B.1.2), and $\{\mu_i, i=1,\cdots,m\}$ are the characteristic roots (eigenvalues) of the matrix

$$Q=\mathbf{I}-(\Delta\mathbf{Y}'\Delta\mathbf{Y})^{-\frac{1}{2}}\Delta\mathbf{Y}'\mathbf{Y}_{-1}(\mathbf{Y}'_{-1}\mathbf{Y}_{-1})^{-1}\mathbf{Y}'_{-1}\Delta\mathbf{Y}\,(\Delta\mathbf{Y}'\Delta\mathbf{Y})^{-\frac{1}{2}}.$$

The representation of the log-likelihood test statistic in terms of eigenvalues is usually referred to as Johansen trace statistic for Case (I).

4.3.2 Case (II)

In this case, since the null hypothesis is the same as in Case (I), the denominator of the test statistics (the log-likelihood function under the null hypothesis) remains

the same as before. The numerator is proportional to the logarithm of the sum of squared residuals when the restriction $\Pi' = \alpha\beta'$ is imposed, where $\alpha, \beta \in \mathbb{R}^m$, and the model is

$$\Delta\mathbf{Y} = \mathbf{Y}_{-1}\Pi' + \mathbf{v} = \mathbf{Y}_{-1}\alpha\beta' + \mathbf{v}.$$

Hence, the log-likelihood function under the alternative hypothesis is proportional to $\frac{T}{2}\ln\left(\left|(\Delta\mathbf{Y} - \mathbf{Y}_{-1}\alpha\beta')'(\Delta\mathbf{Y} - \mathbf{Y}_{-1}\alpha\beta')\right|\right)$, which is to be minimized with respect to β, given α, i.e. concentrating out β, and subsequently with respect to α.

Let $\mathbf{z} = \mathbf{Y}_{-1}\alpha$, which is stationary under the alternative hypothesis, with coefficient vector β; i.e. α is the single co-integrating vector under the alternative hypothesis. Cast in this form, the model under the alternative hypothesis amounts to m LHS variables collected in $\Delta\mathbf{y}_t$ and a single RHS variable z_t, which enters each equation with an individual coefficient β_i, $i = 1, \cdots, m$:

$$\Delta y_{i,t} = \beta_i z_t + v_{i,t}, \quad i = 1, \cdots, m; t = 1, \cdots, T.$$

The coefficients β_i can be estimated by individual OLS regressions. Consequently, and analogously to Case (I), the log-likelihood function under the alternative hypothesis is proportional to $\frac{T}{2}\ln\left(\left|\Delta\mathbf{Y}' M_{\mathbf{Y}_{-1}\alpha}\Delta\mathbf{Y}\right|\right)$, where $M_{\mathbf{Y}_{-1}\alpha} = M_{\mathbf{z}} = I - \frac{\mathbf{z}\mathbf{z}'}{\mathbf{z}'\mathbf{z}}$. Using the result provided in Appendix B.2,

$$\frac{T}{2}\ln\left(\left|\Delta\mathbf{Y}' M_{\mathbf{Y}_{-1}\alpha}\Delta\mathbf{Y}\right|\right) = \frac{T}{2}\left(\frac{\left|\alpha'\mathbf{Y}'_{-1}M_{\Delta\mathbf{Y}}\mathbf{Y}_{-1}\alpha\right|\left|\Delta\mathbf{Y}'\Delta\mathbf{Y}\right|}{\left|\alpha'\mathbf{Y}'_{-1}\mathbf{Y}_{-1}\alpha\right|}\right) \to \min_{\alpha}.$$

This is a ratio of quadratics in α, i.e. of the form $\frac{T}{2}\ln\left(\frac{\alpha' A\alpha}{\alpha' B\alpha}\right)$, where $A = \mathbf{Y}'_{-1}M_{\Delta\mathbf{Y}}\mathbf{Y}_{-1}$ and $B = \mathbf{Y}'_{-1}\mathbf{Y}_{-1}$, which is p.d.s. The FOCs of this minimization problem yield

$$\begin{aligned}
\mathbf{0} &= (\hat{\alpha}' B\hat{\alpha})^{-2}(\hat{\alpha}' B\hat{\alpha} 2A\hat{\alpha} - \hat{\alpha}' A\hat{\alpha} 2B\hat{\alpha}) \\
\Rightarrow \mathbf{0} &= \left(A - \frac{\hat{\alpha}' A\hat{\alpha}}{\hat{\alpha}' B\hat{\alpha}}B\right)\hat{\alpha} \\
&= (A - \hat{r}B)\hat{\alpha} \\
\Leftrightarrow \mathbf{0} &= \left(B^{-\frac{1}{2}}AB^{-\frac{1}{2}} - \hat{r}\mathbf{I}\right)\hat{\gamma},
\end{aligned}$$

where $\hat{r}$ are the characteristic roots (eigenvalues) of $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$ and $\hat{\gamma} = B^{\frac{1}{2}}\hat{\alpha}$. There are m pairs of eigenvalues $\hat{r}$ and associated eigenvectors $\hat{\alpha}$. Minimization with respect to $\hat{\alpha}$ leads to choosing the smallest eigenvalue, $\hat{r}_{\min}$. Hence, the log-likelihood

under the alternative hypothesis is proportional to $\frac{T}{2}\ln\left(|\Delta\mathbf{Y}'\Delta\mathbf{Y}|\,\hat{r}_{\min}\right)$, so that the Johansen likelihood ratio test statistic for Case (II) is

$$\mathcal{LR}_T = -T\ln\left(\hat{r}_{\min}\right).$$

Note that, as a consequence of the result provided in Appendix B.1.3,

$$B^{-\frac{1}{2}}AB^{-\frac{1}{2}} = \left(\mathbf{Y}'_{-1}\mathbf{Y}_{-1}\right)^{-\frac{1}{2}}\mathbf{Y}_{-1}M_{\Delta\mathbf{Y}}\mathbf{Y}_{-1}\left(\mathbf{Y}'_{-1}\mathbf{Y}_{-1}\right)^{-\frac{1}{2}}$$

has the same eigenvalues as

$$\left(\Delta\mathbf{Y}'\Delta\mathbf{Y}\right)^{-\frac{1}{2}}\Delta\mathbf{Y}'M_{\mathbf{Y}_{-1}}\Delta\mathbf{Y}\left(\Delta\mathbf{Y}'\Delta\mathbf{Y}\right)^{-\frac{1}{2}},$$

and these are $1-\mu_i$, $i=1,\cdots,m$ (see Appendix B.1.4), where the μ_i are the eigenvalues of the matrix Q encountered in Case (I). Hence, the Johansen likelihood ratio test statistic can also be expressed as

$$\mathcal{LR}_T = -T\ln\left(1-\mu_{\max}\right).$$

4.3.3 Further Results

Using the same principles as in the preceding two subsections, the Johansen likelihood ratio test statistics for the remaining two test cases can be deduced. For Case (III), $H_0:\ \mathrm{rk}(\Pi)=r$ against $H_A:\ \mathrm{rk}(\Pi)>r$, the test statistic is

$$\mathcal{LR}_T = -T\sum_{i=r+1}^{m}\ln\left(\mu_{(i)}\right),$$

where $\mu_{(1)}<\cdots<\mu_{(m)}$ are the ordered eigenvalues of the matrix Q obtained in Case (I). Similarly, for Case (IV), $H_0:\ \mathrm{rk}(\Pi)=r$ against $H_A:\ \mathrm{rk}(\Pi)=r+1$,

$$\mathcal{LR}_T = -T\ln\left(1-\mu_{(m-r)}\right) = -T\ln\left(\hat{r}_{(r+1)}\right),$$

where $\hat{r}_{(1)}<\cdots<\hat{r}_{(m)}$ are the ordered eigenvalue of $\mathbf{I}-Q$. The critical values depend on m and r and are provided in tables or by statistical software.

5 Supplement: Time Series Models of Heteroskedasticity

5.1 Basic Concepts

Up to this point, it was assumed that the stochastic processes being modelled are propelled by innovations that have constant variances and covariances over time. This assumption impedes the analysis of potential volatility in the series, i.e. changing or heteroskedastic variances (and covariances) over time. Time series models of heteroskedasticity have important applications as a useful tool to capture the volatility of a stochastic process, notably in empirical finance. Recent experience in financial markets shows that - beyond the theory of efficient financial markets which predicts no autocorrelation in asset returns - squared returns vary widely and, to some extent, predictably depend on the past. This suggests that conditional variances may follow a time series process as well, and sometimes this process may be characterized by a distribution with thick tails.

For the purpose of illustration, consider the univariate stationary AR(p) process $y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + u_t$, where u_t is assumed to be white noise, i.e. u_t is i.i.d. with $\mathbb{E}[u_t] = 0$ and $\mathbb{E}[u_t u_s] = \sigma^2 1_{\{t=s\}}$, $\sigma^2 > 0$. The white noise assumption implies that the process' unconditional variance is constant. This does not preclude that the conditional variance may vary over time. One way to model this is as a stationary AR(m) for $\{u_t^2, t = 1, \cdots\}$:

$$u_t^2 = \xi + \sum_{j=1}^{m} \alpha_j u_{t-j}^2 + \omega_t,$$

where ω_t is white noise, i.e. ω_t is i.i.d. with $\mathbb{E}[\omega_t] = 0$ and $\mathbb{E}[\omega_t \omega_s] = \lambda^2 1_{\{t=s\}}$, $\lambda^2 > 0$, for all t. Since $E[u_t | u_{t-s}, s = 1, 2, \cdots] = 0$ this implies for the conditional variance of u_t, given the past,

$$\mathbb{E}[u_t^2 | u_{t-s}^2, s = 1, 2, \cdots] = \xi + \sum_{j=1}^{m} \alpha_j u_{t-j}^2.$$

This AR(m) model for u_t^2 is called Autoregressive Conditional Heteroskedasticity

(ARCH) model (Engle (1982)[24]).

This model requires further restrictions in order to be an adequate representation of volatility and to be compatible with the stationary AR model for the primary series of interest, y_t. (i) To ensure that the conditional variances are positive, it is required that $\alpha_j \geq 0$, $j = 1 \cdots, m$, and $\xi > 0$. (ii) To ensure that u_t^2 is covariance stationary, it is required that $|1 - \alpha(z)| = |1 - \sum_{j=1}^{m} \alpha_j z^j| = 0$ have all roots outside the unit circle. Provided these conditions hold, the unconditional variance of u_t can be expressed in terms of the ARCH model parameters as

$$\sigma^2 = \xi/(1 - \alpha(1)).$$

Further restrictions are required if the model is designed to eliminate thick tails, i.e. to control higher-order moments. To see this, consider the alternative representation of the innovations $u_t = \sqrt{h_t} v_t$, $h_t = \xi + \sum_{j=1}^{m} \alpha_j u_{t-j}^2$, so that $v_t = \frac{u_t}{\sqrt{h_t}}$ have the interpretation of standardized innovations of the primary process y_t, satisfying

$$\begin{aligned} \mathbb{E}[v_t] &= \mathbb{E}\left[\mathbb{E}[v_t | u_{t-s}, s = 1, \cdots]\right] = 0 \\ \text{var}(v_t | u_{t-s}, s = 1, \cdots) &= \frac{1}{h_t} \text{var}(u_t | u_{t-s}, s = 1, \cdots) = \frac{1}{h_t}(h_t + \lambda^2) \\ \text{var}(v_t) &= \mathrm{E}[u_t^2/h_t] = \mathbb{E}\left[\mathbb{E}[u_t^2 | u_{t-s}, s = 1, \cdots]/h_t\right] = 1. \end{aligned}$$

The thickness of the tails of the distribution of v_t is governed by its fourth moment, $\mathbb{E}[(v_t^2 - 1)^2]$. Since $u_t^2 = h_t v_t^2 = h_t + \omega_t$, it follows that $\omega_t = h_t(v_t^2 - 1)$, so that $\mathbb{E}[\omega_t^2] = \lambda^2 = \mathbb{E}[h_t^2 (v_t^2 - 1)^2] = \mathbb{E}[h_t^2]\mathbb{E}[(v_t^2 - 1)^2]$, because v_t is independent. Consider, for simplicity, the case of an ARCH(1) model, for which $h_t = \xi + \alpha_1 u_{t-1}^2$. Then, the

[24]Engle, R.F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation", *Econometrica*, **50(4)**, 987-1009.

unconditional expectation of h_t^2 is

$$\begin{aligned}
\mathbb{E}[h_t^2] &= \mathbb{E}\left[(\xi + \alpha_1 u_{t-1}^2)^2\right] \\
&= \xi^2 + \alpha_1^2 \mathbb{E}[u_{t-1}^4] + 2\alpha_1 \mathbb{E}[u_{t-1}^2] \\
&= \xi^2 + \alpha_1^2 \left(\text{var}(u_{t-1}^2) + (\mathbb{E}[u_{t-1}^2])^2\right) + 2\alpha_1 \xi \frac{\xi}{1-\alpha_1} \\
&= \xi^2 + \alpha_1^2 \left(\frac{\lambda^2}{1-\alpha_1^2} + \frac{\xi^2}{(1-\alpha_1)^2}\right) + 2\alpha_1 \xi \frac{\xi}{1-\alpha_1} \\
&= \frac{\alpha_1^2 \lambda^2}{1-\alpha_1^2} + \frac{(1-\alpha_1)^2 \xi^2 + \alpha_1^2 \xi^2 + 2(1-\alpha_1)\alpha_1 \xi^2}{(1-\alpha_1)^2} \\
&= \frac{\xi^2}{(1-\alpha_1)^2} + \frac{\alpha_1^2 \lambda^2}{1-\alpha_1^2}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[(v_t^2 - 1)^2\right] = \frac{\lambda^2}{\frac{\xi^2}{(1-\alpha_1)^2} + \frac{\alpha_1^2 \lambda^2}{1-\alpha_1^2}}.$$

Suppose the assumptions are slightly strengthened, so that the standardized innovations $v_t = \frac{u_t}{\sqrt{h_t}}$ have a distribution whose tails are not thick, say they be distributed $N(0,1)$. Then, the fourth moment satisfies $\mathbb{E}\left[(v_t^2 - 1)^2\right] = 2$. The above expression for the fourth moment then implies

$$\frac{\lambda^2(1 - 3\alpha_1^2)}{1-\alpha_1^2} = \frac{2\xi^2}{(1-\alpha_1)^2}.$$

The right-hand side is positive. Therefore, for the left-hand side to be positive, it is required that $\alpha_1 \geq 1/\sqrt{3}$.

Empirically, for financial time series, such restrictions on the tails of their distributions are typically rejected. Researchers, therefore, often maintain distributional assumptions that allow for thicker tails, e.g. t-distribution instead of normality.

5.2 Estimation of ARCH(m) Model

Assuming v_t is Gaussian, then estimation can proceed by Maximum Likelihood methods. The likelihood function is thereby set up recursively.

Let $\mathcal{Y}_t = (y_t, y_{t-1}, \cdots)$. Then, the conditional density of y_t, given the past, is

$$f(y_t|\mathcal{Y}_{t-1};\theta) = \frac{1}{\sqrt{2\pi h_t}} \exp\left(-\frac{1}{2h_t}((1-\phi(L))y_t - c)^2\right)$$

where $\theta' = (c, \phi_1, \cdots, \phi_p, \alpha_1, \cdots, \alpha_m, \xi)$. The log-likelihood function $l(\theta; y_{-m+1}, \cdots, y_T)$ can then be expressed as

$$\begin{aligned} l(\theta; y_{-m+1}, \cdots, y_T) &= \ln\left(f(y_1, \cdots, y_T|y_0, \cdots, y_{-m+1};\theta)\right) \\ &= \sum_{t=1}^{T} \ln\left(f(y_t|\mathcal{Y}_{t-1};\theta)\right) \\ &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln(h_t) - \frac{1}{2}\sum_{t=1}^{T}\frac{((1-\phi(L))y_t - c)^2}{h_t} \\ &\to \max_{\theta}! \end{aligned}$$

5.3 Extensions

5.3.1 Generalized ARCH (GARCH)

Consider the model $h_t = \xi + \pi(L)u_t^2$, where $\pi(L) = \sum_{j=1}^{\infty}\pi_j L^j$ is an infinite polynomial in the lag operator L and the u_t are white noise, as above. Parameterize $\pi(L)$ as the ratio of two finite order polynomials in L:

$$\pi(L) = \frac{\alpha(L)}{1-\delta(L)},$$

where

$$\begin{aligned} \alpha(L) &= \sum_{j=1}^{m}\alpha_j L^j \\ \delta(L) &= \sum_{k=1}^{r}\delta_k L^k, \end{aligned}$$

where it is assumed that $|1-\delta(z)| = 0$ has all roots outside the unit circle.

This yields

$$h_t = \xi + \frac{\alpha(L)}{1-\delta(L)}u_t^2,$$

from which it follows that

$$(1-\delta(L))h_t = (1-\delta(1))\xi + \alpha(L)u_t^2,$$

which is equivalent to

$$h_t = (1-\delta(1))\xi + \delta_1 h_{t-1} + \cdots + \delta_r h_{t-r} + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2,$$

i.e. h_t follows an ARMA(r,m) process. This model is referred to as Generalized ARCH (GARCH; Bollerslev (1986)[25]). Estimation proceeds my maximum likelihood, analogous to the case of ARCH.

5.3.2 Integrated GARCH (IGARCH)

Consider the following GARCH model. Suppose $(1-\delta(L))h_t = \xi + \alpha(L)u_t^2$, so that

$$h_t = \xi + \sum_{i=1}^{r} \delta_i h_{t-i} + \sum_{j=1}^{m} \alpha_j u_{t-j}^2.$$

Then,

$$h_t + u_t^2 = \xi - \delta_1(u_{t-1}^2 - h_{t-1}) - \cdots - \delta_r(u_{t-r}^2 - h_{t-r}) + \sum_{i=1}^{r} \delta_i u_{t-i}^2 + \sum_{j=1}^{m} \alpha_j u_{t-j}^2 + u_t^2.$$

Defining the martingale difference sequences $\omega_t = u_t^2 - h_t$ which satisfies $\mathbb{E}[\omega_t|\text{past}] = 0$, and $p = \max\{r,m\}$, this model is equivalent to

$$u_t^2 = \xi + \sum_{s=1}^{p}(\delta_s + \alpha_s)u_{t-s}^2 + \omega_t - \sum_{k=1}^{r}\delta_k \omega_{t-k},$$

where $\delta_s = 0$ for $s > r$ and $\alpha_s = 0$ for $s > m$, $k, s = 1, \cdots, p$. This is an ARMA(p,r) process for u_t^2. It has a unit root if $\sum_{s=1}^{p}(\delta_s + \alpha_s) = 1$. This special case is called IGARCH (Engle and Bollerslev (1986)[26]).

[25]Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, 31, 307-27.

[26]Engle, R.F. and T. Bollerslev (1986): "Modelling the Persistence of Conditional Variances", *Econometric Reviews*, **5**, 1-50.

A Granger Representation Theorem, part (2)

Note, first, that $\beta'\psi(1) = \mathbf{0}$ implies $\Pi\psi(1) = \mathbf{0}$, which, in turn, is equivalent to $(\mathbf{A}(1) - \mathbf{I})\psi(1) = \mathbf{0}$. To prove the assertion, notice that the MA(∞) representation $\Delta\mathbf{y}_t = (\mathbf{I} - L)\mathbf{y}_t = \psi(L)\epsilon_y$implies

$$\begin{aligned} & (\mathbf{I} - \mathbf{A}(L))(\mathbf{I} - L)\mathbf{y}_t &=& (\mathbf{I} - \mathbf{A}(L))\psi(L)\epsilon_t \\ \Leftrightarrow\ & (\mathbf{I} - L)(\mathbf{I} - \mathbf{A}(L))\mathbf{y}_t &=& (\mathbf{I} - \mathbf{A}(L))\psi(L)\epsilon_t \quad \text{(linear operators commute)} \\ & \Leftrightarrow\ (\mathbf{I} - L)\epsilon_t &=& (\mathbf{I} - \mathbf{A}(L))\psi(L)\epsilon_t, \end{aligned}$$

for any realization of the random vector ϵ_t. Therefore, $(\mathbf{I} - L)$ and $(\mathbf{I} - \mathbf{A}(L))\psi(L)$ represent the same polynomials in the lag operator, i.e.

$$(\mathbf{I} - z) = (\mathbf{I} - \mathbf{A}(z))\psi(z) \text{ for any } z;$$

Choosing $z = 1$ yields the desired result. □

B Useful Auxiliary Results

The following results are useful for the development of the Johansen Tests for the number of cointegrating vectors.

1. Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of dimension $n \times n$.

   1.1 Distributive law: $\det(\mathbf{AB}) = |\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$.

   1.2 Eigenvalues and matrix spectrum: The eigenvalues (characteristic roots) λ_i, $i = 1, \ldots, n$, of $\mathbf{A}$ satisfy the characteristic equation

   $$\det(\mathbf{A} - \lambda_i \mathbf{I}_n) = 0.$$

   Furthermore, there exist n eigenvectors (characteristic vectors) $\mathbf{a}_i$, $i = 1, \ldots, n$, of dimension $n \times 1$, such that

   $$(\mathbf{A} - \lambda_i \mathbf{I}_n)\mathbf{a}_i = \mathbf{0}, \ \ i = 1, \ldots, n.$$

The collection of eigenvalues of $\mathbf{A}$, $\lambda(\mathbf{A}) = \{\lambda_i, i = 1, \ldots, n\}$, is called the matrix spectrum of $\mathbf{A}$ and satisfies

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i.$$

1.3 Let λ_i, $i = 1, \ldots, n$, satisfy

$$(1)\ \ |\lambda_i \mathbf{I}_n - (\mathbf{A}'\mathbf{A})^{-\frac{1}{2}}\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-\frac{1}{2}}| = 0.$$

Then, μ_i, $i = 1, \ldots, n$, satisfying

$$(2)\ \ |\mu_i \mathbf{I}_n - (\mathbf{B}'\mathbf{B})^{-\frac{1}{2}}\mathbf{B}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-\frac{1}{2}}| = 0$$

are pairwise identical to λ_i.

Proof: Let $\tilde{\mathbf{A}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-\frac{1}{2}}$, and $\tilde{\mathbf{B}} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-\frac{1}{2}}$. Then, (1) is equivalent to $|\lambda_i \mathbf{I}_n - \tilde{\mathbf{A}}'\tilde{\mathbf{B}}\tilde{\mathbf{B}}'\tilde{\mathbf{A}}| = 0$, while (2) is equivalent to $|\mu_i \mathbf{I}_n - \tilde{\mathbf{B}}'\tilde{\mathbf{A}}\tilde{\mathbf{A}}'\tilde{\mathbf{B}}| = 0$. Letting $\mathbf{C} = \tilde{\mathbf{A}}'\tilde{\mathbf{B}}$, (1) is equivalent to $|\lambda_i \mathbf{I}_n - \mathbf{C}\mathbf{C}'| = 0$, while (2) is $|\mu_i \mathbf{I}_n - \mathbf{C}'\mathbf{C}| = 0$. Denoting the corresponding characteristic vectors by $\mathbf{x}_i$ and $\mathbf{z}_i$,

$$\begin{aligned} \mathbf{C}'\mathbf{C}\mathbf{x}_i &= \lambda_i \mathbf{x}_i, \\ \mathbf{C}\mathbf{C}'\mathbf{z}_i &= \mu_i \mathbf{z}_i, \end{aligned}$$

implying

$$\begin{aligned} \mathbf{z}_i'\mathbf{C}'\mathbf{C}\mathbf{C}'\mathbf{x}_i &= \lambda_i \mathbf{z}_i'\mathbf{C}'\mathbf{x}_i, \\ \mathbf{x}_i'\mathbf{C}\mathbf{C}'\mathbf{C}\mathbf{z}_i &= \mu_i \mathbf{x}_i'\mathbf{C}\mathbf{z}_i, \end{aligned}$$

so that $\mu_i = \lambda_i$.

1.4 Let λ_i, $i = 1, \ldots, n$, be the eigenvalues of $\mathbf{A}$. Then, $\gamma_i = 1 - \lambda_i$, $i = 1, \ldots, n$, are the eigenvalues if $\mathbf{I}_n - \mathbf{A}$.

Proof: This follows immediately from the definition of λ_i,

$$0 = |A - \lambda_i \mathbf{I}_n| = |\mathbf{A} - \mathbf{I}_n - (\lambda_i - 1)\mathbf{I}_n| = (-1)^n |\mathbf{I}_n - \mathbf{A} - (1 - \lambda_i)\mathbf{I}_n|.$$

and so $0 = |\mathbf{I}_n - \mathbf{A} - (1 - \lambda_i)\mathbf{I}_n|$.

2. Let $\mathbf{W} = [\mathbf{U}, \mathbf{V}]$, where the matrices $\mathbf{U}$ and $\mathbf{V}$ have dimensions $n \times a$ and $n \times b$, respectively. Let $\mathbf{M}_U = \mathbf{I}_n - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$, and analogously for $\mathbf{M}_V$. Then,

$$\mathbf{W}'\mathbf{W} = \begin{bmatrix} \mathbf{U}'\mathbf{U} & \mathbf{U}'\mathbf{V} \\ \mathbf{V}'\mathbf{U} & \mathbf{V}'\mathbf{V} \end{bmatrix}.$$

For the case $a = b = 1$, $\mathbf{U}$ and $\mathbf{V}$ are column vectors, hence their inner products are scalars, and so it can readily be verified that

$$\begin{aligned} \det(\mathbf{W}'\mathbf{W}) &= (\mathbf{U}'\mathbf{U})(\mathbf{V}'\mathbf{V}) - (\mathbf{U}'\mathbf{V})^2 \\ &= (\mathbf{U}'\mathbf{U})(\mathbf{V}'\mathbf{V} - (\mathbf{U}'\mathbf{V})^2/\mathbf{U}'\mathbf{U}) = (\mathbf{U}'\mathbf{U})(\mathbf{V}'\mathbf{M}_U\mathbf{V}) \\ &= (\mathbf{V}'\mathbf{V})(\mathbf{U}'\mathbf{U} - (\mathbf{U}'\mathbf{V})^2/\mathbf{V}'\mathbf{V}) = (\mathbf{V}'\mathbf{V})(\mathbf{U}'\mathbf{M}_V\mathbf{U}). \end{aligned}$$

This generalizes for arbitrary integers n:

$$\det(\mathbf{W}'\mathbf{W}) = |\mathbf{W}'\mathbf{W}| = |\mathbf{U}'\mathbf{U}||\mathbf{V}'\mathbf{M}_U\mathbf{V}| = |\mathbf{V}'\mathbf{V}||\mathbf{U}'\mathbf{M}_V\mathbf{U}|.$$

3. Concentrated Likelihood Function: Consider the normal linear regression model $y_n|\mathbf{x}_n \sim$ i.i.d. $N(\mathbf{x}_n'\beta_0, \sigma_0^2)$, $n = 1, \ldots, N$. Let $\mathbf{y} = [y_1, \ldots, y_N]'$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]'$. ML estimation of β_0 and σ_0^2 amounts to maximizing the average log-likelihood function

$$\begin{aligned} L(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) &= \frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \mathbf{x}_n'\beta)^2\right)\right), \\ \text{i.e. } \max_{\beta,\sigma^2} L(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) &\Leftrightarrow \max_{\beta,\sigma^2} \left\{-\frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{N}\sum_{n=1}^{N}(y_n - \mathbf{x}_n'\beta)^2\right\}. \end{aligned}$$

Note that the order of maximization is immaterial. For any value of β, maximization with respect to σ^2 yields the solution $\sigma^2(\beta) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mathbf{x}_n'\beta)^2$. This allows to concentrate out σ^2 in the average log-likelihood function and cast the maximization problem as a maximization over β alone:

$$\begin{aligned} \max_{\beta,\sigma^2} L(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) &\Leftrightarrow \max_{\beta} L(\beta, \sigma^2(\beta); \mathbf{y}, \mathbf{X}) \\ &\Leftrightarrow \max_{\beta} -\frac{N}{2}\ln(\sigma^2(\beta)) - \frac{1}{2} \\ &\Leftrightarrow \max_{\beta} -\frac{N}{2}\ln(\mathbf{u}(\beta)'\mathbf{u}(\beta)), \text{ where } \mathbf{u}(\beta) = \mathbf{y} - \mathbf{X}\beta. \end{aligned}$$

It is straightforward to check that this results in the well-known MLE for β_0, $\hat{\beta}$, equivalent to the OLS estimator, and in the MLE for σ_0^2, $\hat{\sigma}^2 = \sigma^2(\hat{\beta})$.

4. Details on the Sargan-Hansen J-test statistic: Let $\text{rk}(\mathbf{Z}) = m > \text{rk}(\mathbf{X}) = k$. It follows from the definition of the 2SLS estimator that $\text{var}\left(\hat{\beta}_{2\text{SLS}}\right) = \sigma^2 \left(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X}\right)^{-1}$. Therefore,

$$\begin{aligned} \text{var}\left(\mathbf{y} - \mathbf{X}\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) &= \text{var}\left(\mathbf{y}|\mathbf{X}, \mathbf{Z}\right) + \text{var}\left(\mathbf{X}\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) \\ &\quad -\text{cov}\left(\mathbf{y}, \mathbf{X}\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) \\ &\quad -\text{cov}\left(\mathbf{X}\hat{\beta}_{2\text{SLS}}, \mathbf{y} \middle| \mathbf{X}, \mathbf{Z}\right). \end{aligned}$$

Since $\mathbf{X}' P_{\mathbf{Z}} \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X}\right)^{-1} \mathbf{X}' P_{\mathbf{Z}}\right] = \mathbf{0}$ implies $\text{cov}\left(X\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) = \mathbf{0}$, it also follows that $\mathbf{X}\left(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X}\right)^{-1} \mathbf{X} P_{\mathbf{Z}} = \mathbf{X}\left(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X}\right)^{-1} \mathbf{X}'$. Hence,

$$\begin{aligned} \text{cov}\left(\mathbf{y}, \mathbf{X}\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) &= \text{var}\left(\mathbf{X}\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) \\ \Rightarrow \text{var}\left(\mathbf{y} - \mathbf{X}\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) &= \sigma^2 \left[\mathbf{I} - \mathbf{X}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}'\right] \\ &= \sigma^2 \left[\mathbf{I} - \mathbf{X}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}' P_{\mathbf{Z}}\right]. \end{aligned}$$

Also, the matrix in square brackets in the preceding expressions is idempotent and symmetric, and so

$$\begin{aligned} \text{rk}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}'\right) &= \text{tr}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}' P_{\mathbf{Z}}\right) \\ &= N - \text{tr}\left((\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}' P_{\mathbf{Z}} \mathbf{X}\right) \\ &= N - k. \end{aligned}$$

Finally,

$$\begin{aligned} \text{var}\left(\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}_{2\text{SLS}} \middle| \mathbf{X}, \mathbf{Z}\right) &= \sigma^2 \mathbf{Z}' \left(\mathbf{I} - \mathbf{X}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}'\right) \mathbf{Z} \\ &= \sigma^2 \left(\mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{X}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\right). \end{aligned}$$

Then, $\text{rk}(\mathbf{Z}'\mathbf{Z}) = m$, $\text{rk}(\mathbf{Z}'\mathbf{X}) = k$ and $\text{rk}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X}) = k$ imply that the central matrix of the J-statistic satisfies $\text{rk}\left(\mathbf{Z}'\left(\mathbf{I} - \mathbf{X}(\mathbf{X}' P_{\mathbf{Z}} \mathbf{X})^{-1} \mathbf{X}'\right) \mathbf{Z}\right) = m - k$.