Forecasting and Predictive Analytics

Guillaume Chevillon
chevillon@essec.edu

ESSEC Business School
Set of notes on Time Series Models

Contents

Preface

This set of notes was prepared for the course in Forecasting and Predictive Analytics. It is not fully complete yet but provides a reference. It was based on a course on Financial Econometrics – hence the examples presented. Regular handouts will be provided over the weeks.

Those who have not followed the course *Introduction to Econometrics* may want to consult some of the books on Econometrics in the library. Alternatively, I will try to cover in these notes most of what is needed, starting from a basic course in statistics.

The required notions are basic probabilities, random variables, the Law of Large Numbers, the Central Limit Theorem, basic statistical inference and maximum likelihood estimation. A specific handout for the first-year students at ESSEC is also available upon request.

This handout is meant to be used primarily as a reference for time series econometric methods which can be used in finance, economics, marketing, accounting research, management... Hence the specific applications will be covered in class.

Bibliography

Most chapters will also report a specific list of references.

- Brooks, C. (2002), Introductory econometrics for finance. Cambridge University Press.
- Campbell, J. Y., Lo, A. W. & A. C. MacKinlay (1997), The Econometrics of Financial Markets. Princeton University Press.
- Gourieroux, C. & J. Jasiak (2001), Financial Econometrics: problems, models and methods. Princeton University Press.
- Mills, T. (2008), The Econometric Modelling of Financial Time Series. Cambridge University Press.
- Ruppert, D. (2004), Statistics and Finance : an introduction. Springer.
- Tsay, R. S. (2005), Analysis of Financial Time Series. Wiley, second edition.

on ARCH

- Enders, W. (2004), Applied econometric time series, 2nd edn, J. Wiley, Hoboken, NJ.
- Engle, R. F., ed. (1995), ARCH : selected readings, Oxford University Press, Oxford.
- Hamilton, J. D. (1994), Time series analysis, Princeton University Press, Princeton, N.J.
- Taylor, S. J. (2005), Asset price dynamics, volatility, and prediction, Princeton University Press, Princeton, N.J.; Oxford.

Contains more specific topics:

- Alexander, C. (2008), Practical Financial Econometrics. Wiley.
- Hardle, W., Hautsch, N. and L. Overbeck (2009), Applied Quantitative Finance. Springer.
- Hull, J. C. (2009), Options, Futures and Other Derivatives. Pearson/Prentice Hall.
- Jorion, P. (2001), Value at Risk : the New Benchmark for Managing Financial Risk. McGraw-Hill.

Chapter 1

Classical Linear Regression

Linear regression is the foundation of modern econometrics. While the importance of linear regression in financial econometrics has diminished in recent years, it is still widely employed. More importantly, the theory behind least squares estimators is useful in more general contexts and many results of this chapter are special cases of more general specifications presented later in the notes. This chapter covers model specification, estimation, inference, under both the classical assumptions and using asymptotic analysis, and model selection. Linear regression is the most basic tool of any econometrician and is widely used throughout finance and economics. The success of linear regression is derived from two key aspects: simple, closed form estimators and ease of interpretation. However, despite superficial simplicity, the concepts discussed in this chapter will reappear in subsequent chapters.

1.1 Model Description

Linear regression expresses a dependent random variable as a linear function of independent variables, possibly random, and an error.

$$y_n = \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_K x_{nK} + \epsilon_n,$$

where y_n is known as the **regressand**, **dependent variable** or simply the **left-hand-side variable**. The K variables, $x_{1n}, ..., x_{Kn}$ are known as the **regressors**, **independent variables** or **right-hand-side variables**. $\beta_1, \beta_2, ..., \beta_K$ are the **regression coefficients** and ϵ_n is known as the **innovation**, **shock** or **error** and $n = 1, 2, ..., N$ indexes the observation. While this representation clarifies the relationship between y_n and the x's, matrix notation will generally be used to compactly describe models:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{X}$ is an N by K matrix, β is a K by 1 vector, and both Y and ϵ are N by 1 vectors.

The notes will occasionally make use of one of two vector notations, row,

$$\begin{bmatrix} y_1 = \mathbf{x}_1\beta + \epsilon_1 \\ y_2 = \mathbf{x}_2\beta + \epsilon_2 \\ \vdots \\ y_n = \mathbf{x}_n\beta + \epsilon_n \end{bmatrix}$$

or column

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + ... + \beta_K\mathbf{x}_K + \epsilon,$$

Linear regression allows coefficients to interpreted all things equal. Specifically, the effect of a change in one variables can be examined without changing the others. Linear regression also allows for models which contain all information relevant for determining y_n whether it is directly of interest or not. This feature provides the mechanism to interpret the coefficient on a regressor as the unique effect of that regressor under certain conditions. This feature of linear regression makes it very attractive.

1.1.1 What is a model?

What a model is and isn't is a difficult question to answer. One view of a model holds that the model is the data generating process. For instance, if a model postulates

$$y_n = \beta_1 x_n + \epsilon_n$$

one possible interpretation is the regressand, y_n is exactly determined by x_n and some random number coming from somewhere. An alternative view, one that I espouse, holds that x_n is the only relevant variable in the awareness of the researcher that can explain variation of y_n. Everything else that determines y_n cannot be measured and, in the usual case, cannot placed into a framework which would allow the researcher to formulate a model.

For instance, consider monthly returns on the S&P 500, a value weighted index of 500 large firms in the United States. Equity holdings and returns are generated by individuals based on their beliefs and preferences. If one were to take a literal a data generating process (DGP) view of the return on this index, data would have to be collected on the preferences for individuals and

their beliefs and formulated into a model for returns. This would be a daunting task to undertake, depending on how general the beliefs and preference structures are. On the other hand, a model can still be built to explain the variation based on observable quantities (such as oil price changes, macroeconomic news announcements, etc.) without explicitly collecting information on beliefs and preferences. In a model of this type, the effect of relevant explanatory variables can be thought of as an input individuals take into account when forming their beliefs and, subject to their preferences, taking actions which ultimately affect the return on the S&P 500. In the case, the model can still be meaningful even if it is not the data generating process and relationships between regressands and regressors can be explored.

In the context of time-series data, models often postulate that the recent values of a series are useful in determining the value of this series in the near future. Again, suppose that the data were monthly returns on the S&P 500 and that rather than using other explanatory variables past returns were used to explain present and future returns. Taken as a DGP, it would mean that average returns in the near future would be influenced by returns in the immediate past. Alternatively, taken a model, one interpretation would postulate that changes in beliefs other variables which influence holdings of assets change slowly (and in an unobserved manner) which produce returns which can be predicted. Of course, there are other interpretations but these should come from theory rather than data. The model as a proxy interpretation is additionally useful as it allows models to be specified which are only loosely coupled with theory but can capture interesting features of a theoretical model. Careful consideration of what a model is and isn't is a very important step in the development of an econometrician, and one should always consider what assumptions and beliefs are needed to justify a specification.

1.1.2 Example: Cross-section regression of returns on factors

Throughout this chapter, the concepts of linear regression will be explored in the context of a cross-section regression of returns on a set of factors which capture systematic risk. Cross sectional regressions in financial econometrics date back at least to the Capital Asset Pricing Model (CAPM, Markowitz (1959), Sharpe (1964) and Lintner (1965)), a model formulated as a regression of the excess return on individual assets on the excess return on the market. More general specifications with multiple regressors are motivated by the Intertemporal CAPM (ICAPM, Merton (1973)) and Arbitrage Pricing Theory (APT, Ross (1976)).

The basic model postulates that excess returns are linearly related to a set of systematic risk factors. The factors can be returns on other assets, such as the market portfolio, or any other relevant variable, such as interest rates, shocks to inflation or consumption growth.

$$r_n - r_n^f = \mathbf{f}_n\beta + \epsilon_n$$

or more compactly,

$$r_n^e = \mathbf{f}_n\beta + \epsilon_n$$

where $r_n^e = r_n - r_n^f$ is the excess return on the asset and $f_{1n}, ..., f_{Kn}$ are returns on factors capturing systematic variation in excess returns.

Linear factors models have been used in countless papers, the most well known by Fama and French (Fama & French (1992) and Fama & French (1993)) who use returns on specially constructed portfolio as factors to capture specific types of risk. The Fama-French data set is available in Eviews (`ff.wk1`), Excel (`ff.xls`) formats and contains the following variables:

- All data, except the interest rates, are from the CRSP[1] database and are available monthly from January 1927. Returns are calculated as 100 times the logarithmic return (100($\ln(p_n)$ - $\ln(p_{n-1})$)). Portfolios were constructed by sorting the firms into categories based on market capitalization, Book Equity to Market Equity (BE/ME), or past returns over the previous year. For more details, on portfolio construction see Fama & French (1993). A general model for the BH portfolio can be specified

Variable	Description
DATE	Date in format YYYYMM.
VWM	Returns on a value-weighted portfolio of all NYSE, AMEX and NASDAQ stocks
SMB	Returns on the Small minus Big factor, a zero investment portfolio that is long in small market capitalization firms and short in big caps.
HML	Returns on the High minus Low factor, a zero investment portfolio that is long in high BE/ME firms and short in low BE/ME firms.
UMD	Returns on the Up minus Does factor (also known as the Momentum factor), a zero investment portfolio that is long in firms with returns in the top 30% over the past 12 months and short in firms with returns in the bottom 30%.
SL	Returns on a portfolio of small cap and low BE/ME firms.
SM	Returns on a portfolio of small cap and medium BE/ME firms.
SH	Returns on a portfolio of small cap and high BE/M firms.
BL	Returns on a portfolio of big cap and low BE/ME firms.
BM	Returns on a portfolio of big cap and medium BE/ME firms.
BH	Returns on a portfolio of big cap and high BE/ME firms.
RF	Risk free rate (Rate on a 3 month T-bill).

$$BH_n - RF_n = \beta_1 + \beta_2(VWM_n - RF_n) + \beta_3 SMB_n + \beta_4 HML_n + \beta_5 UMD_n + \epsilon_n$$

[1] available on myessec.com and at the library

or

$$BH_n^e = \beta_1 + \beta_2 VWM_n^e + \beta_3 SMB_n + \beta_4 HML_n + \beta_5 UMD_n + \epsilon_n$$

The coefficients in the model can be interpreted as the effect of a change in one variable holding the other variables constant. For example, β_3 captures the effect of a change in the SMB_n risk factor holding VWM_n^e, HML_n and UMD_n constant. The following table contains some descriptive statistics of the factors and the six portfolios included in this data set (the data consist of monthly observations from January 1927 until July 2005(N = 943)):

Portfolio	Mean	Std. Dev	Skewness	Kurtosis
VWM^e	0.64	5.47	0.21	10.65
SMB	0.24	3.35	2.10	23.56
HML	0.40	3.58	1.80	18.26
UMD	0.75	4.71	-3.00	31.04
BHe	0.93	7.34	1.71	21.62
BMe	0.69	5.85	1.43	20.78
BLe	0.61	5.46	-0.08	8.19
SHe	1.21	8.32	2.13	23.59
SMe	1.02	7.18	1.47	18.73
SLe	0.73	7.90	1.01	13.09

1.2 Functional Form

A linear relationship is fairly specific and, in some cases, restrictive and it is important to distinguish specifications which can be examined in the framework of a linear regression from those which cannot. Linear regression requires two key features of any model: each term on the right hand side must have only one coefficient that enters multiplicatively and the error must enter additively.[2] Most specifications satisfying these two requirements can be treated using the tools of linear regression.[3] For instance, any regressor or the regressand can be non-linear transformations of observed data.

Double log (also known as **log-log**) specifications, where both the regressor and the regressands are log transformation of the original (positive) data, are common:

$$\ln y_n = \beta_1 + \beta_2 \ln x_n + \epsilon_n.$$

[2]A third but obvious requirement is that neither y_n nor any of the x_{kn} may be latent (unobservable), $n = 1, 2, ..., N; k = 1, 2, ..., K$.

[3]There are further requirements on the data, both the regressors and the regressand, to ensure that estimators of the unknown parameters are reasonable, but these are treated in a subsequent sections.

In the parlance of a linear regression, the model is specified

$$\widetilde{y}_n = \beta_1 + \beta_2 \ln \widetilde{x}_n + \epsilon_n$$

where $\widetilde{y}_n = \ln y_n$ and $\widetilde{x}_n = \ln x_n$. The usefulness of the double log specification can be illustrated by a Cobb-Douglas production function subject to a multiplicative shock

$$Y_n = \beta_1 K_n^{\beta_2} L_n^{\beta_3} \epsilon_n$$

Using the production function directly, it is not obvious that, given values for output (Y_n), capital (K_n) and labor (L_n) of firm n, the model is consistent with a linear regression. However, taking logs,

$$\ln Y_n = \beta_1 + \beta_2 \ln K_n + \beta_3 \ln L_n + \epsilon_n$$

the model can be reformulated as a linear regression on the transformed data. Other forms, such as semi-log (either log-lin, where the regressand is logged but the regressors are unchanged, or lin-log, the opposite) are often useful to describe certain relationships.

Linear regression does, however, rule out specifications which may be of interest. For instance, linear regression is not an appropriate framework to examine a model of the form

$$y_n = \beta_1 x_{1n}^{\beta_2} + \beta_3 x_{2n}^{\beta_4} + ...$$

Fortunately, more general frameworks, such as generalized method of moments (GMM) or maximum likelihood estimation (MLE), topics of subsequent chapters, can be applied.

Two other constructs, dummy variables and interactions, can be used to generate non-linear (in regressors) specifications.

A **dummy variable** is a special class of regressor that takes the value 0 or 1. In finance, dummy variables (or dummies) are used to model calendar effects, asymmetries (where the magnitude of a coefficient depends on the sign of the regressor), or group-specific effects. **Variable interactions** parameterize non-linearities into a model through products of regressors. Common interactions include powers of regressors ($x_{n1}^2, x_{n1}^3 ...$), cross-products of regressors ($x_{n1} x_{n2}$) and interactions with dummy variables. Considering the range of nonlinear transformation, linear regression is surprisingly general given the restriction of linearity.

The use of non-linear transformation also change the interpretation of the regression coefficients. If only linear function of regressors are included,

$$y_n = \mathbf{x}_n \beta + \epsilon_n$$

and $\partial y_n / \partial x_{nk} = \beta_k$. Suppose a specification includes both x_k and x_k^2 as regressors,

$$y_n = \beta_1 x_n + \beta_2 x_n^2 + \epsilon_n$$

In this specification, $\partial y_n/\partial x_{nk} = \beta_1 + 2\beta_2 x_n$ and the level of the variable enters its partial effect. Similarly, in a simple double log model

$$\ln y_n = \beta_1 \ln x_n + \epsilon_n.$$

$$\beta_1 = \frac{\partial \ln y_n}{\partial \ln x_n} = \frac{\frac{\partial y_n}{y_n}}{\frac{\partial x_{nk}}{x_{nk}}} = \frac{\%\Delta y_n}{\%\Delta x_n}$$

Thus, β_1 corresponds to the elasticity of y_n with respect to x_n. In general, the coefficient on a variable in levels corresponds to the effect of a unit changes in that variable while the coefficient on a variable in logs corresponds to the effect of a percentage change. For instance, in a semi-log model where the regressor is in logs but the regressand is in levels,

$$y_n = \beta_1 \ln x_n + \epsilon_n.$$

β_1 will correspond to a unit change in y_n for a % change in x_n. Finally, in the case of discrete regressors, where there is no differential interpretation of coefficients, β represents the effect of a whole unit change, such as a dummy going from 0 to 1.

Example 1 (Dummy variables and interactions in cross section regression) *Two calendar effects, the January and the December effects, have been widely studied in finance. Simply put, the December effect hypothesizes that returns in December are unusually low due to tax-induced portfolio rebalancing, mostly to realize losses, while the January effects stipulates returns are abnormally high as investors return to the market. To model excess returns on a portfolio (BH_n^e) as a function of the excess market return (VWM_n^e), a constant, and the January and December effects, a model can be specified*

$$BH_n^e = \beta_1 + \beta_2 VWM_n^e + \beta_3 I_{1n} + \beta_4 I_{12n} + \epsilon_n$$

where I_{1n} takes the value 1 if the return was generated in January and I_{12n} does the same for December (they are zero otherwise). The model can be reparameterized into three cases:

$$\begin{aligned} January &: BH_n^e = (\beta_1 + \beta_3) + \beta_2 VWM_n^e + \epsilon_n \\ December &: BH_n^e = (\beta_1 + \beta_4) + \beta_2 VWM_n^e + \epsilon_n \\ Otherwise &: \beta_1 + \beta_2 VWM_n^e + \epsilon_n \end{aligned}$$

Dummy interactions can be used to produce models with both different intercepts and different slopes in January and December,

$$BH_n^e = \beta_1 + \beta_2 VWM_n^e + \beta_3 I_{1n} + \beta_4 I_{12n} + \beta_5 I_{1n} VWM_n^e + \beta_6 I_{12n} VWM_n^e + \epsilon_n$$

Similarly, if excess returns on a portfolio were non-linearly related to returns on the market, a simple model can be specified

$$BH_n^e = \beta_1 + \beta_2 VWM_n^e + \beta_3 \left(VWM_n^e\right)^2 + \beta_4 \left(VWM_n^e\right)^3 + \epsilon_n$$

Dittmar (2002) proposed a similar model to explain the cross-sectional dispersion of expected returns.

1.3 Estimation

Linear regression is also known as ordinary least squares (OLS) or simply least squares, a moniker derived from the usual method of estimating the unknown regression coefficients. Least squares minimizes the squared distance between the fit line (or plane if there are multiple regressors) and the regressand. The parameters are estimated as the solution to

$$\min_{\beta}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)=\sum_{n=1}^{N}(y_n-\mathbf{x}_n\beta)^2.$$

First order conditions of this optimization problem are

$$-2\mathbf{X}'(\mathbf{y}-\mathbf{X}\beta)=-2\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}\beta=-2\mathbf{x}_n\sum_{n=1}^{N}(y_n-\mathbf{x}_n\beta)^2$$

and rearranging, the least squares estimator for β, denoted by $\widehat{\beta}$, is

$$\widehat{\beta}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

Clearly this estimator is only reasonable if $\mathbf{X}'\mathbf{X}$ is invertible. In the linear algebra notes, this is shown to be equivalent to a condition that $rank(\mathbf{X})=K$. This requirement states that no column on $\mathbf{X}$ can be exactly expressed as a combination of the $K-1$ remaining columns. This is a weak condition and is trivial to verify in most econometric software packages; using a less than full rank matrix of regressors will generate a warning or error.

There is one notable case when using dummy variables worthy of independent examination. Suppose dummy variables corresponding to the 4 quarters of the year, $I_{1n},...,I_{4n}, I_{1n}=\{Jan,Feb,Mar\}$, etc., are constructed from a monthly data set on returns. Consider a simple model with a constant and all 4 dummies

$$r_n=\beta_1+\beta_2I_{1n}+\beta_3I_{2n}+\beta_4I_{3n}+\beta_5I_{4n}+\epsilon_n.$$

It it not possible to estimate this model with all 4 dummy variables and the constant; the constant is a perfect linear combination of the dummy variables and the regressor matrix is rank deficient. The solution is to exclude either the constant or one of the dummy variables. It makes no difference in estimation which is excluded, although the interpretation of the coefficients changes. In the case where the constant is excluded, the coefficients on the dummy variables are directly interpretable as quarterly averages. If one of the dummy variables is excluded, for instance the first quarter dummy variable, the interpretation changes. In this parameterization,

$$r_n = \beta_1 + \beta_2 I_{1n} + \beta_3 I_{2n} + \beta_4 I_{3n} + \epsilon_n$$

β_1 is the effect in Q_1, while $\beta_1 + \beta_j$ is the effect in Q_j .

It is also crucial that any regressor, other the constant, be non-constant. Suppose a regression included years since public flotation but the data set only contained assets that had been floated for exactly 10 years. Including this regressor and a constant results in perfect collinearity, but, more importantly, without variability in a regressor it is impossible to determine whether changes in the regressor (years since float) results in a change in the regressand or whether the effect is simply constant across all assets. The role of variability of regressors will be revisited in consideration of the statistical properties of $\widehat{\beta}$.

The second order conditions of the minimization,

$$2\mathbf{X}'\mathbf{X}$$

ensure that the only solution must be a minimum as long as $\mathbf{X}'\mathbf{X}$ is positive definite. Again, positive definiteness of this matrix is equivalent to $rank(\mathbf{X}) = K$.

Once the regression coefficient have been estimated, it is useful to define the fit values,

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$$

and sample residuals

$$\widehat{\epsilon} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{X}\left(\beta - \widehat{\beta}\right) + \epsilon$$

Rewriting the first order condition in terms of the explanatory variables and the residuals provides insight into the numerical properties of the residuals and an equivalent first order condition is

$$\mathbf{X}'\widehat{\epsilon} = \mathbf{0}$$

This set of linear equations is commonly referred to as the normal equations or orthogonality conditions. This set of conditions requires that $\widehat{\epsilon}$ is outside of the span of the columns of $\mathbf{X}$. Moreover, considering the columns of $\mathbf{X}$ separately, $\mathbf{X}_k'\widehat{\epsilon} = \mathbf{0}$ for all $k = 1, 2, ..., K$. When a column contains a constant (an intercept in the model specification),

$$\iota'\widehat{\epsilon} = \sum_{i=1}^{n} \widehat{\epsilon}_i = 0$$

and the mean of the residuals will be exactly zero.

The OLS estimator of the variance of the residuals,[4] s^2 is given by

$$s^2 = \frac{\widehat{\epsilon}'\widehat{\epsilon}}{N - K}$$

[4]The choice of $N - K$ in the denominator will be clear when the properties of this estimator have been examined.

and the standard error of the regression is defined to be $s = \sqrt{s^2}$.

Two useful matrices, one known as the projection matrix, $\mathbf{P_X}$ and the other known as the annihilator matrix, $\mathbf{M_X}$, are defined

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\mathbf{M_X} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

These two matrices have some desirable properties. Both the fitted value of $\mathbf{y}$ ($\widehat{\mathbf{y}}$) and the estimated errors, $\widehat{\epsilon}$ can be simply expressed in terms of these matrices as $\mathbf{P_X y}$ and$\mathbf{M_X y}$ respectively. These matrices are also idempotent:

$$\mathbf{P_X P_X} = \mathbf{P_X} \quad and \quad \mathbf{M_X M_X} = \mathbf{M_X}$$

and orthogonal:

$$\mathbf{P_X M_X} = \mathbf{0}$$

The projection matrix captures the portion of $\mathbf{y}$ that lies in the linear space spanned by $\mathbf{X}$, while the annihilator matrix captures the portion of $\mathbf{y}$ which lies in the null space of $\mathbf{X}$. In essence, the annihilator matrix annihilates any portion of $\mathbf{y}$ which is explainable by $\mathbf{X}$ leaving only the residuals.

The least squares estimator has two final properties worth mentioning. First, non-singular transformations of the $\mathbf{x}$'s and non-zero scalar transformations of the $\mathbf{y}$'s have deterministic effects on the estimated regression coefficients. Suppose $\mathbf{A}$ is an K dimensional non-singular matrix and c is a non-zero scalar. The coefficients of a regression of cy_n on $x_n\mathbf{A}$ are

$$\begin{aligned}
\widetilde{\beta} &= \left[(\mathbf{XA})'(\mathbf{XA})\right]^{-1}(\mathbf{XA})'c\mathbf{y} \\
&= c\left[\mathbf{A}'\mathbf{X}'\mathbf{XA}\right]^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\
&= c\mathbf{A}^{-1}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{A}'^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\
&= c\mathbf{A}^{-1}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} \\
&= c\mathbf{A}^{-1}\widehat{\beta}
\end{aligned}$$

Second, as long as the model contains a constant, the regression coefficients on all terms except the intercept are unaffected by adding arbitrary constant to either the regressor or the regressands. Consider transforming the standard specification,

$$y_n = \beta_1 + \beta_2 x_{1n} + ... + \beta_k x_{kn} + \epsilon_n$$

to

$$\widetilde{y}_n = \beta_1 + \beta_2 \widetilde{x}_{1n} + ... + \beta_k \widetilde{x}_{kn} + \epsilon_n$$

where $\widetilde{y}_n = y_n + c_y$ and $\widetilde{x}_{kn} = x_{kn} + x_{x_{kn}}$. This model is exactly equivalent to

$$y_n = \widetilde{\beta}_1 + \beta_2 x_{1n} + ... + \beta_k x_{kn} + \epsilon_n$$

where $\widetilde{\beta}_1 = \beta_2 c_{x_{1n}} + ... + \beta_k c_{x_{1n}} - c_y$.

Exercise 1 (Estimation of Cross-Section regressions of returns on factors) *Estimation of regression coefficients is straight forward. The table below contains the estimated regression coefficients as well as the standard error of the regression for the 6 portfolios in the Fama-French data set in a specification including all four factors and a constant. Comparing the magnitude of the standard error of the regression to the magnitude of the standard deviation of the original data, there has been a substantial reduction. The next section will formalize how this reduction is interpreted.*

Portfolio	*Constant*	VWM^e	SMB	HML	UMD	σ
BH^e	*-0.06*	*1.07*	*0.01*	*0.81*	*-0.04*	*1.31*
BM^e	*-0.00*	*0.98*	*-0.13*	*0.32*	*-0.03*	*1.26*
BL^e	*0.08*	*1.02*	*-0.10*	*-0.23*	*-0.01*	*0.77*
SH^e	*0.05*	*1.02*	*0.92*	*0.76*	*-0.03*	*0.73*
SM^e	*0.06*	*0.98*	*0.82*	*0.32*	*0.00*	*1.06*
SL^e	*-0.09*	*1.08*	*1.04*	*-0.18*	*-0.06*	*1.25*

This table records the estimated regression coefficients from the model $r_n^{pi} = \beta_1 + \beta_2 VWM^e + \beta_3 SMB + \beta_4 HML + \beta_5 UMD$, where r_n^{pi} is the excess return on one of the six size and BE/ME portfolios. The final column contains the standard error of the regression.

1.4 Assessing Fit

Once the parameters have been estimated, the next natural step is to determine whether or not the model fits the data. The minimized sum of squared errors, the objective of the optimization, is an obvious choice to use to assess fit. However, there is one major drawback to this approach: changes in the scale of y_n alter the minimized value without changing the fit. In order to devise a scale free metric, it is necessary to distinguish between the portions of $\mathbf{y}$ which can be fit by $\mathbf{X}$ and those which cannot. Decomposing $\mathbf{y}$ using the projection and annihilator matrices,

$$\mathbf{y} = \mathbf{P_X y} + \mathbf{M_X y}$$

which follows from $\mathbf{P_X} + \mathbf{M_X} = \mathbf{I}_N$. The squared observations can be decomposed

$$\begin{aligned}
\mathbf{y'y} &= \left(\mathbf{P_X y} + \mathbf{M_X y}\right)' \left(\mathbf{P_X y} + \mathbf{M_X y}\right) \\
&= \mathbf{y'P'_X P_X y} + \mathbf{y'M'_X M_X y} + \mathbf{y'P'_X M_X y} + \mathbf{y'M'_X P_X y} \\
&= \mathbf{y'P_X y} + \mathbf{y'M_X y} + \mathbf{0} + \mathbf{0}
\end{aligned}$$

noting that $\mathbf{P_X}$ and $\mathbf{M_X}$ are idempotent (and hence symmetric) and $\mathbf{P_X M_X} = \mathbf{0}_N$. These three quantities are often referred to as[5]

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= \sum_{i=1}^{n} y_i^2 && \text{Uncentered Total Sum of Squares (TSS}_\text{U}\text{)} \\
\mathbf{y}'\mathbf{P_X y} &= \sum_{i=1}^{n} \widehat{y}_i^2 && \text{Uncentered Regression Sum of Squares (RSS}_\text{U}\text{)} \\
\mathbf{y}'\mathbf{M_X y} &= \sum_{i=1}^{n} \widehat{\epsilon}_i^2 && \text{Uncentered Sum of Squared Errors (SSE}_\text{U}\text{)}
\end{aligned}$$

Dividing through by $\mathbf{y}'\mathbf{y}$

$$\frac{\mathbf{y}'\mathbf{P_X y}}{\mathbf{y}'\mathbf{y}} + \frac{\mathbf{y}'\mathbf{M_X y}}{\mathbf{y}'\mathbf{y}} = 1$$

or

$$\frac{\text{RSS}_\text{U}}{\text{TSS}_\text{U}} + \frac{\text{SSE}_\text{U}}{\text{TSS}_\text{U}} = 1$$

This identity expresses the scale-free total variation in y that is captured by $\mathbf{X}$ ($\mathbf{y}'\mathbf{P_X y}$) and that which is not ($\mathbf{y}'\mathbf{M_X}$). The portion of the total variation explained by $\mathbf{X}$ is known as the uncentered $\mathbf{R}^2$ ($\mathbf{R}^2_U$),

Definition 1 (Uncentered $\mathbf{R}^2$ ($\mathbf{R}^2_U$))

$$\mathbf{R}^2_U = \frac{RSS_U}{TSS_U} = 1 - \frac{SSE_U}{TSS_U}$$

While this measure is scale free it suffers from one major shortcoming. Suppose a constant is added to y, so that the TSS_U changes to $(\mathbf{y} + c)'(\mathbf{y} + c)$. The identity still holds and $(\mathbf{y} + c)'(\mathbf{y} + c)$ must increase for a sufficiently large c and one of the right-hand side variables must also grow larger. In the usual case where the model contains a constant, the increase will occur in the RSS_U and as c grows arbitrarily large, **uncentered $\mathbf{R}^2$ will asymptotically tend to to one**. To overcome this limitation, a measure can be constructed which depends on deviations from the mean rather than on levels.

Let $\widetilde{\mathbf{y}} = \mathbf{y} - \overline{y} = \mathbf{M}_\iota \mathbf{y}$. Then

$$\begin{aligned}
\mathbf{y}'\mathbf{M}_\iota\mathbf{P_X}\mathbf{M}_\iota\mathbf{y} + \mathbf{y}'\mathbf{M}_\iota\mathbf{M_X}\mathbf{M}_\iota\mathbf{y} &= \mathbf{y}'\mathbf{M}_\iota\mathbf{y} \\
\frac{\mathbf{y}'\mathbf{M}_\iota\mathbf{P_X}\mathbf{M}_\iota\mathbf{y}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}} + \frac{\mathbf{y}'\mathbf{M}_\iota\mathbf{M_X}\mathbf{M}_\iota\mathbf{y}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}} &= 1
\end{aligned}$$

[5]There is no consensus on the names of these quantities. In some texts, the component capturing the fit portion is known as the Regression Sum of Squares (RSS) while in other it is known as the Explained Sum of Squares (ESS), while the portion attributable to the errors is known as the Sum of Squared Errors (SSE), the Sum of Squared Residuals (SSR) ,the Residual Sum of Squares (RSS) or the Error Sum of Squares (ESS). The choice to use SSE and RSS in this text was to ensure the reader that SSE must be the component of the squared observations relating to the error variation.

or more compactly

$$\frac{\widetilde{\mathbf{y}}'\mathbf{P}_{\mathbf{X}}\widetilde{\mathbf{y}}}{\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}} + \frac{\widetilde{\mathbf{y}}'\mathbf{M}_{\mathbf{X}}\widetilde{\mathbf{y}}}{\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}} = 1$$

Centered $\mathbf{R}^2$ ($\mathbf{R}^2_C$) is defined analogously to uncentered replacing the uncentered sums of squares with their centered counterparts. The same holds for TSS_C, RSS_C and SSE_C.

Definition 2 (Centered $\mathbf{R}^2$ ($\mathbf{R}^2_C$))

$$\mathbf{R}^2_C = \frac{RSS_C}{TSS_C} = 1 - \frac{SSE_C}{TSS_C}$$

Generally, when the unqualified expressions $\mathbf{R}^2$, SSE, RSS and TSS are used, they should all be assumed to be centered versions.

With two versions of $\mathbf{R}^2$ available, which generally differ, which should you use? Centered should be used if the model is centered (contains a constant) and uncentered should be used when it does not. Failing to chose the correct $\mathbf{R}^2$ can lead to incorrect inference about the fit of the model. Mixing the definitions can lead to nonsensical $\mathbf{R}^2$ which fall outside of [0, 1]. For instance, computing $\mathbf{R}^2$ using the centered version when the model does not contain a constant often results in a negative when

$$\mathbf{R}^2 = 1 - \frac{\text{SSE}_\text{C}}{\text{TSS}_\text{C}}$$

Most software will return centered $\mathbf{R}^2$ and caution is warranted if a model is fit without a constant.

$\mathbf{R}^2$ does have some caveats. Assuming the correct version is being used, adding an additional regressor will always (weakly) increase the $\mathbf{R}^2$; the sum of squared errors cannot increase by the inclusion of an additional regressor. This renders $\mathbf{R}^2$ useless in discriminating between two models where one is nested in the other. One solution to this problem is to use the degree of freedom adjusted $\mathbf{R}^2$.

Definition 3 (R-bar squared ($\overline{\mathbf{R}}^2$))

$$\overline{\mathbf{R}}^2 = 1 - \frac{\frac{SSE}{N-K}}{\frac{TSS}{N-1}} = 1 - \frac{SSE}{TSS}\frac{N-1}{N-K}$$

$\overline{\mathbf{R}}^2$ will increase if the decrease in the SSE is large enough to compensate for a loss of 1 degree of freedom, captured by the $N - K$ term. However, if the SSE doesn't change, $\overline{\mathbf{R}}^2$ will decrease. $\overline{\mathbf{R}}^2$ is preferable to $\mathbf{R}^2$ for comparing models, although the topic of model selection will be more formally considered at the end of this chapter. $\overline{\mathbf{R}}^2$, like $\mathbf{R}^2$ should be constructed from the appropriate versions of the RSS, SSE and TSS.

Secondly, $\mathbf{R}^2$ **is not invariant to changes in the regressand**. One fairly common mistake is to use $\mathbf{R}^2$ to compare the fit from two models with different regressands, for instance y_n and $ln(y_n)$.

These numbers are completely incompatible and these comparisons should be avoided. Moreover, $\mathbf{R}^2$ is even sensitive to more benign transformations. Suppose a simple model is postulated

$$y_n = \beta_1 + \beta_2 x_n + \epsilon_n$$

A model logically consistent with the original model,

$$y_n - x_n = \beta_1 + (\beta_2 - 1)\, x_n + \epsilon_n$$

can be formulated. However, the two $\mathbf{R}^2$s will differ in general. Suppose the original coefficient on x_n was 1. Subtracting x_n will reduce the explanatory power of x_n, rendering it effectively useless and should produce a $\mathbf{R}^2$ close to 0 irrespective of the $\mathbf{R}^2$ in the original model.

Thus far, all of these derivations and identities are purely numerical. They do not tell us whether $\widehat{\beta}$ is a reasonable way to estimate the unknown β. To make sense of $\widehat{\beta}$, we need to make some assumptions about the innovations and the regressors.

Example 2 (R^2 and $\overline{R}^2$ in Cross-Sectional Factor models) *To illustrate the use of R^2, and the problems with it, consider a model for BH^e which can depend on one or more of the risk factors. The following table records Centered and uncentered R^2 and $\overline{R}^2$ from a variety of factors models. Bold indicates the correct version (centered or uncentered) for that model. R^2 is monotonically increasing in larger models, while $\overline{R}^2$ is not.*

Regressors	R^2_C	R^2_U	$\overline{R}^2_C$	$\overline{R}^2_U$
VWM^e	***0.809***	*0.805*	***0.809***	*0.805*
VWM^e, SMB	***0.809***	*0.806*	***0.809***	*0.806*
VWM^e, HML	***0.967***	*0.966*	***0.967***	*0.966*
VWM^e, SMB, HML	***0.967***	*0.967*	***0.967***	*0.966*
VWM^e, SMB, HML, UMD	***0.968***	*0.967*	***0.968***	*0.967*
$1, VWM^e$	*0.809*	***0.806***	*0.809*	***0.806***
$1, VWM^e, SMB$	*0.809*	***0.806***	*0.809*	***0.806***
$1, VWM^e, SMB, HML$	*0.967*	***0.967***	*0.967*	***0.967***
$1, VWM^e, SMB, HML, UMD$	*0.968*	***0.967***	*0.968*	***0.967***

The R^2 above show two things. First, the excess return on the market portfolio alone can explain 80% of the variation in excess returns on the big-high portfolio. Second, the HML factor appears to have additional explanatory power on top of the excess market evidenced by increases in R^2 from 0.80 to 0.96. The centered and uncentered R^2 are very similar because the intercept in the model is near zero. Suppose that the dependent variable is changed to $10 + BH^e$ or $100 + BH^e$ and attention is restricted to the CAPM.

Using the incorrect definition for R^2 can lead to nonsensical (negative) and misleading (artificially near 1) values. Finally, the table also illustrates the problems of changing the regressand, consider

replacing BH_n^e with $BH_n^e - VWM_n^e$. The R^2 decreases from a respectable 0.80 to 0.10. However, the interpretation of the model is identical:

Regressand	*Regressors*	R_C^2	R_U^2	$\overline{R}_C^2$	$\overline{R}_U^2$
BH^e	VWM^e	***0.809***	*0.805*	***0.809***	*0.805*
BH^e	$1, VWM^e$	*0.809*	***0.806***	*0.809*	***0.806***
$10 + BH^e$	VWM^e	***0.353***	*-1.080*	***0.353***	*-1.080*
$10 + BH^e$	$1, VWM^e$	*0.939*	***0.806***	*0.939*	*0.806*
$100 + BH^e$	VWM^e	***0.032***	*-182.7*	***0.032***	*-182.7*
$100 + BH^e$	$1, VWM^e$	*0.999*	***0.806***	*0.999*	***0.806***
BH^e	$1, VWM^e$	*0.809*	***0.806***	*0.809*	***0.806***
$BH_n^e - VWM_n^e$	$1, VWM^e$	*0.112*	***0.106***	*0.112*	***0.105***

Centered and uncentered R^2 and $\overline{R}^2$ from models with regressor changes. Using the wrong R^2 can lead to nonsensical values (negative) or a false sense of fit (R^2 near one). The bottom two lines examine the effect of subtracting a regressor before fitting the model: the R^2 decreases sharply. However, this should not be viewed as problematic since models with different regressands cannot be compared using R^2. The values are negative.

1.5 Assumptions

Estimation of parameters and assessing fit are purely numerical problems. However, the goal of most regressions is to produce interpretable coefficients which requires assumptions on the properties of the model, innovations and regressors. Two broad classes of assumptions can be used to analyze the behavior of $\widehat{\beta}$, classical (also known as small sample or finite sample) and asymptotic analysis.

Neither method is ideal. Small sample is precise in that the exact distribution for any number of observations N is known. However, this precision comes at the cost of many restrictive assumptions; assumption not usually plausible in most financial applications. On the other hand, asymptotic analysis requires few restrictive assumptions and is broadly applicable to financial data. However, as the word asymptotic implies, the results are only exactly true when the number of observations is infinite. However, asymptotic analysis can still be used to examine the behavior in finite samples where it is assumed that the sample size is large enough for the asymptotic distribution to reasonably approximate the small (but unknown) sample distribution. This leads to the most important question of asymptotic analysis: How large does N need to be before the approximation is reasonable? Unfortunately, the answer to this question is "It depends". In simple cases, where residuals are independent and identically distributed, as few as 30 observations are sufficient for the asymptotic distribution to be a good approximation to the finite sample distri-

bution. In more complex cases, anywhere from 100 to 1,000 can be needed, while in the most extreme cases, where the data is extremely heterogenous and highly dependent, the approximation can be bad with more than 1,000,000 observations.

Fortunately, the latter are reasonably easy to detect and are examined in more detail consider in the time series and volatility sections.

The properties of $\widehat{\beta}$ will be examined under both sets of assumptions. While small sample results are not generally applicable, it is important to understand these results as lingua franca of econometric, the limitations of tests based on the classical assumptions, and to be able to detect when a test statistic may not have the intended asymptotic distribution. Six assumptions are required to examine the distribution of $\widehat{\beta}$ and examine the optimality of the OLS procedure, although many properties only require a subset.

Assumption 1.1 (Linearity) $y_n = \mathbf{x}_n\beta + \epsilon_n$

This assumption simply states the conditions necessary for least squares to be a reasonable method to estimate the unknown parameters of the model and imposes a less obvious condition: $\mathbf{x}_n$ must be observed and measured without error. Many applications in financial econometrics include latent variables. Linear regression, as described in this text, is not appropriate for treating these cases. In other cases, the true value of x_{nk} is not observed and a proxy must be used, $\widetilde{x}_{nk} = x_{nk} + \nu_{nk}$ where ν_{nk} is some noise process. In these cases, standard regression theory is misleading and a modified procedure (Two-stage-least-squares (2SLS) or instrumental variables (IV)) must be used.

Assumption 1.2 (Conditional Mean) $\mathsf{E}[\epsilon_n|\mathbf{X}] = 0, n = 1, 2, ..., N$

This assumption states that mean of each ϵ_n is zero given any x_{nk}, any function of any x_{nk} or combinations of these. It's stronger than the assumption used in the asymptotic section and is not appropriate for some applications. Specifically, when the regressand and regressor consist of time series data, this assumption may be violated and $\mathsf{E}[\epsilon_n|\mathbf{x}_{n+j}] \neq 0$ for some j. This assumption also implies that the correct form of x_{nk} enters the regression, that $\mathsf{E}[\epsilon_n] = 0$ (simple application of the law of iterated expectations), and that the innovations are uncorrelated with the regressors, so that $\mathsf{E}[\epsilon_n x_{jk}] = 0, n = 1, 2, ..., N; j = 1, 2, ..., N; k = 1, 2, ..., K$.

Assumption 1.3 (Rank) *The rank of* $\mathbf{X}$ *is* K *with probability 1.*

This assumption is needed to ensure that $\widehat{\beta}$ is identified and can be estimated. In practice, it requires no exact collinearity, the number of observations be at least as large as the number of regressors ($N \geq K$) and that variables other than a constant have non-zero variance.

Assumption 1.4 (Conditional Homoskedasticity) $\mathsf{E}[\epsilon_n|\mathbf{X}] = \sigma^2$

Homoskedasticity is rooted in *homo* (meaning same) and *skedannumi* (meaning scattering) and in modern English means identical variance. This assumption is needed to examine the optimality

of the OLS estimator. This assumption specifically rules out the case where the variance of an innovation is a function of a regressor.

Assumption 1.5 (Conditional Uncorrelatedness) $\mathsf{E}[\epsilon_i \epsilon_j | \mathbf{X}] = 0, i = 1, 2, ..., N; j = 1, 2, ..., N; i \neq j$.

Assuming the residuals are conditionally uncorrelated is convenient when coupled with the homoskedasticity assumption: the covariance of the residuals will be $\sigma^2 \mathbf{I}_N$. Like homoskedasticity, this assumption is needed for discussing optimality of the estimator.

Assumption 1.6 (Conditional Normality) $\epsilon_n | \mathbf{X} \sim \mathsf{N}\left(0, \sigma^2_{\epsilon_n}\right)$

The final assumption is by far the strongest. Assuming a distribution for the error terms makes the model, and anything (such as inference) based on this assumption, very restrictive but allows for precise statements on the finite sample distribution of the estimator and test statistics. This assumption, when combined with the Homoskedasticity and Conditional Uncorrelatedness assumptions provides a simple distribution for the innovation,

$$\epsilon_n | \mathbf{X} \sim \mathsf{NID}\left(0, \sigma^2\right)$$

1.6 Small Sample Properties of OLS estimators

Building off of these assumptions, important properties of $\widehat{\beta}$ can be derived. Recall that $\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$.

Theorem 1.1 (Bias of $\widehat{\beta}$) *Under assumptions 1.1-1.3,*

$$\mathsf{E}\left[\widehat{\beta} | \mathbf{X}\right] = \beta$$

Thus, the OLS estimator is unbiased. This is a desirable property, but it is not particularly meaningful without further results. For instance, an estimator which is unbiased, but does not increase in precision as the sample size increases is generally not desirable. Fortunately, under the assumptions above, $\widehat{\beta}$ is not only unbiased, it has a variance that goes to zero.

Theorem 1.2 (Variance of $\widehat{\beta}$) *Under assumptions 1.1-1.5,*

$$\mathsf{V}\left[\widehat{\beta} | \mathbf{X}\right] = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Under the conditions necessary for unbiasedness for β, plus assumptions about homoskedasticity and conditional uncorrelatedness of the residuals, the form of the variance is simple. $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ will generally decrease as the sample size increases and the variance of the parameter estimates will decrease.

However, $\widehat{\beta}$ has an even stronger property under the same assumptions. It is BLUE: Best Linear Unbiased Estimator. Best, in this context, means that it has the lowest variance among all

other linear unbiased estimators. While this is a strong result, a few words of caution are needed. The class of Linear Unbiased Estimators (LUEs) is small in the universe of all estimators. Saying OLS is the 'best' is akin to a one-armed boxer claiming to be the best one-arm boxer. While possibly true, she probably wouldn't stand a chance against two-armed boxers.

Theorem 1.3 (Gauss-Markov Theorem) *Under assumptions 1.1-1.5, $\widehat{\beta}$ is the minimum variance estimator among all linear unbiased estimators. That is $\mathsf{V}\left[\widetilde{\beta}|\mathbf{X}\right] - \mathsf{V}\left[\widehat{\beta}|\mathbf{X}\right]$ is positive semi-definite where $\widetilde{\beta} = \mathbf{C}\mathbf{y}$ and $\mathsf{E}\left[\widetilde{\beta}|\mathbf{X}\right] = \beta$.*

Letting $\widetilde{\beta}$ be any other linear, unbiased estimator of β, it must have a (weakly) larger covariance. However, many estimators, including most maximum likelihood estimators, are nonlinear and are not necessarily less efficient. Finally, making use of the normality assumption, it is possible to determine the conditional distribution of $\widehat{\beta}$.

Theorem 1.4 (Distribution of $\widehat{\beta}$) *Under assumptions 1.1-1.6*

$$\widehat{\beta} \sim \mathsf{N}\left(\beta, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

Theorem 1.4 should not be surprising. $\widehat{\beta}$ is a linear combination of (jointly) normally distributed random variables and thus are also normally distributed. Normality is useful for establishing the relationship between the estimated residuals $\widehat{\epsilon}$ and the estimated parameters $\widehat{\beta}$.

Theorem 1.5 (Conditional independence of $\widehat{\epsilon}$ and $\widehat{\beta}$) *Under assumptions 1.1-1.6, $\widehat{\epsilon}$ is independent of $\widehat{\beta}$, conditionally on $\mathbf{X}$.*

One implication of this theorem is that $\mathsf{Cov}\left[\widehat{\epsilon}_n, \widehat{\beta}_k|\mathbf{X}\right] = 0, n = 1, 2, ..., N; k = 1, 2, ..., K$. Independence is useful because functions of only $\widehat{\epsilon}$ will be independent of function of only $\widehat{\beta}$, a property very useful in deriving distributions of test statistics which depend on both.

Finally, in the small sample setup, the exact distribution of

$$s^2 = \frac{\widehat{\epsilon}'\widehat{\epsilon}}{N-K}$$

the sample error variance estimators can be deduced.

Theorem 1.6 (Distribution of $\widehat{\sigma}^2$)

$$(N-K)\, s^2/\sigma^2 \sim \chi^2_{N-K}$$

where $s^2 = (N-K)^{-1}\, \mathbf{y}'\mathbf{M}_{\mathbf{X}}\mathbf{y} = (N-K)^{-1}\, \widehat{\epsilon}'\widehat{\epsilon}$

Since $\widehat{\epsilon}_n$ is a normal random variable, once it is standardized and squared, it should be a χ^2. The change in the divisor from N to $N-K$ reflects the loss in degrees of freedom owing to the fact that K parameters are estimated. Essentially, the $\widehat{\epsilon}_n$ are "over-fitted".

1.7 Maximum Likelihood

Once the assumption that the innovations are conditionally normal has been made, conditional maximum likelihood is an obvious method to estimate the unknown parameters $\left(\beta,\sigma^2\right)$. Conditioning on $\mathbf{X}$, and assuming that the innovations are normal, homoskedastic, and conditionally uncorrelated, the likelihood is given by

$$f\left(\mathbf{y}|\mathbf{X};\beta,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-\frac{N}{2}}\exp\left(-\frac{\left(\mathbf{y}-\mathbf{X}\beta\right)'\left(\mathbf{y}-\mathbf{X}\beta\right)}{2\sigma^2}\right)=L\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right)$$

and, taking logs, the log likelihood

$$\ell\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right)=-\frac{N}{2}\log\left(2\pi\right)-\frac{N}{2}\log\left(\sigma^2\right)-\frac{\left(\mathbf{y}-\mathbf{X}\beta\right)'\left(\mathbf{y}-\mathbf{X}\beta\right)}{2\sigma^2}$$

Recall that the logarithm is a monotonic, strictly increasing transformation, and the extremums of the log-likelihood and the likelihood will occur at the same points. Maximizing the likelihood with respect to the unknown parameters, there are $K+1$ first order conditions

$$\begin{aligned}\frac{\partial\ell}{\partial\beta}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right)&=\frac{\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\beta\right)}{2\sigma^2}=\mathbf{0}\\ \frac{\partial\ell}{\partial\sigma^2}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right)&=-\frac{N}{2\sigma^2}+\frac{\left(\mathbf{y}-\mathbf{X}\beta\right)'\left(\mathbf{y}-\mathbf{X}\beta\right)}{2\sigma^4}\end{aligned}$$

The first set of conditions is identical to the first order conditions of the least squares estimator ignoring the scaling by σ^2, assumed to be greater than 0. Thus, the solution to the $K+1$ first order conditions is given by

$$\begin{aligned}\widehat{\beta}_{MLE}&=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\\ \widehat{\sigma}^2_{MLE}&=N^{-1}\left(\mathbf{y}-\mathbf{X}\beta\right)'\left(\mathbf{y}-\mathbf{X}\beta\right)=N^{-1}\widehat{\epsilon}'\widehat{\epsilon}\end{aligned}$$

The coefficients are identical under maximum likelihood and OLS. Does this mean that normality is implicitly assumed any time OLS is used? No. The derivation of the OLS estimator does not require any assumptions about normality. Moreover, the unbiasedness, variance, and BLUE properties do not rely on conditional normality of residuals. However, if the innovations are homoskedastic, uncorrelated and normal, the results of the Gauss-Markov theorem can be strengthened using the Cramer-Rao lower bound.

Theorem 1.7 (Cramér-Rao Inequality) *Let $f\left(\mathbf{z};\theta\right)$ be the joint density of $\mathbf{z}$ where θ is a K dimensional parameter vector. Let $\widehat{\theta}$ be an unbiased estimator of θ with finite covariance. Under some regularity condition on $f(\ )$*

$$\mathsf{V}\left[\widehat{\theta}\right]>\mathcal{I}^{-1}\left(\theta\right)$$

where

$$\mathcal{I}(\theta) = \mathsf{E}\left[\frac{\partial \ln f}{\partial \theta}\frac{\partial \ln f}{\partial \theta'}\right]$$

Let

$$\mathcal{J}(\theta) = -\mathsf{E}\left[\frac{\partial^2 \ln f}{\partial \theta \partial \theta'}\right]$$

the under some additional regularity conditions

$$\mathcal{I}(\theta) = \mathcal{J}(\theta)$$

The last part of this theorem is known as the **information matrix equality**. Essentially, when a model is correctly specified in its entirety, the expected covariance of the scores (the score function is the first-order derivative of the log likelihood) is equal to negative of the expected hessian.[6] The **IME** will be revisited in later chapters. Thus, to find a lower bound for the variance of the MLE estimator, expectation of the second order conditions, given by

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta \partial \beta'}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right) &= -\frac{\mathbf{X}'\mathbf{X}}{\widehat{\sigma}^2_{MLE}} \\
\frac{\partial^2 \ell}{\partial \beta \partial \sigma^2}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right) &= -\frac{\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\beta\right)}{2\sigma^4} \\
\frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right) &= \frac{N}{2\sigma^4} - \frac{\left(\mathbf{y}-\mathbf{X}\beta\right)'\left(\mathbf{y}-\mathbf{X}\beta\right)}{2\sigma^6}
\end{aligned}$$

Taking expectations of the second derivatives,

$$\begin{aligned}
\mathsf{E}\left[\frac{\partial^2 \ell}{\partial \beta \partial \beta'}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right)\right] &= -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} \\
\mathsf{E}\left[\frac{\partial^2 \ell}{\partial \beta \partial \sigma^2}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right)\right] &= 0 \\
\mathsf{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2}\left(\beta,\sigma^2;\mathbf{y}|\mathbf{X}\right)\right] &= -\frac{N}{2\sigma^4}
\end{aligned}$$

Thus the lower bound for the variance of $\widehat{\beta}_{MLE}$ is $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. However, this is the variance of the OLS estimator $\widehat{\beta}$ and the Gauss-Markov theorem can be strengthened in the case of conditionally homoskedastic, uncorrelated normal residuals.

Theorem 1.8 (Best Unbiased Estimator) *Under assumptions 1.1-1.6, $\widehat{\beta} = \widehat{\beta}_{MLE}$ is the best unbiased estimator for β.*

The difference between this theorem and theGauss-Markov theorem is subtle but important. The class of estimators is no longer restricted to just linear estimators and this result is both broad and powerful: MLE (or OLS) is an ideal (in the sense that no other unbiased estimator, linear or not, has a lower variance) estimator under these assumptions.

Finally, because $\mathsf{E}\left[\widehat{\sigma}^2_{MLE}\right] = \frac{N}{N-K}\sigma^2$, the optimality of $\widehat{\sigma}^2_{MLE}$ cannot be established using the Cramér-Rao theorem.

[6] There are quite a few regularity conditions for the IME to hold, but discussion of these is beyond the scope of this course. Interested readers could see White (1996) for a thorough discussion.

1.8 Small Sample Hypothesis Testing

Most regressions are estimated with the goal of testing implications of economic or finance theory. Hypothesis testing is the mechanism used to determine whether data are congruent to these by measuring whether the data are consistent with the theory. Formalized in terms of β, the null hypothesis (also known as the maintained hypothesis) is formulated as

$$\mathsf{H}_0 : \mathbf{R}(\beta) - \mathbf{r} = 0$$

where $\mathbf{R}(\cdot)$ is a function from $\mathbb{R}^K$ to $\mathbb{R}^M$, $M \leq K$. Initially, a subset of all hypotheses, those in the linear equality hypotheses class, formulated

$$\mathbf{R}\beta - r = 0$$

will be examined, where $\mathbf{R}$ is an $M \times K$ matrix. In subsequent chapters, more general specification including non-linear restrictions on the parameters will be considered. All hypotheses in this class can be written as weighted sums of the regression coefficients,

$$\begin{aligned} R_{11}\beta_1 + R_{12}\beta_2 ... + R_{1K}\beta_K &= r_1 \\ &\vdots \\ R_{M1}\beta_1 + R_{M2}\beta_2 ... + R_{MK}\beta_K &= r_M \end{aligned}$$

Each constraint is represented as a row in the above set of equations. Linear equality constraints can be used to test parameter restrictions such as

$$\begin{aligned} \beta_1 &= 0 \\ 3\beta_2 + \beta_3 &= 1 \\ \sum_{k=1}^{K} \beta_k &= 0 \\ \beta_1 = \beta_2 = \beta_3 &= 0 \end{aligned}$$

For instance, if the unrestricted model was

$$y_n = \beta_1 + \beta_2 x_{1n} + \beta_3 x_{3n} + \beta_4 x_{4n} + \epsilon_n$$

the hypotheses can be described in terms of $\mathbf{R}$ and $\mathbf{r}$ as

H_0	$\mathbf{R}$	$\mathbf{r}$
$3\beta_2 + \beta_3 = 1$	$\begin{bmatrix} 0 & 3 & 1 & 0 \end{bmatrix}$	1
$\sum_{k=1}^{K} \beta_k = 0$	$\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$	0
$\beta_1 = \beta_2 = \beta_3 = 0$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

In the case of linear equality constraints, alternatives are generally formulated as $\mathsf{H}_1 : \mathbf{R}\beta - \mathbf{r} \neq$. Once both the null the alternative hypotheses have been postulated, it is necessary to discern whether the data are consistent with the null hypothesis. Three classes of statistics will be described to test these hypotheses: **Wald**, **Lagrange Multiplier** and **Likelihood Ratio**. Wald tests are perhaps the most intuitive: they directly test whether $\mathbf{R}\beta - \mathbf{r}$ is close to zero.

Lagrange Multiplier (LM) test incorporate the constraint into the least squares problem using a lagrangian. If the constraint has little effect on the minimized sum of squares, the lagrange multipliers, often described as the shadow price of the constraint in economic applications, should be close to zero. The magnitude of these forms the basis of the test statistic.

Finally, likelihood ratios (LR) test whether the data are less likely under the null than they are under the alternative. If these restrictions are not statistically meaningful this ratio should be close to one and the difference in the log-likelihoods should be small.

Only the Wald test is covered here. The LR and LM tests will be seen later in this course.

1.8.1 T-tests

Before plunging into general tests with many constraints on many parameters, t-tests, a special cases of a Wald tests, warrant independent treatment. T-tests can be utilized to test a single hypothesis involving one or more coefficients,

$$\mathsf{H}_0 : \mathbf{R}\beta = r$$

where $\mathbf{R}$ is a 1 by K vector. Recall that $\widehat{\beta}-\beta \sim \mathsf{N}\left(\mathbf{0},\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$. Noting that $\mathbf{R}\left(\widehat{\beta}-\beta\right) = \mathbf{R}\widehat{\beta} - \mathbf{R}\beta = \mathbf{R}\widehat{\beta} - r$ and applying the properties of normal random variables,

$$\mathbf{R}\widehat{\beta} - r \sim \mathsf{N}\left(0,\sigma^2\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)$$

A simple test that the null hypothesis is true, $\mathbf{R}\beta - r = 0$, can be constructed

$$z = \frac{\mathbf{R}\widehat{\beta} - r}{\sqrt{\sigma^2\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'}}$$

Where

$$z \sim \mathsf{N}\left(0,1\right)$$

To perform a test with size α, the value of z can be compared to the critical values of the standard normal and rejected if $|z| > q_{1-\alpha}$ where $q_{1-\alpha}$ is the $1-\alpha$ quantile. However, z is an infeasible statistic since it depends on an unknown quantity, σ^2. The natural solution is to replace the unknown parameter with an estimate. Dividing z by $\sqrt{s^2/\sigma^2}$ and simplifying,

$$t = \frac{\mathbf{R}\widehat{\beta} - \mathbf{r}}{\sqrt{s^2\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'}}$$

Note that the ratio,

$$(N-K)\sqrt{\frac{s^2}{\sigma^2}} \sim \chi^2_{N-K}$$

so t is the ratio of a standard normal to the square root of a χ^2_ν normalized by its standard deviation. As long as the standard normal in the numerator and the χ^2_ν are independent, this ratio will have a Student's T distribution.

Definition 4 (Student's T distribution) *Let $z \sim \mathsf{N}(0,1)$ (standard normal) and let $w \sim \chi^2_\nu$ where z and w are independent. Then*

$$\frac{z}{\sqrt{w/\nu}} \sim T_\nu$$

Under assumption 1.5 $\widehat{\beta}$ is a function of $\mathbf{X}$ and hence is independent of $\widehat{\epsilon}$ and hence $\widehat{\beta}$ and s^2, which is only a function of $\widehat{\epsilon}$, are independent and t has a Student's T distribution.

Theorem 1.9 (T-test) *Under assumptions 1.1-1.6*

$$\frac{\mathbf{R}\widehat{\beta} - r}{\sqrt{s^2\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'}} \sim T_{N-K}$$

As $\nu \to \infty$, the Student's T distribution converges to a standard normal. As a practical matter, when $\nu > 30$, the T distribution is very close to a normal. Any single linear restriction can be tested with a t-test. The expression t-stat has become synonymous with a specific null hypothesis.

Definition 5 (t-stat) *The t-stat of a coefficient, β_k, is the t-test value of a test of the null H_0 : $\beta_k = 0$ against H_1 : $\beta_k \neq 0$, and is computed*

$$\frac{\widehat{\beta}_k}{\sqrt{s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}_{[kk]}}}$$

where $\left(\mathbf{X}'\mathbf{X}\right)^{-1}_{[kk]}$ indicates the kth diagonal element of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

T-tests is also unique among the three main classes of test statistics in that they can be applied against both **one-sided alternatives** and **two-sided alternatives**. The previous examples were all two-sided; the null would be rejected if the parameters differed in either direction from the null hypothesis. However, there is often a good argument to test a one-sided alternative. For instance, in tests of the market premium, theory indicates that it must be positive to induce investment. Thus, when testing the null hypothesis that a risk premium is zero, a two-sided alternative could reject in cases which are not theoretically interesting. More importantly, a one-sided alternative, when appropriate, will generally have more power than a two-sided alternative; the direction information in the null hypothesis can be used to tighten confidence intervals. The two types of tests involving a one-sided hypothesis are upper tail tests which test nulls of the form H_0 : $\mathbf{R}\beta \leq r$ against alternatives of the form H_1 : $\mathbf{R}\beta > r$, and lower tail tests which test H_0 : $\mathbf{R}\beta \geq r$ against H_0 : $\mathbf{R}\beta < r$.

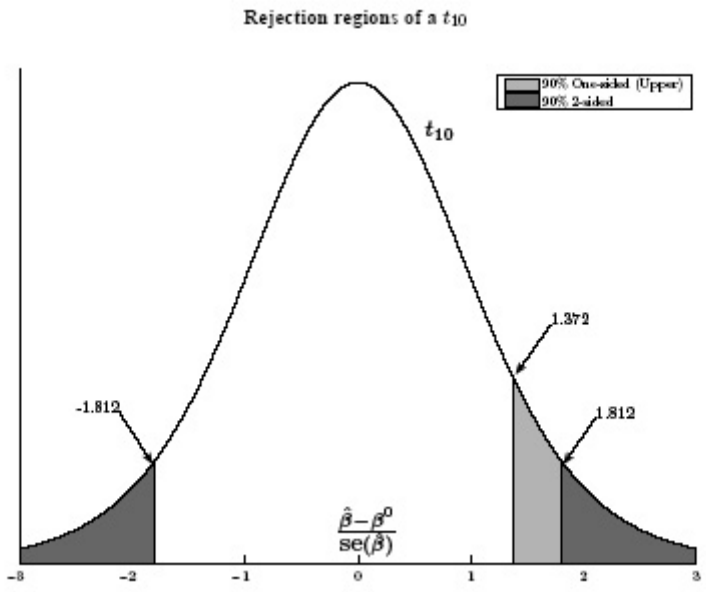

Figure 1.1: Rejection region for a t-test of the nulls $\mathsf{H}_0 : \beta = \beta_0$ (two-sided) and $\mathsf{H}_0 : \beta \leq \beta_0$. The two-sided rejection region is indicated by dark grey while the one sided (upper) rejection region includes both the light and dark grey areas in the right tail.

Figure 1.1 contains the rejection regions of a t_{10} distribution. The dark grey region corresponds to the rejection region of a two-sided alternative to the null that $\mathsf{H}_0 : \beta = \beta_0$ for a 10% test. The light grey region, combined with the upper dark grey region corresponds to the rejection region of a one-sided upper tail test. Evidenced in the picture, and test statistic between 1.372 and 1.812 would be rejected using a one-sided alternative but with a two-sided one. Thus, to use a T test to conduct a test:

- Compute $\widehat{\beta}$ and $s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ using least squares.
- Compute $t = \frac{\mathbf{R}\widehat{\beta}-r}{\sqrt{s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}}$
- Compare t to the critical value, c_α, of the T_{N-K} distribution at for a test size α. In the case of a two tailed test, reject null hypothesis if $|t| > c_\alpha = q_{1-\alpha/2}$. In the case of a one-sided upper tail test, reject if $t > c_\alpha = q_{1-\alpha}$ and in the case of a one-sized lower tailed test, reject if $t < c_\alpha = q_\alpha$. (q_α is the quantile of the T_{N-K} distribution).

1.8.2 Wald Tests

Wald test directly examine how close, in a statistically meaningful sense, $\mathbf{R}\widehat{\beta}$ is to $\mathbf{r}$. Intuitively, if the null hypothesis is true, then $\mathbf{R}\widehat{\beta} - \mathbf{r} \approx \mathbf{0}$. In the small sample framework, the distribution of

$\mathbf{R}\widehat{\beta} - \mathbf{r}$ follows directly from the properties of normal random variables. Specifically,

$$\mathbf{R}\widehat{\beta} - \mathbf{r} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2 \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)$$

Thus, to test the null $\mathsf{H}_0 : \mathbf{R}\beta = \mathbf{r}$ against the alternative $\mathsf{H}_1 : \mathbf{R}\beta \neq \mathbf{r}$, a test statistic can be based on

$$W_{Infeasible} = \frac{\left(\mathbf{R}\widehat{\beta} - \mathbf{r}\right)' \left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1} \left(\mathbf{R}\widehat{\beta} - \mathbf{r}\right)}{\sigma^2}$$

which has a χ^2_M (it is not straightforward to prove) . However, this statistic depends on an unknown quantity, σ^2, so to operationalize it σ^2 must be replaced with an estimate, s^2

$$W = \frac{\left(\mathbf{R}\widehat{\beta} - \mathbf{r}\right)' \left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1} \left(\mathbf{R}\widehat{\beta} - \mathbf{r}\right)}{s^2}$$

However, the replacement of σ^2 with s^2 has an affect on the distribution of the estimator. First, recall the definition of an F distribution.

Definition 6 (F distribution) *Let $z_1 \sim \chi^2_{\nu_1}$ and $z_2 \sim \chi^2_{\nu_2}$ where z_1 and z_2 are independent. Then*

$$\frac{z_1/\nu_1}{z_2/\nu_2} \sim F_{\nu_1,\nu 2}$$

The conclusion that W follows an F distribution follows directly due to the independence of $\widehat{\beta}$ and $\widehat{\epsilon}$, which in turn implies that $\widehat{\beta}$ and s^2 are independent since s^2 is only a function of $\widehat{\epsilon}$.

Theorem 1.10 (Wald test) *Under assumptions 1.1-1.6*

$$W = \frac{\left(\mathbf{R}\widehat{\beta} - \mathbf{r}\right)' \left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1} \left(\mathbf{R}\widehat{\beta} - \mathbf{r}\right)}{s^2} \sim F_{M,N-K}$$

Analogous to the t_ν distribution, an $F_{\nu_1,\nu 2}$ distribution converged to a $\chi^2_{\nu_1}$ as $\nu_2 \to \infty$. Figure 1.2 contains *failure* to reject regions (FTR) for some hypothetical Wald tests. The shape of the region depends crucially on the correlation between the hypotheses being tested. For instance, the upper left corresponds to testing a joint hypothesis where the tests are independent and have the same variance. In the case, the FTR region is a circle. The bottom right figure shows the FTR region for a highly correlated tests where one has larger variance. Once W has been computed, the test statistic should be compared to a table of $F_{M,N-K}$ and rejected if the test statistic is larger than the critical value.

Figure 1.3 contains an $F_{5,30}$ distribution. Any $W > 2.049$ would lead to rejection of the null hypothesis. The Wald test has a more common expression in terms of the SSE from both the restricted and unrestricted models. Specifically,

$$W = \frac{\left(SSE^R - SSE^U\right)/M}{SSE^U/\left(N-K\right)} = \frac{\left(SSE^R - SSE^U\right)/M}{s^2}$$

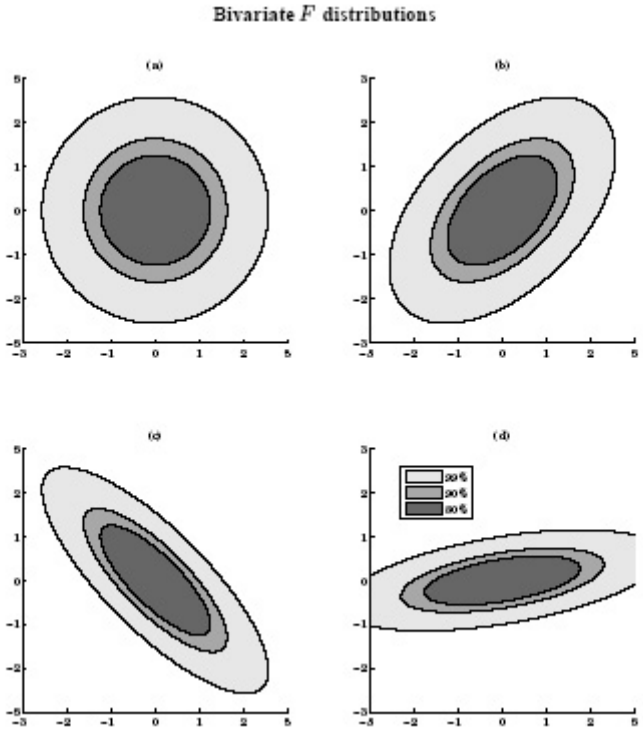

Figure 1.2: Bivariate plot of F distribution. The fox boxes cont the failure-to-reject regions corresponding to 20, 10 and 1% sized tests. Box (a) contains the region for uncorrelated variables. Box (b) contains the region for variables with the same variance but correlation 0.5. Box (c) contains the region for variables with a correlation of -.8 and box (d) contains the region for variables with correlation of 0.5 but with variances of 2 and 0.5 (Variable with variance 2 is along the x-axis).

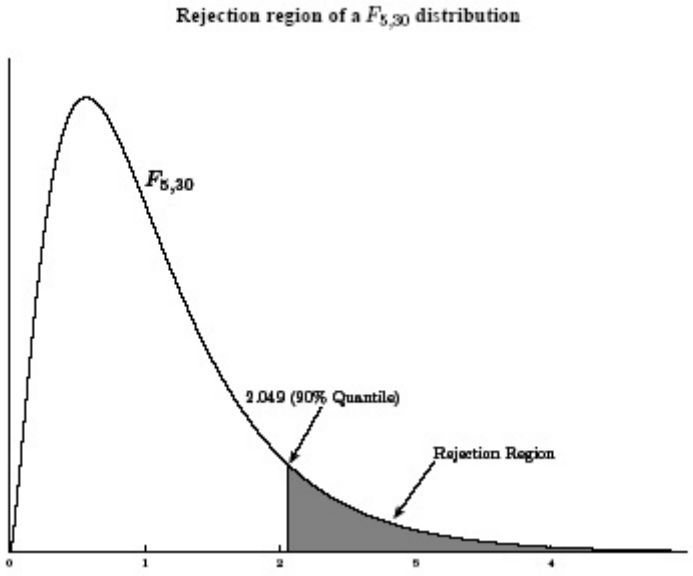

Figure 1.3: Rejection region for an $F_{5,30}$ distribution when using a test with a size of 10%. If the null hypothesis is true, the test statistic should be relatively small (would be 0 if exactly true). Large test statistics lead to rejection of the null hypothesis. In the example, a test statistic with a value greater than 2.049 would lead to a rejection of the null at the 10% level.

While the Wald test technically only requires the unrestricted model to be estimated, this form is useful because it can be computed from the output of any standard regression package. The usefulness of this result is augmented by the observation that any linear regression subject to linear restrictions can be estimated using transformed regressors and unconstrained estimation. The constraint can be directly implemented into the structure of the model.

Thus, to use a Wald statistic to test a hypothesis:

- ComputeSSE^R and SSE^U using least squares.
- Compute $W = \frac{\left(SSE^R - SSE^U\right)/M}{SSE^U/(N-K)}$
- Compare W to the critical value, c_α, of the $F_{M,N-K}$distribution at size α. Reject the null hypothesis if $W > c$.

Finally, in the same way the t-stat is a test that $\mathsf{H}_0 : \beta = 0$ against $\mathsf{H}_1 : \beta \neq 0$, the F-stat of a regression tests whether all coefficients are zero (except the intercept).

Definition 7 (F-stat) *The F-stat of a regression is the value of a Wald test that all coefficient are zero except the coefficient on the constant (if it is included). Specifically, if the model is*

$$y_n = \beta_1 + \beta_2 x_{1n} + ... + \beta_k x_{(k-1)n} + \epsilon_n$$

the F-stat is the value of a Wald test of the null $\mathsf{H}_0 : \beta_2 = ... = \beta_k = 0$ against the alternative $\mathsf{H}_1 : \beta_1 \neq 0$ and corresponds to a test based on the restricted regression

$$y_n = \beta_1 + \epsilon_n$$

Example 3 (T and Wald Tests in Cross-Sectional Factor models) *Returning to the factor regression example, the t-stats of in the general model can be computed*

$$t_j = \frac{\widehat{\beta}_j}{\sqrt{s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{[jj]}}}$$

Consider a regression of BH^e on the set of four factors and a constant,

$$BH^e_n = \beta_1 + \beta_2 VWM^e_n + \beta_3 SMB_n + \beta_4 HML_n + \beta_5 UMD_n + \epsilon_n$$

The fitted coefficients, t-stats and p-values are contained in the table below. The p-value is defined as the probability the data was generated from a model where the null was true. p-values have the advantage that they are independent of the distribution of the test statistic. For example, when using a 2-sided t-test, the p-value of a test statistic t is $2\left(1 - F_{t_\nu}\left(|t|\right)\right)$ where $F_{t_\nu}\left(|t|\right)$ is the CDF of a t-distribution with ν degrees of freedom. In a Wald test, the p-value is $\left(1 - F_{f\nu_1,\nu_2}\left(W\right)\right)$ where $F_{f\nu_1,\nu_2}\left(W\right)$ is the CDF of an $f\nu_1, \nu_2$ distribution.

The critical value, c_α for a 2-sided α =10% t-test with $N - 5 = 938$ degrees of freedom is 1.645. Thus, if $|t| > c_\alpha$ the null hypothesis would be rejected. The data indicate that the null hypothesis for two of the explanatory variables, the constant and SMB, cannot be rejected the 10% level. Also, notice the p-values. The p-values provide the smallest test size where the null hypothesis could be rejected. For example, the null that the constant was 0 could be rejected at an alpha of 16% but not one of 15%.

The table also contains theWald test statistics and p-values for a variety of hypotheses, some economically interesting, such as restriction that the 4 factor model reduce to the CAPM, $\mathsf{H}_0 : \beta_2 = 1$, and $\beta_j = 0$, $j = 1, 3, ..., 5$. Only one regression, the completely unrestricted regression, was needed to compute all of the test statistics using Wald tests, where $\mathbf{R}$ and $\mathbf{r}$ depend on the null being tested. For example, to test whether strict CAPM was likely valid,

$$\mathbf{R} = \mathbf{I}_5 \text{ and } \mathbf{r} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}'$$

All of the null hypotheses save one are strongly rejected with p-values of 0 to three decimal places. The sole exception is $\mathsf{H}_0 : \beta_1 = \beta_3 = 0$, which produced a Wald test statistic of 3.884. The 10% critical value of an $F_{2,938}$ is 2.308 so the null hypothesis would be rejected at the 10% level. The p-value indicates that the test would be rejected at the 15% level but not at the 14% level. One further peculiar number appears in the table. The Wald test statistic for the null $\mathsf{H}_0 : \beta_5 = 0$ is exactly the square of the t-test statistic for the same null. With a little consideration, this shouldn't

be deeply surprising as $W = t^2$ when testing a single linear hypothesis. Moreover, if $z \sim t_\nu$, $z^2 \sim F_{1,\nu}$ can be proven by examining the square and applying the definition of an F-distribution.

t-tests

Parameter	$\widehat{\beta}$	$\widehat{\sigma}_{\widehat{\beta}}$	$t_{\widehat{\beta}}$	p-value
Constant	-0.064	0.0449	-1.435	0.151
VWM^e	1.078	0.0087	123.508	0.000
SMB	0.019	0.0136	1.419	0.155
HML	0.818	0.0131	62.358	0.000
UMD	-0.043	0.0104	-4.218	0.000

Wald tests

Null	Alternative	W	M	p-value
$\beta_j = 0, j = 1, ..., 5$	$\exists j, \beta_j \neq 0$	28701	5	0.000
$\beta_2 = 1; \beta_j = 0, j = 1, 3, ..., 5$	$\beta_2 \neq 1 \cup \exists j, \beta_j \neq 0, j = 1, 3, ..., 5$	5429	5	0.000
$\beta_2 = 1; \beta_j = 0, j = 1, 5$	$\beta_2 \neq 1 \cup \exists j, \beta_j \neq 0, j = 1, 5$	129.9	3	0.000
$\beta_j = 0, j = 1, 5$	$\exists j, \beta_j \neq 0, j = 1, 5$	24.47	2	0.000
$\beta_j = 0, j = 1, 3$	$\exists j, \beta_j \neq 0, j = 1, 3$	3.884	2	0.143
$\beta_5 = 0$	$\beta_5 \neq 0$	17.79	1	0.000

Chapter 2

Deterministic dynamic models

2.1 Time Series Decomposition and Forecasting

2.1.1 What is a Time Series?

Time Series Analysis studies the *dynamics* of a variable. The latter is essential for three reasons:

1. Advances in econometrics have shown that it is only possible to relate variables that exhibit similar properties, in particular as concerns stability (or instability);

2. Mathematical properties of models that allow for estimation of the link between variables depends on their dynamics;

3. Causality or forecasting can only be understood if some pre-determination (i.e. it happens before) is allowed.

Remark 1 *A Time Series is defined as a sequence of observations, or as a sequence of random variables . In this course, we will come to see time series as one or the other but it is important always to remember that they can be both.*

A stochastic process is an ordered sequence of random variables $\{y_t(\omega), \omega \in \Omega, t \in \mathbb{T}\}$, such that for all $t \in \mathbb{T}$, y_t is a random variable on Ω and that for all $\omega \in \Omega$, $y_t(\omega)$ is a realization of the stochastic process on the indexation set $\mathbb{T}$. It is possible to study a time series according to its trajectory, as a function of t (i.e. for a given event ω, we focus on a *realization* of the stochastic process), or to its distribution at a given time t (see figure 2.1).

A time Series is hence **any succession of observations that correspond to the same variable**: it can be e.g. from

- macroeconomics (a country's GDP, inflation, exports...),

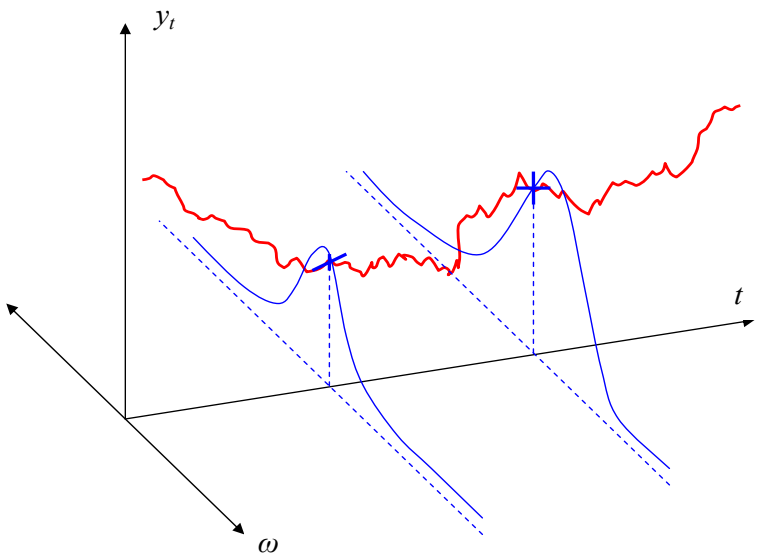

Figure 2.1: Representation of the stochastic process along the three dimensions: both (t, y_t) axes define a trajectory of the process (an observation of the series); the (t, ω) axes represent at each time the density of the r.v. Y_t.

- microeconomics (a firm's sales, total number of employees, a person's income, her/his number of children...),
- financial (S&P500, DJIA, CAC40, the price of a put or call option, a stock return),
- meteorological (rainfall, number of sunny days per year),
- political (voters' turnout, number of votes received by a candidate...),
- demographic (average height in the population, age...).
- electromagnetic (transmission of sound signals, speech modeling, image analysis and transmission)
- geophysical (seismic data, oil extraction...),
- and so on.

In practice anything that can be expressed as numbers and that varies across time or space. The **time dimension** matters in practice as it is then the analysis of a historical chronicle: variations of a unique variable through time in order to understand the dynamics. (Panel data by contrast focus on the variation of characteristics across individuals, agents, firms, assets...).Periodicity does not matter in general: the data can be recorded daily, monthly, quarterly, annually, or even without

periodicity. Notice though that financial trades that are recorded every few seconds can generate some problems (this issue is called market microstructure).

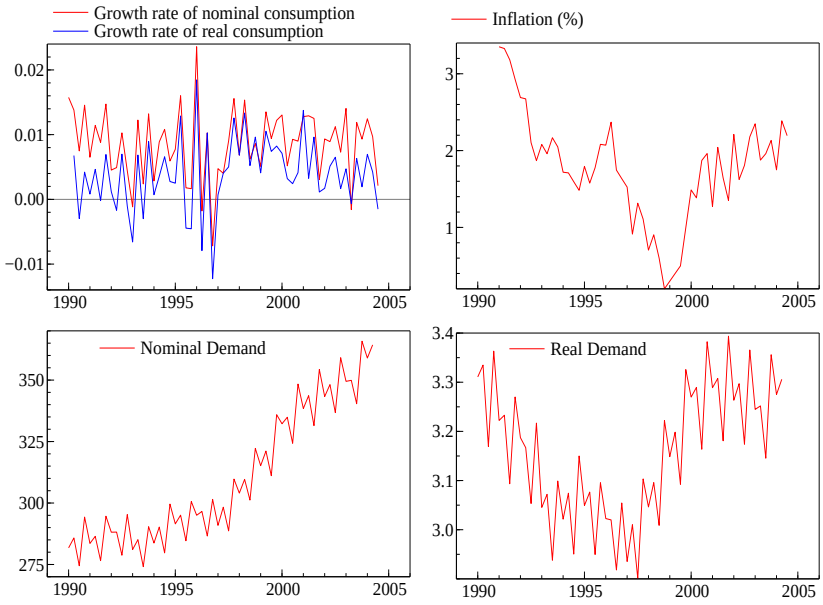

Figure 2.2: French quarterly data for: (a) Nominal and real growth in consumption; (b) Inflation; (c) Nominal demand; (d) Real demand.

Time Series are generally represented on graphs (vertical axis) as a function of time (horizontal axis). Such a representation constitutes an essential tool that allows the experienced modeler directly to observe the key dynamic properties and to be guided through her analysis. Figure 2.2, for instance, presents four graphs of time series with differing properties. Panel (a) *records two series that oscillate about a fixed value (between 0 and 0.01): they are stable about their mean and are called stationary. In* (b)*, inflation decreases strongly until 1999 and then increases: it does not oscillate around a give mean, although it is never away from 2%. Statistical tests should be carried out to check whether it is not less stable than the series in* (a) *, it might hence be non-stationary. Series* (c)*, by contrast, grows throughout the sample, it is said to exhibit a trend and its mean is not constant (its mean between 1990 and 1995 differs strongly from that which is measured between 2000 and 2004). Finally panel* (d) *reproduces the same series without the effect of prices: the upward trend was caused by the increase in prices; real demand actually decreases in the early 1990s. The latter series also presents regularities in each quarter of the year; it is called a seasonal pattern.*

- The characteristics of these graphs can all be modeled and analyzed through time series analysis
- We will introduce later the concepts of seasonality, stationarity, trends that will allow to test

various hypotheses and relate variables.

2.1.2 What are the aims of this analysis?

- **Extract information:** dynamic models allow to analyze evolving tendencies of the data. It is for instance possible to decompose the series for GDP into a long run trend and the position of the business cycle. It is also possible to estimate unobservable components such as *volatility*.

- **Adjust for seasonal variations:** this leads to seasonally adjusted data.

- **Forecast:** this is the key reason why focus on time series data. By observing historical chronicles, we get information about some regularities that can be extrapolated into the future. It is even possible to obtain forecasts that are "robust" against some unexpected events.

2.1.3 How does this work in practice?

- **Aim:** observe the effects of the passage of time (trends, seasonality) to use (forecast, extract information) or correct (seasonal adjustments) some dynamic aspects.

- **Approach:** It is in practice impossible to know the distribution of a time series $\{y_t\}_{t\geq 0}$, we therefore focus on the modeling on the variable y_t **as a function of its past** through its **conditional distribution** (a priori constant through time): for a given $(y_{t-1}, y_{t-2}, ..., y_0)$, what can be said about y_t? We study the density of the variable:

$$y_t | y_{t-1}, y_{t-2}, ..., y_0.$$

conditional on the history of the process: $(y_{t-1}, y_{t-2}, ..., y_0)$.

Note: the conditional expectation corresponds to the expectation of a variable when one already knows the value taken by another variable. For instance, if $X = Y + 3$, with $\mathsf{E}[Y] = 0$, then $\mathsf{E}[X] = 3$. But the expectation of X, given knowledge of the event $\{Y = y\}$ is written as

$$\mathsf{E}[X | \{Y = y\}] = \mathsf{E}[X|Y = y] = y + 3$$

it is a function of y,written $h(y) = \mathsf{E}[X|Y = y]$. We write the function of the random variable. $Y : \mathsf{E}[X|Y] = h(Y)$.

- **Result:** The conditional approach gives an **Error-Prevision Decomposition**, according to which

$$y_t = \mathsf{E}[y_t|Y_{t-1}] + \epsilon_t,$$

$$\text{where} \begin{cases} (i) & \widehat{y}_t = \mathsf{E}\left[y_t | Y_{t-1}\right] \text{ is the component} \\ & \text{of } y_t \text{ that can be predicted} \\ & \text{when the history of the process,} \\ & Y_{t-1} = \{y_{t-1}, y_{t-2}, y_{t-3}, ..., y_0\}, \text{ is known ; and} \\ (ii) & \epsilon_t \text{ represents informations} \\ & \text{that are not modeled:} \\ & \text{themodel error.} \end{cases}$$

2.1.4 Some usual processes

Definition 8 *A **White Noise Process** is a process whose distribution is identical at all times, whose expectation is zero and that is dynamically uncorrelated:*

$$u_t \sim \mathsf{BB}\left(0, \sigma^2\right).$$

Thus, $\{u_t\}$ is White Noise if for all $t \in \mathbb{T}$: $\mathsf{E}[u_t] = 0$, $\mathsf{E}\left[u_t^2\right] = \sigma^2 < \infty$, with u_t and u_{t-h} uncorrelated if $h \neq 0$, t and $(t-h) \in \mathbb{T}$.

Definition 9 *If the White Noise $\{u_t\}$ is Normally distributed, it is then called a **Gaussian White Noise:***

$$u_t \sim \mathsf{NID}\left(0, \sigma^2\right).$$

The hypothèsis of independence is then equivalent to that of uncorrelatedness: $\mathsf{E}[u_t u_{t-h}] = 0$ if $h \neq 0$, t and $(t-h) \in \mathbb{T}$.

Note that the Normality assumption implies dynamic independence. NID processes can be generalized into IID with moments of higher orders that are constant be left undefined.

Definition 10 *A process $\{u_t\}$ whose componentsare independent and identically distributed is written IID:*

$$u_t \sim \mathsf{IID}\left(\mu, \sigma^2\right).$$

All the u_t are drawn from the same distribution with expectation μ and variance σ^2, where u_t and u_{t-h} are independent if $h \neq 0$, t and $(t-h) \in \mathbb{T}$.

Example 4 (Time Series Models) *see figures 2.3 (examples 1-2) and 2.4 (example 3).*

1. *Linear Deterministic Trend:*

$$y_t = \alpha + \beta t + \epsilon_t, \quad \textit{where} \quad \mathsf{E}\left[\epsilon_t\right] = 0.$$

The variable increases or decreases continuously (with small errors at each period). The trend is called deterministic because the value of $(\alpha + \beta t)$ is known in advance ad vitam eternam,it is the expectation of y_t : $\mathsf{E}\left[y_t\right] = \mathsf{E}\left[\alpha + \beta t + \epsilon_t\right] = \alpha + \beta t$.

2. *Quarterly seasonal model*

$$y_t = \alpha_1 1_{q1} + ... + \alpha_4 1_{q4} + \epsilon_t$$

where 1_{qj} *takes value 1 in quarter* j *and zero elsewhere.*

3. *Autoregressive process of order 1,* $AR(1)$*:*

$$y_t = \alpha y_{t-1} + \epsilon_t,$$
$$\epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right) \text{ (white noise)}$$

The value of y_t *only depends on its predecessor. Its properties are function of* α *which controls the inertia of the process: when* $\alpha = 0$, y_t *becomes* ϵ_t *and does not depend on the past, it is then a white noise; if* $\alpha \in]-1, 1[$, y_t *is stable about zero; if* $|\alpha| = 1$, y_t *is unstable and its variations* $y_t - y_{t-1}$ *are unpredictable; finally is* $|\alpha| > 1$, y_t *is explosive.*

Examples are drawn figures 2.3 (ex 1-3) and 2.4 (ex 4).

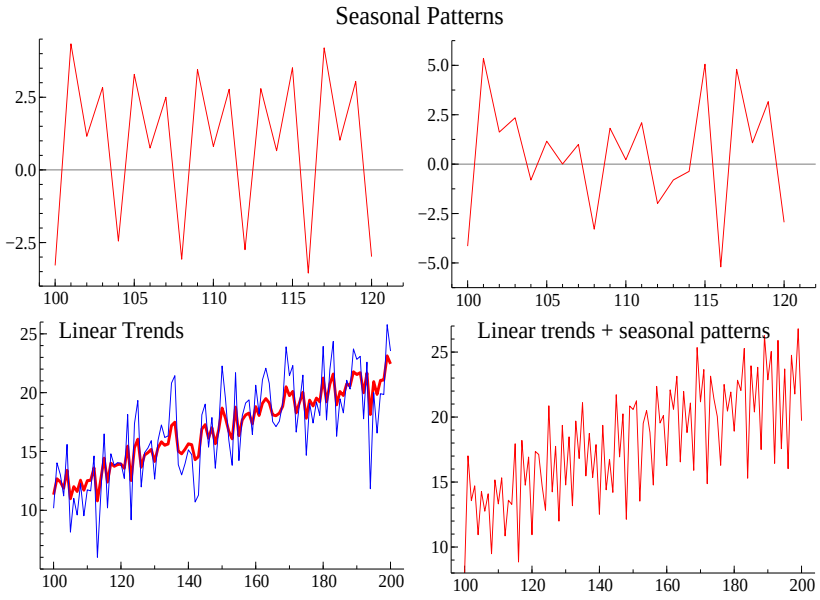

Figure 2.3: Simulated variables presenting (top) a seasonal pattern over a period of four observations. Panel (c) : the two variables present a linear upward trend (with erreurs whose variances differ). In (d) the sum of a trend and a seasonal pattern.

2.2 Decomposition

In this section, we aim to decompose variables into three elements (see figure 2.5 for French internal demand):

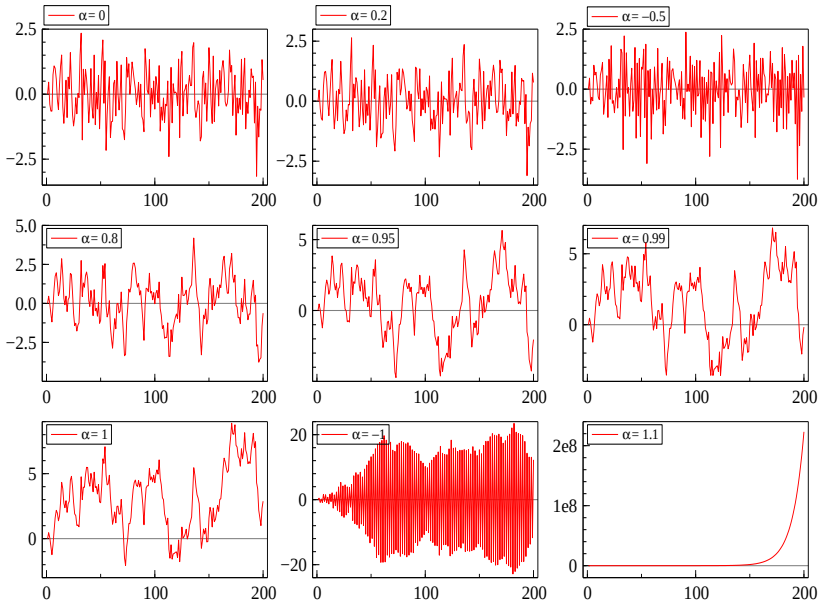

Figure 2.4: Time Series simulated from an AR(1) model $y_t = \alpha y_{t-1} + \varepsilon_t$ for various values of α. Notice the continuity in the characteristics as α tends to unity although the statistical properties of y_t are radically different whether $|\alpha| < 1$ (constant expectation and variance) or $|\alpha| = 1$ ($E[y_t] = E[y_0]$ et $V[y_t] = tV[\varepsilon_t]$). Notice how explosive the series is for $\alpha > 1$.

1. A trend $\mathcal{T}_t$: this is the element that is most persistent and determines the general orientation (upwards, downwards). The tend is *smoother*, less erratic, than the original series. This smoothness is often used to estimate the trend.

2. A seasonal component $\mathcal{S}_t$ that modifies in a regular and predictable fashion the "underlying" behavior.

3. An irregular element $\mathcal{I}_t$ with zero expectation: this element is the main focus of the course.

Example 5 *The quarterly time series generated by*

$$y_t = \alpha + \beta t + \alpha_1 1_{q1} + ... + \alpha_4 1_{q4} + \epsilon_t, \quad o\grave{u} \quad \epsilon_t \sim \mathsf{BB}\left(0, \sigma^2\right)$$

can be decomposed as $y_t = \mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t$ *with*

$$\begin{aligned} \mathcal{T}_t &= \alpha + \beta t, \\ \mathcal{S}_t &= \alpha_1 1_{q1} + ... + \alpha_4 1_{q4}, \\ \mathcal{I}_t &= \epsilon_t. \end{aligned}$$

Here $\mathsf{E}[y_t] = \mathsf{E}[\mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t] = \mathcal{T}_t + \mathcal{S}_t$.

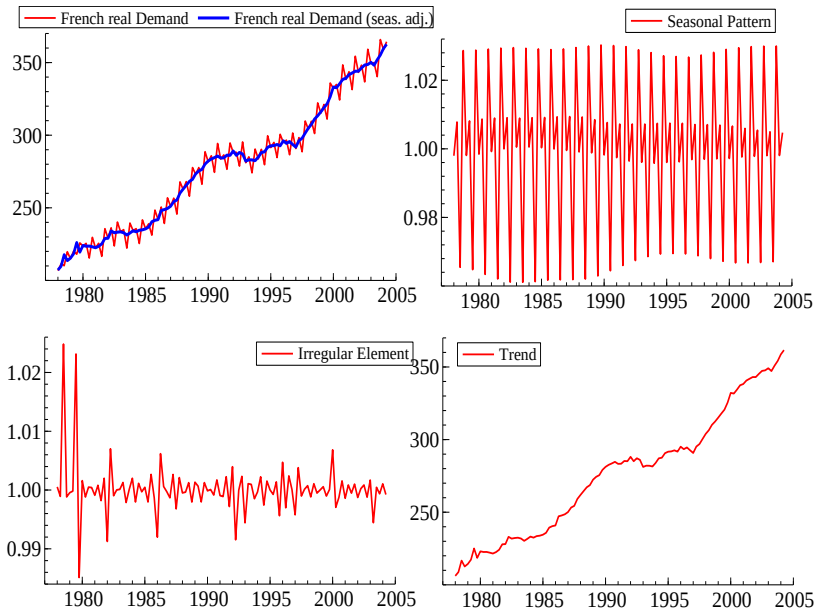

Figure 2.5: Example of decomposition of French internal Demand using the X11 method.

There exist two main ways to decompose time series: additively (A) or multiplicatively (M).

$$\text{A: } Y_t = \mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t$$
$$\text{M: } Y_t = \mathcal{T}_t \times \mathcal{S}_t \times \mathcal{I}_t$$

The difference between these two representations lies in that the trend and seasonal component can be expressed as units of Y_t or percent (see figure 2.6). There are numerous manners to filter the trend and seasonal variations of a series and the results also differ according to the order of extraction ($\mathcal{T}$ then $\mathcal{S}$ or reversely).

2.2.1 Trend Extraction, first case: no seasonality

Two main methods consist in using either a regression or a filtration.

Regression on a deterministic trend.

- The simplest method consists in using a **polynomial function of time**:

$$\mathcal{T}_t = f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + ...\alpha_n t^n \tag{2.1}$$

the (maximal) order n is chosen by visual inspection. We then regress y_t on $\left(1, t, t^2, t^3, ..., t^n\right)$. One should be careful in the choice of n : it should be too large. For any sample of T observations, there exist a polynomial of order $T-1$ that passes through all of $y_1, y_2, ..., y_{T-1}, y_T$. Unfortunately there is no reason why $f(T+1) = \widehat{y}_{T+1}$ should be any close to y_{T+1}. Hence

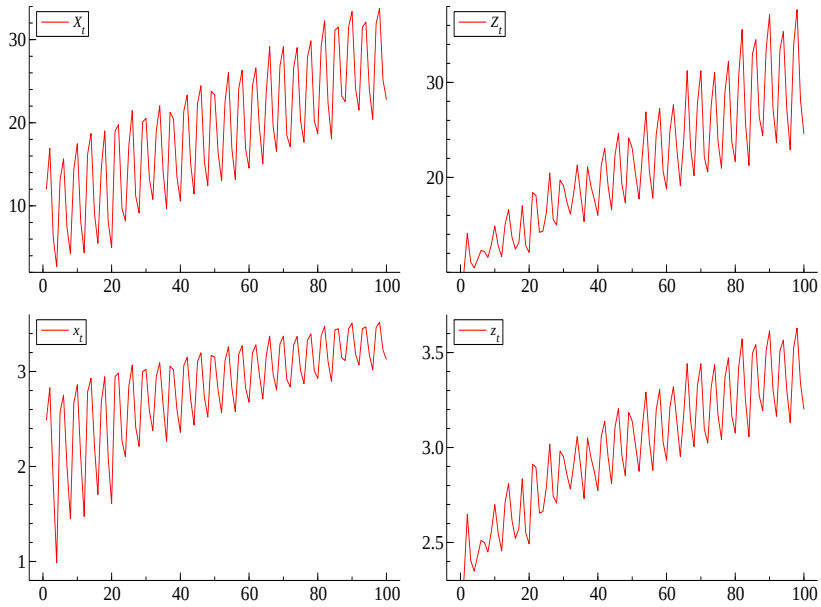

Figure 2.6: Example of series that exhibit linear trends and additive (X_t) or multiplicative (Z_t) seasonal patterns. Séries x_t and z_t correspond to the logarithm of X_t and Z_t : note the how the patterns vary (logarithmic trend, additive seasonality for z_t).

(2.1) does not *describe* y_t, it only *reproduces it* and captures no salient characteristic usable in forecasting. Instead of a polynomial, any function of t can be used.

Moving Average filters

Filters are functions that produce an *output* signal (the filtered variable) from any *input*. Some can extract trends.

- Consider the centered moving average of Y_t of order p, written as $\mathsf{ma}_p^c(t)$, and defined, according to whether p is even or odd as:
  - odd $p = (2q+1)$:

$$\mathsf{mm}_{2q+1}^c(t) : W_t = \frac{1}{2q+1}\sum_{j=-q}^{q} Y_{t-j}, \quad q \in \mathbb{N}^*.$$

  - even $p = 2q$:

$$\mathsf{mm}_{2q}^c(t) : W_t = \frac{1}{2q}\left[\frac{1}{2}Y_{t-q} + \sum_{j=-(q-1)}^{q-1} Y_{t-j} + \frac{1}{2}Y_{t+q}\right], \quad q \in \mathbb{N}^*$$

- The uncentered moving average of Y_t of order p, written as $\mathsf{ma}_p^u(t)$ is defined as

$$\mathsf{mm}_p^d(t) : Z_t = \frac{1}{p}\sum_{j=0}^{p-1} Y_{t-j}.$$

- Then , for $Y_t = \mathcal{T}_t + \mathcal{I}_t$, and p odd, ma_p^c is:

$$W_t = \frac{1}{2q+1}\sum_{j=-q}^{q}\mathcal{T}_{t-j} + \frac{1}{2q+1}\sum_{j=-q}^{q}\mathcal{I}_{t-j}$$

since $\mathcal{T}_{t-j}$ is quite persistent, the centered $\frac{1}{2q+1}\sum_{j=-q}^{q}\mathcal{T}_{t-j} \approx \mathcal{T}_t$ if $\mathcal{T}_t$ is approximately linear (see exercise 6). The moving average of the irregular element is close to its expectation (law of large numbers) which is zero: $\frac{1}{2q+1}\sum_{j=-q}^{q}\mathcal{I}_{t-j} \approx 0$ and

$$W_t \approx \mathcal{T}_t$$

The greater q is, the closer $\frac{1}{2q+1}\sum_{j=-q}^{q}\mathcal{I}_{t-j}$ is to zero, but if the trend is not exactly linear, then by increasing q, the trend is more strongly "smoothed" and can give a poor estimator of $\mathcal{T}_t$.

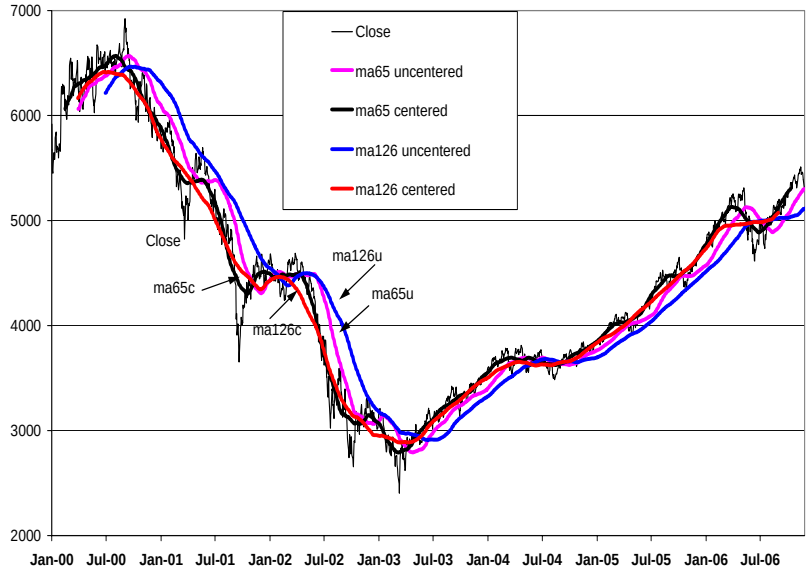

Figure 2.7: Moving Averages on the closing daily level of the (on working days) of CAC40 index (source: yahoo finance). Moving averages are computed over rolling windows of three months (ma65) and six months (ma126). They are either centered or uncentered (using only past or present data at each point in time). Centered moving averages cannot be computed until the end of the sample.

Example 6 *Centered Moving Average of order 3 on* $Y_t = \mathcal{T}_t + \mathcal{I}_t$ *with* $\mathcal{T}_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ *and* $\mathcal{I}_t = \epsilon_t$, *where* $\epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right)$.

1. *Moving average* $W_t = \frac{1}{3}\left(Y_{t-1} + Y_t + Y_{t+1}\right) = \frac{1}{3}\sum_{j=-1}^{1}\mathcal{T}_{t-j} + \frac{1}{3}\sum_{j=-1}^{1}\mathcal{I}_{t-j}$.

2. *For the trend, this becomes*

$$\begin{aligned}\frac{1}{3}\sum_{j=-1}^{1}\mathcal{T}_{t-j} &= \frac{1}{3}\sum_{j=-1}^{1}\left(\alpha_0+\alpha_1(t-j)+\alpha_2(t-j)^2\right)\\ &= \alpha_0+\alpha_1\frac{t-1+t+t+1}{3}+\alpha_2\frac{(t-1)^2+t^2+(t+1)^2}{3}\\ &= \alpha_0+\alpha_1 t+\alpha_2 t^2+\frac{2}{3}\alpha_2\end{aligned}$$

3. *Moving average of* $\mathcal{I}_t$:

$$\frac{1}{3}\sum_{j=-1}^{1}\mathcal{I}_{t-j}=\frac{\epsilon_{t+1}+\epsilon_t+\epsilon_{t-1}}{3}.$$

The Law of Large Numbers yields, when $N\to\infty$, $\frac{1}{2N+1}\sum_{j=-N}^{N}\epsilon_{t-j}\to\mathsf{E}[\epsilon_t]=0$ *since all* ϵ_t *have the same expectation and are independent. We can hence assume that* $\frac{1}{3}\sum_{j=-1}^{1}\mathcal{I}_{t-j}\approx 0$.

4. $W_t=\left(\alpha_0+\frac{2}{3}\alpha_2\right)+\alpha_1 t+\alpha_2 t^2$ *is overestimated for the trend if* $\alpha_2>0$ *since* $W_t-\mathcal{T}_t=\frac{2}{3}\alpha_2$.

- The moving average can also be "weighted" e.g.

$$\overline{Y}_t=\frac{0.5Y_{t-2}+Y_{t-1}+Y_t+Y_{t+1}+0.5Y_{t+2}}{4} \tag{2.2}$$

$$\overline{Z}_t=\frac{Z_{t-2}+2Z_{t-1}+3Z_t+2Z_{t+1}+Z_{t+2}}{9} \tag{2.3}$$

Expression (2.2) is used for seasonal adjustment (see below)..

Hodrick-Prescott Filter

The Hodrick-Prescott filter is commonly used in trend/cycle decompositions, as with the GDP for instance. It consists in generating an output signal ($\mathcal{T}_t$) by minimizing a criterion that weights the proximity to the input and the smoothness (via the second order derivative, $\Delta^2\mathcal{T}_t=\Delta(\mathcal{T}_t-\mathcal{T}_{t-1})=\mathcal{T}_t-2\mathcal{T}_{t-1}+\mathcal{T}_{t-2}$); $\mathcal{T}_t$ is obtained by minimizing the following criterion:

$$\mathcal{T}_t=\underset{g(\cdot)}{\arg\min}\sum_t\left\{(Y_t-g(t))^2+\lambda\left(\Delta^2 g(t)\right)\right\},$$

where λ is chosen according to the periodicity of the the variable ($\lambda=14400$ for monthly data, 1600 for quarterly series, or $\lambda=100$ for annual observations). The difference between the input and output is called cycle $C_t=Y_t-\mathcal{T}_t$. Note that the HP filter uses all the available data: it is very sensitive to the observations on the boundaries of the sample (first and last observations).

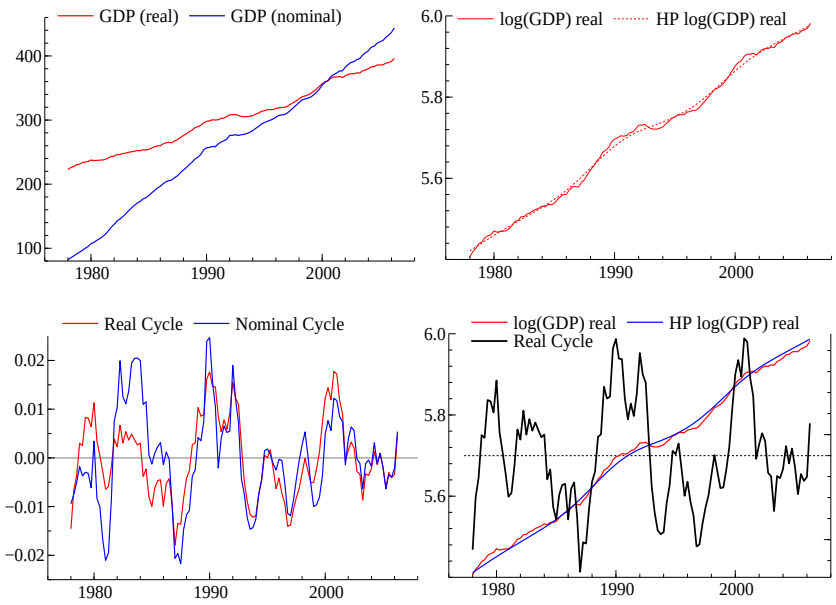

Figure 2.8: French Nominal and Real GDP. The HP filter is applied to the logarithms to get cycles in percent (0.02 indicates a quarterly rate of growth of 2% above the trend rate). The real cycle is computed using real GDP, nominal cycle using nominal GDP. Data source: INSEE.

2.2.2 Trend Extraction, second case: with seasonality

Assume that the seasonal periodicity s is a known integer. Then by definition (in the additive case)

$$\mathcal{S}_{t+s} = \mathcal{S}_t$$
$$\sum_{j=1}^{s} \mathcal{S}_{t+j} = 0, \quad \text{for all } t.$$

and in the quarterly model:

$$Y_t = \mathcal{S}_t + \epsilon_t \tag{2.4}$$
$$\mathcal{S}_t = \alpha_1 1_{q1} + ... + \alpha_4 1_{q4}$$

where 1_{qj} takes value 1 in quarter j and zero elsewhere; it must hold that $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$.

Seasonal adjustment by smoothing in three steps

1. Apply a weighted and centered moving average filter to get a first estimate of the trend

$$X_t = \frac{1}{s}\left(\frac{Y_{t-q} + Y_{t+q}}{2} + \sum_{j=-q+1}^{q-1} Y_{t-j}\right), \text{ if } s = 2q \text{ even.}$$
$$X_t = \frac{1}{s}\sum_{j=-q}^{q} Y_{t-j}, \text{ if } s = 2q+1 \text{ odd.}$$

i.e.

$$X_t = \begin{cases} \left(0.5Y_{t+6} + Y_{t+5} + ... + Y_t + ... + Y_{t-5} + 0.5Y_{t-6}\right)/12 \\ \text{for monthly data} \\ \left(0.5Y_{t+2} + Y_{t+1} + Y_t + Y_{t-1} + 0.5Y_{t-2}\right)/4 \\ \text{for quarterly data} \end{cases}$$

2. Estimate the seasonal component. For each $k = 1, 2, ..., s$, compute the average of $Y_t - X_t$ using only the kth observation k in each period (e.g. the average of the January observations, then that of the February and so on). Denote these averages ω_k. Let then

$$\widehat{\mathcal{S}}(k) = \omega_k - \sum_{i=1}^{s} \omega_i,$$

so that $\sum_{i=1}^{s} \widehat{\mathcal{S}}(k) = 0$. In each period (e.g. each year n) the series Y_{ns+k} equals the seasonally adjusted (SA) series $\widehat{Y}_{ns+k}$ to which $\widehat{\mathcal{S}}(k)$ is added. We therefore let for all $t = ns + k$, with $1 \leq k \leq s$:

$$\mathcal{S}_t = \mathcal{S}_{ns+k} = \widehat{\mathcal{S}}(k).$$

and the seasonally adjusted series is

$$\widehat{Y}_t = Y_t - \mathcal{S}_t.$$

3. Compute the trend using the seasonally adjusted series as seen above (no seasonality).

Example 7 *In the example of expression (2.4), $s = 4$. $X_t = \left(0.5Y_{t+2} + Y_{t+1} + Y_t + Y_{t-1} + 0.5Y_{t-2}\right)/4$ is given by*

$$X_t = \frac{1}{2}\mathcal{S}_{t+2} + \mathcal{S}_{t+1} + \mathcal{S}_t + \mathcal{S}_{t-1} + \frac{1}{2}\mathcal{S}_{t-2} + \frac{1}{2}\epsilon_{t+2} + \epsilon_{t+1} + \epsilon_t + \epsilon_{t-1} + \frac{1}{2}\epsilon_{t-2}$$

where at each point in time S_t equals $\alpha_1 1_{q1} + ... + \alpha_4 1_{q4}$,hence

$$X_t = (\alpha_1 1_{q1} + ... + \alpha_4 1_{q4}) + \frac{1}{2}\epsilon_{t+2} + \epsilon_{t+1} + \epsilon_t + \epsilon_{t-1} + \frac{1}{2}\epsilon_{t-2}$$

and on a sample $t = 1, 2, ..., T$ (we assume that T is a multiple of 4 to simplify the notation), then for all $j \in \{1, 2, 3, 4\}$

$$\omega_j = \alpha_j + \frac{4}{T-1}\sum_{i=1}^{T/4}\left\{\frac{1}{2}\epsilon_{4i+j+2} + \epsilon_{4i+j+1} + \epsilon_{4i+j} + \epsilon_{4i+j-1} + \frac{1}{2}\epsilon_{4i+j-2}\right\}$$

$$\begin{aligned} \sum_{j=1}^{4}\omega_j &= \frac{4}{T-1}\sum_{j=1}^{4}\alpha_j \\ &\quad + \frac{4}{T-1}\sum_{j=1}^{4}\sum_{i=1}^{T/4}\left\{\frac{1}{2}\epsilon_{4i+j+2} + \epsilon_{4i+j+1} + \epsilon_{4i+j} + \epsilon_{4i+j-1} + \frac{1}{2}\epsilon_{4i+j-2}\right\} \\ &\approx \frac{16}{T-1}\sum_{t=1}^{T}\epsilon_t \end{aligned}$$

as $\sum_{j=1}^{4} \alpha_j = 0$, *since* ϵ_t *has zero expectation, the law of large numbers shows that* $\sum_{t=1}^{T} \epsilon_t \approx 0$, *hence* $\widehat{\mathcal{S}}(j) \approx \alpha_j$. *the smoothed series is close to* ϵ_t.

Other methods

- There exist many methods to correct for seasonal patterns, the most commonly used is called x11 (or x12) and was designed by the U.S. Census Bureau. It allows for an evolution of seasonal patterns through time and it also corrects for the number of working days (essential in France where bank holidays are not shifted to the closer week-end). An additional working day corresponds to about 1/250 the annual activity, i.e. 0.4%. This is very significant in a country where GDP grows by about 2% per year on average.

- Multiplicative method: very similar to the additive method presented above but using geometric means and ratios rather than arithmetic means and differences. X_t is computed in the same way, then $r_t = Y_t/X_t$, seasonal indices are ω_i, adjusted as $\omega_m / \sqrt[12]{\omega_1\omega_2...\omega_{12}}$ to get $\widehat{\mathcal{S}}(j)$ for instance for monthly series so that the product of the $\widehat{\mathcal{S}}(j)$ equals unity. The seasonally adjusted series is Y_t/S_t.

2.3 Forecasting

In order to forecast time series, the most sensible method consists in extrapolating from past behavior, hoping that they keep on. Some methods are better than others though!

The simplest method consists using the decomposition model presented above and prolong its behavior assuming that "unpredictable" variables are set to their expectations.

- For instance, assuming that we dispose of a sample of T observations that follow a white noise: $\epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right)$, if we wish to obtain forecasts of the future values of ϵ_{T+h} for the forecast horizons $h \geq 1$ from a forecast origin at T, which we write $\widehat{\epsilon}_{T+h|T}$, then

$$\widehat{\epsilon}_{T+h|T} = \mathsf{E}\left[\epsilon_{T+h}\right] = 0.$$

  and a confidence interval at probability 95% around $\widehat{\epsilon}_{T+h|T}$ is $\pm 1.96\sigma$ in the case of Gaussian white noise.

- If we wish to forecast, from T, and at horizon $h \geq 1$ a variable that is linearly trending, $y_t = \alpha + \beta t + \epsilon_t$, it is natural to choose

$$\widehat{y}_{T+h|T} = \mathsf{E}\left[\alpha + \beta\left(T+h\right) + \epsilon_t\right] = \alpha + \beta\left(T+h\right)$$

  and we notice that

$$\widehat{y}_{T+h|T} = \widehat{y}_{T+h-1|T} + \beta$$

Forecasts can therefore be defined by iterations from a model:

$$y_t = f\left(t, y_{t-1}, y_{t-2}, ..., y_0; \beta\right) + \epsilon_t \tag{2.5}$$

where β represents a vector of parameters. The forecast is the product of an estimation $\widehat{\beta}$ and a model $f(\cdot)$ as

$$\widehat{y}_{T+1|T} = f\left(T+1, y_T, y_{T-1}, ..., y_0; \widehat{\beta}\right)$$

where by iterations, where future values are replaced by their forecasts

$$\widehat{y}_{T+h|T} = f\left(T+h, \widehat{y}_{T+h-1|T}, \widehat{y}_{T+1|T}, y_T, ..., y_0; \widehat{\beta}\right)$$

Define the **forecast error** (once it becomes available, i.e. at time $T+h$) the difference

$$\widehat{e}_{T+h|T} = y_{T+h} - \widehat{y}_{T+h|T}$$

whose properties depend on whether the model is adequate, i.e. if (2.5) is *well specified* (e.g. if ϵ_t is effectively white noise) then

$$\widehat{e}_{T+h|T} = f\left(t, y_{t-1}, y_{t-2}, ..., y_0; \beta\right) - f\left(T+h, \widehat{y}_{T+h-1|T}, \widehat{y}_{T+1|T}, y_T, ..., y_0; \widehat{\beta}\right) + \epsilon_t,$$

ideally

$$\mathsf{E}\left[\widehat{e}_{T+h|T}\right] = 0.$$

Owing to the (e.g. additive) decomposition of time series

$$Y_t = \mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t \tag{2.6}$$

we can generate the forecasts

$$\begin{aligned}\widehat{Y}_{T+h|T} &= E\left[\mathcal{T}_{T+h} + \mathcal{S}_{T+h} + \mathcal{I}_{T+h}\right] \\ &= \mathcal{T}_{T+h} + \mathcal{S}_{T+h}\end{aligned}$$

The forecast error is

$$\widehat{e}_{T+h|T} = Y_t - \widehat{Y}_{T+h|T}.$$

it is clear that the uncertainty around the forecast is related to the variance of $\widehat{e}_{T+h|T}$, but that the latter can only be known via knowledge of the way Y_t is generated. In practice, the uncertainty that surrounds the mechanism at work (the so called **Data Generating Process**) leads us to using (2.6) for Y_{T+h}, i.e.

$$\widehat{e}_{T+h|T} \approx \mathcal{I}_{T+h}$$

and the confidence interval at probability $(1-\alpha)$ about $Y_{T+h} = \widehat{Y}_{T+h|T} + \widehat{e}_{T+h|T}$ is

$$Y_{T+h} \underset{(1-\alpha)}{\in} \left[\widehat{Y}_{T+h|T} + t_{\alpha/2}\sqrt{\mathsf{Var}\left[\widehat{e}_{T+h|T}\right]}, \widehat{Y}_{T+h|T} + t_{1-\alpha/2}\sqrt{\mathsf{Var}\left[\widehat{e}_{T+h|T}\right]}\right].$$

where $t_{\alpha/2}$ is a quantile at the probability $\alpha/2$ from the Student distribution (T degrees of freedom) or from the Normal distribution if $T \geq 30$ and $\mathsf{Var}\left[\widehat{e}_{T+h|T}\right]$ is estimated in-sample (it is different from the variance of the residuals when $h \geq 2$). The uncertainty that surrounds the forecast is therefore related to the second moment of the forecast error (this is called **Mean-Square Forecast Error, MSFE or MSE**), and we can have an estimate thereof using the available sample.

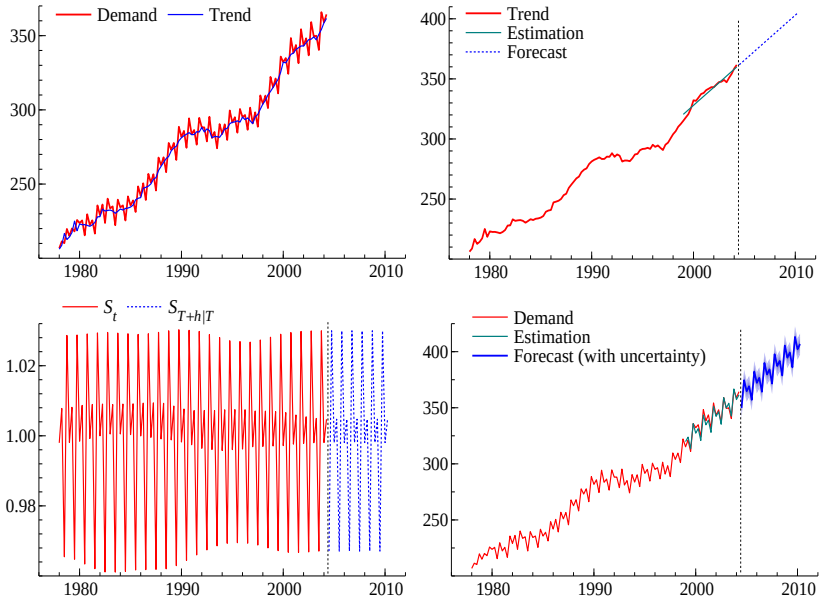

Figure 2.9: Decomposition and forecasts (from 2004Q3) of the French internal Demand. The trend is prolonged by a linear trend estimated over 1999(1)-2004(2). This estimation allows to get (panel d) an estimat and a forecast for the Demand. The uncertainty that surrounds the forecast is represented as blurry area whose amplitude equals $2\times\sqrt{\mathsf{MSFE}}$.

Example 8 *Forecasting using moving averages, example of the exponential smoother*

$$\widehat{y}_{T+1|T} = \alpha \sum_{i=0}^{T} (1-\alpha)^i \, y_{T-i}.$$

then the forecast is the same at all horizons $h \geq 1$:

$$\widehat{y}_{T+h|T} = \alpha \sum_{i=0}^{T} (1-\alpha)^i \, y_{T-i}, \quad \forall h \geq 1.$$

Indeed, take the case $h = 2$: *then if* y_{T+1} *is known,*

$$\begin{aligned} \widehat{y}_{T+2|T} &= \alpha \sum_{i=0}^{T+1} (1-\alpha)^i \, y_{T+1-i} \\ &= \alpha y_{T+1} + (1-\alpha)\, \widehat{y}_{T+1|T} \end{aligned}$$

but for forecasting, we must replace the future value with its forecast and

$$\begin{aligned}\widehat{y}_{T+2|T} &= \alpha\widehat{y}_{T+1|T} + (1-\alpha)\,\widehat{y}_{T+1|T} \\ &= \widehat{y}_{T+1|T}.\end{aligned}$$

by induction

$$\begin{aligned}\widehat{y}_{T+h|T} &= \alpha\widehat{y}_{T+h-1|T} + (1-\alpha)\,\widehat{y}_{T+h-1|T} \\ &= \widehat{y}_{T+h-1|T} = ... = \widehat{y}_{T+1|T}.\end{aligned}$$

Unfortunately, moving averages are limited in their forecasting ability in general.

Chapter 3

Univariate time series: definitions and concepts

3.1 What is a time series? again! to make sure it is clear.

A time series is a sequence of variables ordered in time, e.g. $x_t,\ t = 1, 2, \ldots$, which can be *deterministic*, meaning that the entire path of the process is known with certainty, or *stochastic*. The latter is called a *stochastic process*, that is, a sequence of *random* variables ordered in time. Since it cannot be argued that economic variables are perfectly predictable, they can be represented as stochastic processes. Here, we will consider stochastic processes that are observed discretely, or 'in discrete time' which we can denote as $\{x_t, t \in \mathbb{Z}\}$, where $\mathbb{Z}$ is the set of integers, or equivalently $\{x_t\}_{t=-\infty}^{\infty}$.[1] Very often, we think of the process as having started at a given point in time (usually normalized to 0), so $t = 0, 1, 2, ...$, but there are cases in which we consider processes having started a very long time ago, which we can think of as doubly infinite sequences $t = 0, \pm1, \pm2, ...$ We will often denote processes simply by $\{x_t\}$, omitting the domain of t, except where it is necessary to avoid confusion.

A deterministic time series is any deterministic function of time. For instance, a time trend is the identity mapping $x_t[= x(t)] = t$, or a constant, $x_t = c$. An example of a stochastic time series process is a sequence of i.i.d. standard normal random variables. This process is known as *Gaussian White Noise*.

Many economic variables are observed over time (e.g. most macro-economic aggregates, consumption, GDP, prices), and can therefore be thought of as time series. Economic time series exhibit several distinct patterns, such as temporal dependence (the level of real GDP growth this quarter is likely to correlate with its value in the previous quarter); persistence (GDP growth tends to stay above or below its long-run average for several periods); cycles (e.g., business cycles);

[1] Continuously-observed stochastic processes are typically denoted $x(t)$, where $x(\cdot)$ is a function on the real line.

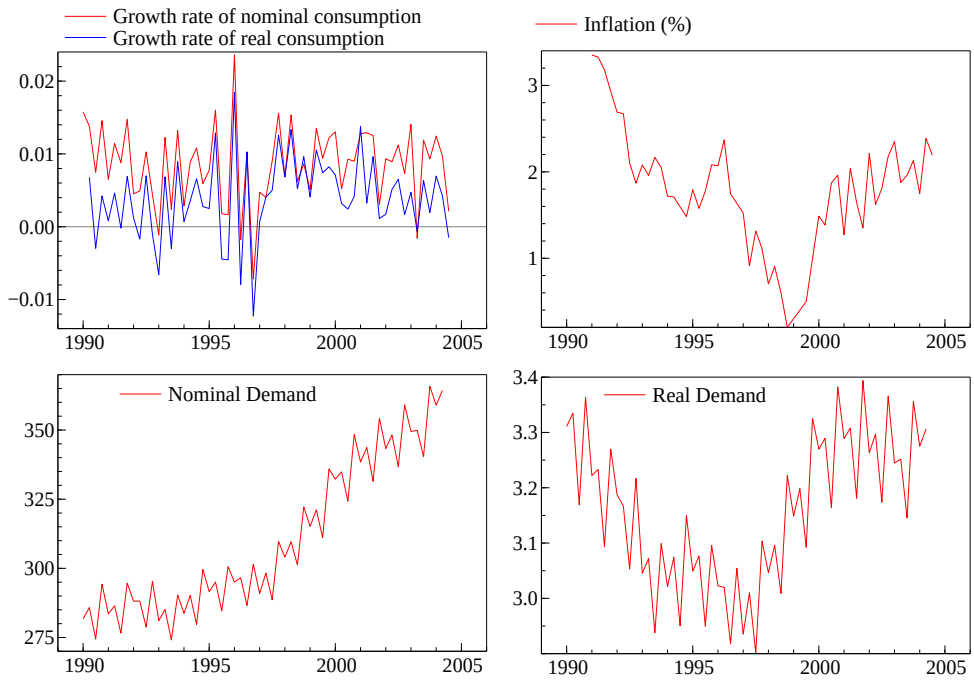

Figure 3.1: French quarterly data: (a) nominal and real growth in domestic consumption; (b) inflation; (c) nominal demand; (d) real demand.

seasonal variation (demand for electricity is higher in winter and summer than in other seasons); trends (per capita income tends to increase over time).

Time series are generally plotted against time. Observing such graphs is invaluable to the experienced modeler: dynamic properties that are noticeable constitute a guide to the analysis. In the plethora of possible statistical tests, it can be useful to have an idea about the possible outcomes: guidance from the data is welcome.

In figure 3.1, four plots show series with different properties. Panel (a) presents two series that oscillate around a value of about 0 to 0.01: they are stable around their mean, we will introduce the concept of stationary series. In (b) inflation decreases until 1999 and goes back up afterwards: it does not oscillate around a constant mean, yet it is never very far from 2%. Conditional on proper statistical testing, this series seems less stable than seen in panel (a), it might be non-stationary. The series in (c) is globally increasing over the whole sample: it is upward *trending*, and its mean is non constant (it mean over 1990-95 differs radically from that over 2000-04). Finally, panel (d) exhibit the same series, corrected for the price increase over the period: the trend in nominal demand was due to inflation, and in real terms, demand decreases in the early 1990s. Moreover, this series shows a regular pattern; quarterly levels reproduce year in, year out: this is a seasonal pattern.

To model economic time series adequately, we need to develop models that can explain those

features. Alternatively, we could, to some extent, transform our data so that certain of those special features disappear (e.g., seasonal adjustment, de-trending transformations). Such a practice is common, but one needs to be aware of the fact that transformations may have unwanted consequences if done incorrectly. In general, data-dependent transformations may be described as a sequential estimation problem that we studied earlier (e.g., removing the sample mean from the data prior to estimating a linear regression), and in that case, due allowance must be made for the uncertainty induced in estimating the transformed variables (this is, for instance, relevant in seasonal adjustment). In other cases, transformations may be actually misleading (e.g., generating a measure of output gap, the deviation of real output from potential, by subtracting a deterministic trend from a real GDP series: this assumes potential output is perfectly predictable, which is rather unrealistic).

3.2 Deterministic Linear Difference equations

The outcomes of the actions of economic agents typically take time to materialize. For instance, an unanticipated increase in the policy rate by the Fed may take a few months or quarters before it impacts on inflation and output. Hence, when studying, say, the marginal effect of a particular variable w on an another variable y, (e.g. a change in interest rates on inflation) we need to use a dynamic model. Such a model can take the form of a difference equation in y_t given w_t, say. For instance, suppose that y_t is generated by the following model:

$$y_t = \phi y_{t-1} + w_t. \tag{3.1}$$

This is a *linear first-order difference equation.* Though it is an unrealistic assumption for economic time series, we will start by assuming that w_t (and hence y_t) is a deterministic sequence in order to discuss the mechanics of deterministic difference equations. In particular, our goal is to characterize the effects on y of changes in the value of w.

3.2.1 Solving a difference equation

Assuming that equation (3.1) holds at all dates $t = 0, 1, 2, ...$ say, and that we know its starting value y_{-1}, say, we can solve it for y_t given w_t by recursive substitution as follows:

$$\begin{aligned} y_0 &= \phi y_{-1} + w_0 \\ y_1 &= \phi y_0 + w_1 \\ &= \phi\left(\phi y_{-1} + w_0\right) + w_1 \\ &\vdots \\ y_t &= \phi^{t+1} y_{-1} + \phi^t w_0 + \phi^{t-1} w_1 + ... + \phi w_{t-1} + w_t \\ &= \phi^{t+1} y_{-1} + \sum_{j=0}^{t} \phi^j w_{t-j}. \end{aligned} \tag{3.2}$$

Equation (3.2) expresses the entire path of the sequence $\{y_t, t = 0, 1, ...\}$ as a function contemporaneous and lagged values of $\{w_t\}$, i.e., the history of w up to time t.

3.2.2 Dynamic multipliers

Consider now the effect of a change in w_0 on y_t, holding all other w_t fixed. This is given by

$$\frac{\partial y_t}{\partial w_0} = \phi^t.$$

By the same reasoning, the effect of a change in w_t on y_{t+j} (i.e., j periods ahead), holding everything else fixed (including y_{t-1}) is

$$\frac{\partial y_{t+j}}{\partial w_t} = \phi^j.$$

The quantity $\partial y_{t+j}/\partial w_t$ is called the (jth) *dynamic multiplier* of w_t on y_t. It represents the response of the process $\{y_t\}$ to a temporary change or 'impulse' in w_t. Thus, $\partial y_{t+j}/\partial w_t$ as a function of j is also referred to as the *impulse response function.*

The behavior of impulse response function for the difference equation (3.1) depends on the value of ϕ. If $0 < \phi < 1$, the effect of a change in w_t decays geometrically and monotonically towards zero, while if $-1 < \phi < 0$ the multiplier will alternate sign (the effect is positive 2,4,6,etc. periods ahead, but negative on odd periods). If $|\phi| > 1$, then the dynamic multiplier increases exponentially over time.

If $\partial y_{t+j}/\partial w_t \to 0$ as $j \to \infty$, then the effects of a change in w_t eventually die out. In that case, we say that the $\{y_t\}$ is a stable process. For the difference equation (3.1), the stability condition is $|\phi| < 1$. In the special case that $\phi = 1$, solving for y_{t+j} given y_{t-1} yields

$$y_{t+j} = y_{t-1} + \sum_{i=0}^{j} w_{t+i}.$$

This shows that $\partial y_{t+j}/\partial w_t = 1$ for all j, which implies that shocks last for ever. A transitory one-unit increase in w_t will cause a permanent one-unit shift in $\{y_t\}$.

The effects of a permanent shift in w_t can also be characterized in the above model. A permanent shift can be expressed as

$$\lim_{j\to\infty} \sum_{i=0}^{j} \frac{\partial y_{t+j}}{\partial w_{t+i}} \tag{3.3}$$

which for model (3.1) is given by

$$\lim_{j\to\infty} \sum_{i=0}^{j} \phi^{j-i} = \sum_{i=0}^{\infty} \phi^i = \frac{1}{1-\phi}$$

provided that infinite sum converges. This happens when $|\phi| < 1$.

3.2.3 pth order difference equations

A pth order difference equation

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + w_t \tag{3.4}$$

can be written as a first order difference equation by defining the vector processes

$$\underset{p\times 1}{\xi_t} = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix}, \underset{p\times 1}{v_t} = \begin{pmatrix} w_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and the *companion matrix*

$$\underset{p\times p}{F} = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_p \\ 1 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{3.5}$$

such that (3.4) can be written as

$$\xi_t = F\xi_{t-1} + v_t. \tag{3.6}$$

Solving the *p*th order difference equation (3.4), is equivalent to solving the first-order difference equation (3.6) for the vector ξ_t as a function of ξ_{-1} and $\{v_t\}$,

$$\xi_t = F^{t+1}\xi_{-1} + \sum_{i=0}^{t} F^i v_{t-i}$$

which eventually yields $\{y_t\}$ as a function of $y_{-1}, ..., y_{-p}$ and $\{w_t\}$. Notice that we now need to specify p starting values to get the solution.

The dynamic multipliers can be found as before, by noting that $\partial y_{t+j}/\partial w_t = \partial \xi_{1,t+j}/\partial v_{1,t}$ is the element $(1,1)$ of the matrix

$$\frac{\partial \xi_{t+j}}{\partial v_t} = F^j,$$

usually denoted $\left(F^j\right)_{11}$ (but $f_{11}^{(j)}$ in Hamilton's notation). As before, the stability of the process depends on the behavior of F^j as j rises. If $\lim_{j\to\infty} F^j = 0$, then $\{y_t\}$ is stable; if it is non-zero but bounded, then y_t is unstable, while if it is unbounded, y_t is explosive.

It would be useful to be able to characterize those cases in terms of the parameters ϕ_i of the difference equation (3.4). It turns out that the stability condition depends on the modulus of the eigenvalues of F.

Proposition 1 *If the eigenvalues of the companion matrix F are all less than 1 in modulus, then the process $\{y_t\}$ is stable.*

(for a proof, see Hamilton, chap. 1).

This result is useful, because it provides a straightforward check of stability.

3.3 The lag operator

The *lag operator* L is a function that maps on time series into another time series, and is defined as

$$Lx_t = x_{t-1},$$

where $Lc = c$ for c constant (think of the lag operator as transforming a *whole process* into another). Obviously, applying the lag operator j times on x_t yields x_{t-j}. This is denoted by $L^j x_t = x_{t-j}$. We can also define the *forward* operator as L^{-1}, which is the inverse operation

$$L^{-1}x_t = x_{t+1}.$$

Another useful operator is the *difference operator*, $\Delta = 1 - L$,

$$\Delta x_t = x_t - x_{t-1}.$$

This can be seen as a special case of $1 - \alpha L$, with $\alpha = 1$. More generally, it is useful to define polynomials in the lag operator, e.g.,

$$\alpha(L) = \alpha_0 + \alpha_1 L + \alpha_2 L^2 + \ldots + \alpha_p L^p. \tag{3.7}$$

Infinite-order lag polynomials arise when $p \to \infty$. Infinite-order polynomials are meaningful if, when applied to a bounded sequence $\{x_t\}_{t=-\infty}^{\infty}$, the result is another bounded sequence (a

sequence $\{x_t\}_{t=-\infty}^{\infty}$ is *bounded* if there exists a finite number $\bar{x}$ such that $x_t < \bar{x}$ for all t). A sufficient condition for this to happen is that the series $\sum_{i=0}^{\infty} \alpha_i$ is absolutely convergent:[2]

$$\sum_{i=0}^{\infty} |\alpha_i| < \infty. \tag{3.8}$$

3.3.1 Application to linear difference equations

The difference equation (3.1) can be written, using the lag operator as:

$$(1 - \phi L)\, y_t = w_t.$$

Consider pre-multiplying both sides of this with the tth order polynomial

$$1 + \phi L + \phi^2 L^2 + \ldots + \phi^t L^t.$$

After a bit of algebra, we can show that

$$\left(1 + \phi L + \phi^2 L^2 + \ldots + \phi^t L^t\right)(1 - \phi L) = 1 - \phi^{t+1} L^{t+1},$$

so that the resulting expression would be

$$\left(1 - \phi^{t+1} L^{t+1}\right) y_t = \left(1 + \phi L + \phi^2 L^2 + \ldots + \phi^t L^t\right) w_t.$$

This is exactly the same as the solution to the difference equation (3.2).

Now, if $|\phi| < 1$, then

$$\left(1 + \phi L + \phi^2 L^2 + \ldots + \phi^t L^t\right)(1 - \phi L) \approx 1$$

that is, the identity operator. In that sense, $\left(1 + \phi L + \phi^2 L^2 + \ldots + \phi^t L^t\right)$ is approximately the inverse operator to the lag polynomial $(1 - \phi L)$. We can make that approximation arbitrarily accurate by letting

$$(1 - \phi L)^{-1} = \lim_{s \to \infty} \left(1 + \phi L + \phi^2 L^2 + \ldots + \phi^s L^s\right)$$

which is an infinite-order lag polynomial. Hence, the solution of the model can be written as

$$y_t = \frac{w_t}{1 - \phi L} = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \tag{3.9}$$

This can be interpreted as saying that the process started in the infinite past, or that the initial value y_{-1} can be expressed as a particular function of the history of w_t up to $t = -1$.

[2] This is equivalent to saying that the polynomial $\alpha(z)$ is analytic on or inside the unit circle, i.e. $\alpha(z) < \infty$, for all (complex numbers) $|z| < 1$.

Hence, we have just seen another method of solving the first-order difference equation (3.1) by inverting lag polynomials.

The pth order difference equation can be written as

$$\phi(L) y_t = w_t \tag{3.10}$$

where $\phi(L) = \alpha(L)$ given in (3.7) with $\alpha_0 = 1$ and $\alpha_i = -\phi_i, i = 1, ..., p$.

Evaluated at a complex number z, a lag polynomial $\alpha(z)$ is referred to as the *characteristic polynomial*. Its roots are the solutions of the polynomial equation

$$\alpha(z) = 0.$$

Consider the pth order polynomial (3.7). It can be shown that its (nonzero) roots z_i, say, are the inverses of the eigenvalues λ_i of the companion matrix F given in (3.5) where $\phi_i = -\alpha_i/\alpha_0$. By definition, λ_i are the p roots of the polynomial equation:

$$|\lambda I_p - F| = 0.$$

These roots can be used to *factor* the lag polynomial $\alpha(L)$ into

$$\alpha(L) = \prod_{i=1}^{p} (1 - \lambda_i L). \tag{3.11}$$

Stability

A difference equation in y_t can be expressed in terms of a lag polynomial. The difference equation is stable, (or the process $\{y_t\}$ is stable) if it is bounded for all t (when $\{y_t\}$ is a stochastic process, it is understood that bounded is in an appropriate sense, e.g. in probability, meaning that for any $0 < \delta < 1$, there exists a finite number $B > 0$ such that $pr(|y_t| > B) \leq \delta$). The necessary condition for stability that was stated in proposition 1 was that the eigenvalues of the companion matrix are all less than 1 in modulus, or 'lie inside the unit circle'. This can be equivalently restated in terms of the roots of the characteristic polynomial, which have to be *outside* the unit circle.

Invertibility

'Inverting' a lag-polynomial $\alpha(L)$ means finding another lag polynomial $\beta(L)$ such that $\alpha(L)\beta(L) = 1$ (and similarly for matrix polynomials). Of course, we require that the resulting 'inverse' polynomial, $\beta(z) \equiv \alpha^{-1}(z) = \left(\beta_0 + \beta_1 z + \beta_2 z^2 + ...\right)$ satisfies

$$\sum_{i=0}^{\infty} |\beta_i| < \infty,$$

so that, when applied to a bounded sequence, the result is also a bounded sequence. From the factorization of the lag polynomial (3.11), we see that inverting $\alpha(L)$ is equivalent to inverting

each of the factors $(1 - \lambda_i L)$. From the discussion of the first-order difference equation, we know that invertibility of each of those first-order polynomials will require that $|\lambda_i| < 1$, or equivalently, that $\alpha(z)$ has all of its roots *outside the unit circle*. This is known as the *invertibility condition.*

3.3.2 Long-run multipliers using lag polynomials

The derivation of the long-run multiplier (3.3) for the process $\{y_t\}$ following a pth order difference equation can be simplified by using lag polynomials. When the equation (3.10) is stable, we can find its solution by inverting $\phi(L)$, i.e., letting $\psi(L) = \phi^{-1}(L)$:

$$y_t = \psi(L) w_t.$$

It can be seen that (3.3) is given by $\psi(1)$, that is, the infinite-order polynomial $\psi(z)$ evaluated at $z = 1$. But this is simply $1/\phi(1)$, and hence the long run dynamic multiplier is given by

$$\frac{1}{1 - \phi_1 - ... - \phi_p}.$$

3.4 Stochastic processes

Now, let us turn to stochastic processes. A stochastic process is a collection of random variables ordered in time. Suppose that we have observed as sample of size T from such a process

$$\{y_1, y_2..., y_T\}.$$

We will use Y_t to denote the random variable and y_t a particular observation drawn from the distribution of Y_t.

Such a sample differs in a fundamental way from a sample of observations in a cross section. The key difference is that all the observations come from a single entity (an individual, firm or country) measured at different points in time, and thus they are related. Unlike in cross-sectional samples, **these observations are ordered in a precise way**, and this ordering is likely to matter a lot for inference. Therefore, we have to view the entire sample $\{y_1, y_2..., y_T\}$ as a single draw out of the distribution of all possible paths or histories of Y_t that were likely to occur over that sample. One draw from the distribution of $\{Y_1, Y_2..., Y_T\}$, is not the same as T draws from the distribution of Y_t for any t.

Consider for instance, annual observations on US real GDP over the last 40 years. This would be different from observations of real GDP in the same year from 40 different countries, even if those countries were identical to the US. Except in special cases, there is absolutely no reason to believe that the distribution of GDP, say in 1979 is the same as that in 2006. This is a particular realization of history, and, if we knew the process generating the data, we could conceive of alternative (counterfactual) realizations over the same sample. In other words, we can think of the

entire sample of T observations as a *single draw* from the distribution of all possible histories of US real GDP.

Any finite subsequence of size T ('sample') from a stochastic process $\{Y_t\}_{t=1}^{T}$ can be characterized by a joint probability density function $f_{Y_1 Y_2 \ldots Y_T}(y_1, \ldots, y_T)$.[3] The distribution of any particular point in the series, Y_t is given by the marginal (also referred to as 'unconditional') density function $f_{Y_t}(y_t)$.

The problem of inference can then be expressed in terms of aspects of the aforementioned probability distributions. For instance, we might be interested in the mean of Y_t, if it exists, i.e.

$$\mu_t \equiv \mathsf{E}(Y_t) = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t)\, dy_t.$$

Obviously, if we had a sample of random draws $y_t^{(j)}$ from f_{Y_t}, we could appeal to a LLN and a CLT to estimate $E(Y_t)$ by the sample average of the $y_t^{(j)}$. This is known as the *ensemble average* of Y_t. Such an estimator would be consistent and asymptotically normal, as the number of observations N, say, drawn from f_{Y_t}, grew.

The trouble is that we only have a single draw from f_{Y_t}, y_t, which is simply the realization of the stochastic process at time t. Since history cannot be repeated, there is no sense in which we can ever hope to have repeated (counterfactual) observations on t. (An exception occurs if we actually know f_{Y_t}, and use the computer to draw random numbers from it. That is only relevant in Monte Carlo simulation settings, though.)

So, in order to be able to say anything about μ_t, or any other aspect of the distribution of the data, we need to make use of the single realization $\{y_t\}_{t=1}^{T}$. To be able to do that, we need to impose some restrictions on the density function f. These concern the *dependence* across different points in the sequence, and the *heterogeneity* of their respective distributions.

Suppose that we are interested in estimating μ_t, when we know that $\mu_t = \mu$ for all t. Intuitively, if the random variables $\{Y_t\}$ are highly dependent, there will be very little independent information across the T observations in the sample to estimate μ (in the limiting case when $\{y_t\}$ are perfectly dependent, we have effectively a single observation). So, we need to impose some restrictions on dependence to get consistency (which requires that the information in the data increases with the sample size). On the other hand, if the sample mean μ_t is changing 'too much' over the sample, we will not be able to have several observations in periods in which it is constant. The limiting case is one in which we have no idea how μ_t changes with t, so, effectively we can treat only observation y_t as having come from a sample with mean μ_t. Again, we would not have increasing information, as T increases, so we wouldn't be able to estimate and do inference on μ_t.

[3] We have assumed the Y_t to be continuous random variables (r.v.) for simplicity.

3.4.1 The autocovariance function

Before proceeding, it would be useful to define the second (unconditional) moments of $\{Y_t\}$, namely its variance at each t, and the covariance between two different points in the sequence. This is known as the *autocovariance function* and it defined by

$$\gamma_{h,t} = \mathsf{Cov}\left[Y_t, Y_{t-h}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(y_t - \mu_t\right)\left(y_{t-h} - \mu_{t-h}\right) f_{Y_t Y_{t-h}}\left(y_t, y_{t-h}\right) dy_t dy_{t-h}.$$

Clearly, $\gamma_{0,t} = \mathsf{Var}\left(Y_t\right)$, and we can define the *autocorrelation function* (ACF) by

$$\rho_{h,t} = \frac{\gamma_{h,t}}{\sqrt{\gamma_{0,t}\gamma_{0,t-h}}}.$$

The plot of the ACF is known as the *correlogram,* and it is an essential starting point for the empirical analysis of the process.

The partial autocorrelation function (PACF) is defined similarly as follows. The hth partial autocorrelation is the last coefficient on a hth order autoregression of x_t

$$x_t = a_{1,t}^{(h)} x_{t-1} + \ldots + a_{h,t}^{(h)} x_{t-h} + e_t^{(h)},$$

namely $\widetilde{\rho}_{h,t} = a_{h,t}^{(h)}$, which forms the PACF for $h = 1, \ldots$

3.4.2 Stationarity

Since the random sampling and identical distribution (i.i.d.) assumption is too strong for time series data, we need to replace them with weaker restrictions. Stationarity concerns the 'identical' part of the i.i.d. assumption, i.e., it places restrictions on how heterogeneous (how different) $f_{Y_t}\left(y_t\right)$ can be over t.

There are two main concepts of stationarity. A stochastic process $\{Y_t\}$ is said to be *strictly stationary* if the joint distribution of any collection $(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k})$ is the same as the distribution of the time-shifted collection $(Y_{t_1+h}, Y_{t_2+h}, \ldots, Y_{t_k+h})$, for all h and $(t_1, t_2, \ldots, t_k)$. It is *weakly stationary* (or *second-order stationary*, or *covariance-stationary*), if its first and second moments exist, and they do not depend on time, i.e.,

$$\mathsf{E}\left[Y_t\right] = \mu, \quad \forall t, \tag{3.12a}$$

$$\mathsf{Var}\left[Y_t\right] = \gamma_0, \quad \forall t, \tag{3.12b}$$

$$\mathsf{Cov}\left[Y_t, Y_{t-h}\right] = \gamma_h, \quad \forall t, h. \tag{3.12c}$$

Note that stationarity implies that $\gamma_h = \gamma_{-h}$ for all h. Strict stationarity implies weak stationarity, provided that the first and second moments exist. Weak stationarity need not imply strict stationarity, but for Gaussian processes the two notions coincide. A process is said to be Gaussian, if the joint density of $(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k})$ is normal for any $(t_1, t_2, \ldots, t_k)$.

Weak stationarity implies that the ACF is time-independent. Autocorrelation is also known as *serial correlation*, and we say that a stationary process is serially correlated at order h, if $\rho_h \neq 0$. A process is asymptotically uncorrelated if $\rho_h \to 0$ as $h \to \infty$.

3.4.3 Ergodicity

The autocorrelation function is a natural measure of serial dependence, that is, dependence between different point in the sequence. Hence, dependence restrictions for a stationary process can be imposed directly on ρ_h, or equivalently γ_h. This concerns the "independent" part of the i.i.d. assumption. Of course, we know that, in general, lack of correlation does not imply independence, but sometimes uncorrelatedness is sufficient for consistent estimation and inference. (For example, recall that OLS is consistent for the coefficients of a linear projection).

A covariance-stationary process is said to be *ergodic for the mean* if its *time* average $\overline{y}$ converges in probability to $\mathsf{E}(Y_t)$ as $T \to \infty$. A sufficient condition for this is that the ACF ρ_h decays sufficiently quickly with h, so that it is 'absolutely summable', i.e.

$$\sum_h |\rho_h| < \infty. \tag{3.13}$$

Ergodicity for higher moments is defined analogously. For stationary Gaussian processes, the above condition is sufficient to ensure ergodicity for all moments.

Even though in many cases, stationarity and ergodicity may amount to the same requirement, it is important to bear in mind that they are different. A process can be stationary and non-ergodic (see the example below).

Stronger dependence restrictions can be placed on the joint distribution of the process

$$f_{Y_1 Y_2 \ldots Y_T}(y_1, \ldots, y_T),$$

by considering the *conditional* distribution of Y_t given other values of Y_s in the sequence. For example, a process is Markovian if

$$f_{Y_t|Y_{t-1},Y_{t-2},\ldots}(y_t|y_{t-1}, y_{t-2}, \ldots) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}).$$

This means the distribution of the conditional on its path up to $t-1$, $\{Y_s\}_{s=-\infty}^{t-1}$ only depends on the immediate past, Y_{t-1}.

Example 9 (Stationary and non-ergodic process) *Consider the process*

$$\begin{aligned} x_t &= x_0 + \varepsilon_t, \text{ for } t \geq 1 \\ \varepsilon_t &\sim \mathsf{NID}\left(0, \sigma_\varepsilon^2\right) \\ x_0 &\sim \mathsf{NID}\left(0, \lambda^2\right) \end{aligned}$$

with ε_t and x_0 independent on one another for $t \geq 1$. Then x_t is stationary

$$\begin{aligned}
\mathsf{E}[x_t] &= \mathsf{E}[x_0] + \mathsf{E}[\varepsilon_t] = 0 \\
\mathsf{V}[x_t] &= \mathsf{V}[x_0] + \mathsf{V}[\varepsilon_t] + 2\mathsf{C}[x_0, \varepsilon_t] = \sigma_\varepsilon^2 + \lambda^2 \\
\mathsf{C}[x_t, x_{t-h}] &= \mathsf{C}[x_0 + \varepsilon_t, x_0 + \varepsilon_{t-h}] = \mathsf{V}[x_0] = \lambda^2,
\end{aligned}$$

but non ergodic since

$$\begin{aligned}
\overline{x}_t &= \frac{1}{t+1}\sum_{i=0}^{t} x_i = x_0 + \frac{1}{t+1}\sum_{i=0}^{t} \varepsilon_i \\
&\to^L x_0,
\end{aligned}$$

hence

$$\overline{x}_t \not\to^p \mathsf{E}[x_t] = 0.$$

3.5 Examples of stochastic processes

3.5.1 White Noise

The building block of time series processes is the white noise process. This is the sequence $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ with

$$\begin{aligned}
\mathsf{E}[\varepsilon_t] &= 0, \text{ and} \\
\mathsf{E}[\varepsilon_t \varepsilon_s] &= \begin{cases} \sigma^2, t = s \\ 0, t \neq s \end{cases}
\end{aligned}$$

denoted $\varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right)$. This process is covariance-stationary, and serially uncorrelated. The above definition does not imply serial independence. When this is required, we will write $\varepsilon_t \sim iid\left(0, \sigma^2\right)$ (this involves the strongest dependence and homogeneity restrictions). The Gaussian White noise process (GWN) is such a process, that we can refer to as $\varepsilon_t \sim \mathsf{NID}\left(0, \sigma^2\right)$, where NID stands for 'normally and independently distributed'.

3.5.2 Linear Processes

More general serial dependence patterns can be modeled by linear combinations of white noise processes. These processes are known as linear or *moving average* processes. For instance

$$Y_t = \varepsilon_t + \mu\varepsilon_{t-1}$$

is a first order moving average process, denoted MA(1). This model expresses Y_t as a distributed lag of white noise errors. We can generalize it by including more lags of ε_t, to order q, and this

is called an MA(q) process, or even consider applying an infinite order lag polynomial to ε_t, and the resulting model would be MA(∞). We can also add deterministic terms, such as a constant μ, which will induce a non-zero mean in y_t.

Another type of model is the Autoregressive (AR) process, which is simply a stochastic difference equation for Y_t in terms of ε_t. This has the same form as (3.4) with w_t replaced by the stochastic process ε_t.

The aforementioned models are *univariate* time series models, because they involve a single observable series $\{Y_t\}$, and express it in terms a sequence of unobserved disturbances ε_t. Univariate time series model have some limited appeal from an economic perspective, because they can be viewed as decomposing the variation in y_t in terms of a sequence of 'shocks'. However, for the disturbances ε_t to admit a structural interpretation as demand, supply or policy shocks, typically a multivariate dynamic structural model is needed. Yet, when two variables are to be related, the properties of estimators depend on whether these variables are stationary or not. Univariate analysis is convenient for the study of the variables' dynamic and stability properties. You will see later in the course that most of the univariate analysis is transferable to multivariate, simply replacing scalar parameter with matrices.

Chapter 4

ARMA models

In this chapter, we describe in some detail a class of linear time series models known as Autoregressive Moving Average (ARMA) models. These are pure time series models, i.e., statistical representations of the first conditional moment of a time series, i.e. $\mathsf{E}[Y_t|\left\{Y_{t-i}\right\}, i>0]$. They do not derive from some underlying economic model. Nonetheless, these models provide a useful starting point for the analysis of economic data because they can be used to describe certain aspects of the data (such as time-dependence, cyclical behavior, trends), as well as for forecasting, i.e., making out-of-sample predictions based on historical information. Both of these objectives can be achieved by models that lack any structural, or causal interpretation. Such models are also referred to as *reduced-form* models to distinguish them from *structural* models that aspire to describe the underlying structure of the economic mechanism that generated the data..

4.1 The MA(1) model

The process

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \quad \varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right) \tag{4.1}$$

is called first-order moving average, denoted $\mathrm{MA}(1)$. Its mean and variance are given by:

$$\begin{aligned}
\mathsf{E}\left[Y_t\right] &= \mu + \mathsf{E}\left[\varepsilon_t\right] + \theta\mathsf{E}\left[\varepsilon_{t-1}\right] = \mu \\
\mathsf{V}\left[Y_t\right] &= \mathsf{V}\left[\varepsilon_t + \theta\varepsilon_{t-1}\right] = \mathsf{E}\left[\varepsilon_t^2\right] + \theta^2\mathsf{E}\left[\varepsilon_{t-1}^2\right] + 2\theta\mathsf{E}\left[\varepsilon_t\varepsilon_{t-1}\right] \\
&= \sigma^2 + \theta^2\sigma^2 + 2\theta\, 0 = \left(1+\theta^2\right)\sigma^2.
\end{aligned}$$

Its first autocovariance is given by

$$\begin{aligned}\gamma_1 &= \mathsf{E}\left[(Y_t - \mu)(Y_{t-1} - \mu)\right] \\ &= \mathsf{E}\left[(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-1} + \theta\varepsilon_{t-2})\right] \\ &= \mathsf{E}\left[\varepsilon_t\varepsilon_{t-1} + \theta\varepsilon_{t-1}^2 + \varepsilon_t\theta\varepsilon_{t-2} + \theta^2\varepsilon_{t-1}\varepsilon_{t-2}\right] \\ &= \theta\sigma^2.\end{aligned}$$

Higher-order autocovariances are zero, since

$$\begin{aligned}\gamma_h &= \mathsf{E}\left[(Y_t - \mu)(Y_{t-h} - \mu)\right] \\ &= \mathsf{E}\left[(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-h} + \theta\varepsilon_{t-h-1})\right] \\ &= \mathsf{E}\left[\varepsilon_t\varepsilon_{t-h} + \theta\varepsilon_{t-1}\varepsilon_{t-h} + \varepsilon_t\theta\varepsilon_{t-h-1} + \theta^2\varepsilon_{t-1}\varepsilon_{t-h-1}\right]\end{aligned}$$

and $\mathsf{E}\left[\varepsilon_t\varepsilon_s\right] = 0$ for all $t \neq s$.

4.1.1 The ACF

The MA(1) is clearly covariance-stationary, since its first two moments are time-invariant. Its ACF is given by

$$\rho_h = \frac{\gamma_h}{\gamma_0} = \begin{cases} \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} = \frac{\theta}{(1+\theta^2)}, & h = 1 \\ 0 & h > 1. \end{cases}$$

Moreover, the MA(1) is ergodic, since the ACF satisfies the ergodicity condition (3.13).

One important implication of the MA(1) process is that, for all values of θ, $-0.5 \leq \rho_1 \leq 0.5$. A plot of ρ_1 as a function of θ is given in Figure 4.1.

This fact is useful because it can sometimes help whether an MA(1) is a suitable model for a given series just by looking at its correlogram.

Note that the ACF of the MA remains unchanged if we replace θ by $1/\theta$. In other words, the parameters $\left(\theta, \sigma^2\right)$ are not globally identifiable from the variance and first autocovariance of Y_t, namely γ_0, γ_1. Since only (a sample from) $\{Y_t\}$ is observable, we can typically learn about (γ_0, γ_1) from the observations on Y_t (we will discuss suitable estimators later), and we can then back out $\left(\theta, \sigma^2\right)$ by solving the above equations. The lack of identification of θ means that there will always be two different values of θ, σ^2 corresponding to the same γ_0, γ_1 and we can never distinguish between them on the basis of available information (unless $|\theta| = 1$, in which case the two solutions coincide).

To rule out this identification problem, we will only consider $|\theta| \leq 1$. Of course, we could have equivalently restricted $|\theta| \geq 1$. But there is a reason why $|\theta| \leq 1$ is preferable.

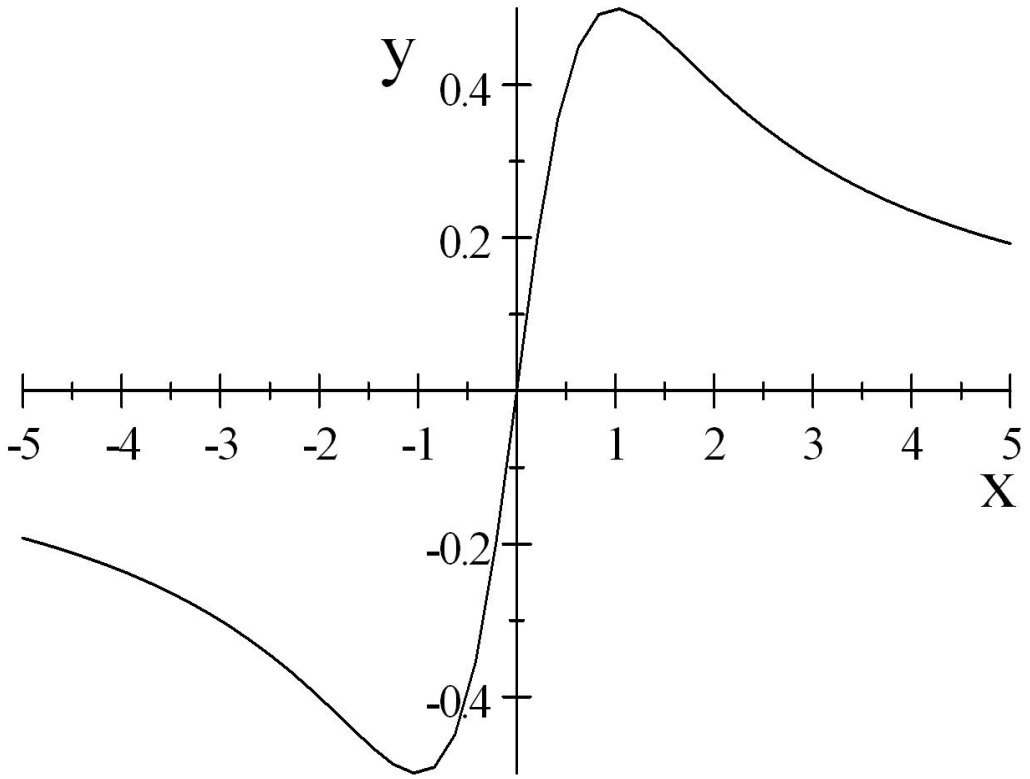

Figure 4.1: θ is on the x-axis, and $\rho_1 = \theta/\left(1+\theta^2\right)$ is on sthe y-axis.

4.1.2 Invertibility and the AR(∞) representation

We can write equation (4.1) using the lag polynomial $(1+\theta L)$, and then think about solving the resulting difference equation for $\{\varepsilon_t\}$. This exercise is useful if we think of $\{Y_t\}$ as an observable series, say, US real GDP, and $\{\varepsilon_t\}$ as a sequence of unobserved shocks that are thought to be driving $\{Y_t\}$. Solving the difference equation (4.1) this way enables us to identify the shocks ε_t from observations on Y_t. This can be done by inverting $(1+\theta L)$, i.e., finding a polynomial $\psi(L) = \left(1+\psi_1 L+\psi_2 L^2+...\right) = (1+\theta L)^{-1}$. The coefficients of the infinite-order lag polynomial $\psi(L)$ can be obtained by matching the coefficients on the powers of the L in the following identity:

$$(1+\theta L)\left(1+\psi_1 L+\psi_2 L^2+...\right) \equiv 1.$$

This yields an infinite set of equations, one for each coefficient of $L^i, i = 1,2,3,...$

$$\begin{aligned} \theta+\psi_1 &= 0 \\ \theta\psi_1+\psi_2 &= 0 \\ &\vdots \end{aligned}$$

Letting $\psi_0 = 1$, all those equations can be written recursively as

$$\psi_i = -\theta\psi_{i-1}$$

which is a difference equation with solution

$$\psi_i = (-\theta)^i \psi_0 = (-\theta)^i .$$

Now, if $|\theta| < 1$, the coefficients of the inverse lag polynomial $\psi(L)$ are all finite, and in fact, geometrically decline to 0. Moreover, the series $\sum \psi_j$ is absolutely summable, i.e., satisfies the condition (3.8). This ensures that applying $\psi(L)$ to a bounded sequence $\{Y_t\}$ produces another bounded sequence. Hence, we see that if $|\theta| < 1$, it is meaningful to invert $(1 + \theta L)$ and write ε_t as an infinite moving average of current and lagged Y_ts

$$\varepsilon_t = (1 + \theta L)^{-1} (Y_t - \mu)$$

or

$$\varepsilon_t = -\frac{\mu}{1+\theta} + \sum_{j=0}^{\infty} (-\theta)^j Y_{t-j} \tag{4.2}$$

where the first term follows by noting that, for any constant c, $\psi(L)\, c = \psi(1)\, c$. The condition $|\theta| < 1$ is known as the *invertibility condition* and an MA(1) representation that satisfies this is called an *invertible* MA(1). Equation (4.2) is called the *infinite-order autoregressive representation* of the process $\{Y_t\}$, and it is denoted AR(∞).

Now, notice that the entire sequence $\{Y_t\}$ also admits a *non-invertible* MA(1) representation, namely, there exist a White Noise series $\{\widetilde{\varepsilon}_t\}$ such that

$$\begin{aligned} Y_t &= \mu + \widetilde{\varepsilon}_t + \theta^{-1}\widetilde{\varepsilon}_{t-1} \\ &= \mu + \left(1 + \theta^{-1}L\right)\widetilde{\varepsilon}_t, \quad \widetilde{\varepsilon}_t \sim \mathsf{WN}\left(0, \widetilde{\sigma}^2\right) \end{aligned} \tag{4.3}$$

Representations (4.1) and (4.3) are *observationally equivalent*, meaning that they both describe the same data generating process for $\{Y_t\}$. The difficulty with the non-invertible representation is that we cannot solve it for $\widetilde{\varepsilon}_t$ using current and lagged values of Y_t, i.e., it does not admit an AR(∞) representation. We could instead solve the difference equation (4.3) by *forward recursion*, i.e., writing it as

$$\begin{aligned} \widetilde{\varepsilon}_{t-1} &= \theta(Y_t - \mu) - \theta\widetilde{\varepsilon}_t \\ &= \theta(Y_t - \mu) - \theta\left[\theta(Y_{t+1} - \mu) - \theta\widetilde{\varepsilon}_{t+1}\right] \\ &\vdots \\ &= \theta \sum_{j=0}^{N} (-\theta)^j (Y_{t+j} - \mu) + (-\theta)^{N+1} \widetilde{\varepsilon}_{t+N} \end{aligned}$$

and observe that if $|\theta| < 1$, this converges to

$$\widetilde{\varepsilon}_t = \theta \sum_{j=0}^{N} (-\theta)^j (Y_{t+j+1} - \mu).$$

This means that $\{Y_t\}$ can be equally characterized in terms of the innovations $\widetilde{\varepsilon}_t$, which can be recovered from future values of Y_t. Since we typically want to identify the innovations in Y_t from historical data, we need to use the invertible MA representation. The innovations associated with the invertible MA representation are referred to as *fundamental innovations*.

4.2 The MA(q) model

The process

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right) \tag{4.4}$$

is called qth-order moving average, denoted MA(q). Let $\theta_0 = 1$ and define the lag polynomial

$$\theta(L) = \theta_0 + \theta_1 L + \ldots + \theta_q L^q \tag{4.5}$$

so that the process can be written as

$$Y_t = \mu + \theta(L)\, \varepsilon_t.$$

Its mean and variance are given by:

$$\begin{aligned}
\mathsf{E}[Y_t] &= \mu + \sum_{j=0}^{q} \theta_j \mathsf{E}[\varepsilon_{t-j}] = \mu, \text{ and} \\
\mathsf{V}[Y_t] &= \mathsf{V}\left[\sum_{j=0}^{q} \theta_j \varepsilon_{t-j}\right] = \mathsf{E}\left[\left(\sum_{j=0}^{q} \theta_j \varepsilon_{t-j}\right)^2\right] \\
&= \sum_{j=0}^{q} \theta_j^2 \underbrace{\mathsf{E}\left[\varepsilon_{t-j}^2\right]}_{=\sigma^2} + 2 \sum_{j=1}^{q} \sum_{i=0}^{j-1} \theta_j^2 \underbrace{\mathsf{E}\left[\varepsilon_{t-j}\varepsilon_{t-i}\right]}_{=0} \\
&= \sigma^2 \sum_{j=0}^{q} \theta_j^2.
\end{aligned} \tag{4.6}$$

4.2.1 The ACF

The autocovariances γ_h are zero at all lags $h > q$, and for $h \leq q$ they are given by

$$\begin{aligned}\gamma_h &= \mathsf{E}\left[(Y_t - \mu)(Y_{t-h} - \mu)\right] \\ &= \mathsf{E}\left[\left(\sum_{j=0}^{q} \theta_j \varepsilon_{t-j}\right)\left(\sum_{j=0}^{q} \theta_j \varepsilon_{t-j-h}\right)\right] \\ &= \mathsf{E}\left[\theta_h \varepsilon_{t-h}^2 + \theta_{h+1}\theta_1 \varepsilon_{t-h-1}^2 + \ldots + \theta_q \theta_{q-h} \varepsilon_{t-q}^2\right] \\ &= \left(\theta_h + \theta_{h+1}\theta_1 + \ldots + \theta_q \theta_{q-h}\right)\sigma^2. \end{aligned} \tag{4.7}$$

Since the first two moments are time-invariant, the MA(q) process is covariance stationary. The autocorrelation function is given by γ_h/γ_0, and so **all autocorrelations beyond q are 0.**

The MA(q) is clearly covariance-stationary, since its first two moments are time-invariant. Moreover, the MA(q) is ergodic, since the ACF satisfies the ergodicity condition (3.13).

4.2.2 Invertibility and AR(∞) representation

If the roots of the lag polynomial (4.5) are all outside the unit circle, then the MA(q) representation (4.4) is invertible. We can find the coefficients of the inverse of $\psi(L) = \theta^{-1}(L)$ as before by matching coefficients in the identity $\psi(L)\theta(L) \equiv 1$. The AR($\infty$) representation would be

$$\varepsilon_t = \theta^{-1}(L)(Y_t - \mu).$$

If any of the roots, z_i, of the characteristic equation $\theta(z) = 0$ were in fact $|z_i| < 1$, the MA(q) process would be non-invertible.

The previous discussion of the identification and observational equivalence readily generalizes to the MA(q) process, namely, for any non-invertible MA(q) representation, we can always find an observationally equivalent representation that is invertible (except if any $|z_i| = 1$). Suppose that we have a noninvertible MA(q) process characterized by $\widetilde{\theta}(L)$ (and associated $\widetilde{\sigma}^2$), and let $|z_i| < 1$ for $i \leq n$, where $n \leq q$ (i.e., order the roots in ascending order), and define $\lambda_i = z_i^{-1}$, so that $\widetilde{\theta}(L)$ can be factored as follows:

$$\widetilde{\theta}(L) = \prod_{i=1}^{n}(1 - \lambda_i L)\prod_{j=n+1}^{q}(1 - \lambda_j L).$$

Then, we can choose an observationally equivalent representation $\theta(L)$ that uses λ_i^{-1} instead of λ_i for $i \leq n$, i.e.

$$\theta(L) = \prod_{i=1}^{n}\left(1 - \lambda_i^{-1} L\right)\prod_{i=n+1}^{q}(1 - \lambda_i L).$$

Clearly $\theta(L)$ is invertible whenever $|\lambda_j| < 1$ for $j > n$. Clearly, there are several observationally equivalent noninvertible MA(q) representations, but at most one invertible representation (none if any $|\lambda_i| = 1$).

4.2.3 MA(∞)

If we let $q \to \infty$ in (4.4) we obtain an infinite order moving average process. Its mean is given by μ, and its variance would be obtained as the limit of (4.6) as $q \to \infty$. Clearly, for the process to have finite variance (this is akin to being bounded in *mean square*) the series $\sum_{i=0}^{\infty} \theta_i^2$ must be convergent. Sufficient for this is that the sequence of MA coefficients $\{\theta_j\}_{j=0}^{\infty}$ is absolutely summable

$$\sum_{i=0}^{\infty} |\theta_i| < \infty. \tag{4.8}$$

The autocovariances of the MA(∞) will be given by (4.7) with $q \to \infty$. Since the first two moments of an MA(∞) are time-invariant, the process is covariance stationary.

It can be shown (see Hamilton p.70) that absolute summability (4.8) implies ergodicity (3.13).

4.3 The AR(1) model

A first-order autoregressive process $\{Y_t\}$, denoted AR(1) is given the first-order difference equation (3.1) with w_t replaced by a constant white noise, i.e.,

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right). \tag{4.9}$$

Using lag operators, we can write this as

$$\left(1 - \phi L\right) Y_t = c + \varepsilon_t. \tag{4.10}$$

Assuming that the process started at $t = -\infty$, we can attempt to solve it in terms of $\{\varepsilon_t\}_{t=-\infty}^{\infty}$. This can be done by inverting the polynomial $(1 - \phi L)$. The condition for this to work is that the same as the stability condition for the solution of the first-order difference equation (3.1), namely $|\phi| < 1$. If that holds, then we can define the infinite-order lag polynomial $\psi(L)$, with coefficients $\psi_j = \phi^j$, and observe that $\{\psi_j\}$ would be absolutely summable. We would then obtain an MA(∞) representation of $\{Y_t\}$

$$\begin{aligned} Y_t &= \left(1 - \phi L\right)^{-1} c + \left(1 - \phi L\right)^{-1} \varepsilon_t \\ &= \frac{c}{1 - \phi} + \psi\left(L\right) \varepsilon_t \end{aligned} \tag{4.11}$$

with $\mu = c/\left(1 - \phi\right)$. Absolute summability of $\left\{\phi^j\right\}$ also implies ergodicity of $\{Y_t\}$.

Existence of an MA(∞) representation implies that $\{Y_t\}$ is covariance stationary. Hence, we can derive its mean variance and autocovariances directly from equation (4.9). The mean is found by taking expectations and letting $\mu = \mathsf{E}\left[Y_t\right] = \mathsf{E}\left[Y_{t-1}\right]$

$$\mu = c + \phi\mu + \mathsf{E}\left[\varepsilon_t\right] \Rightarrow \mu = \frac{c}{1 - \phi}.$$

Similarly, $\gamma_0 = \mathsf{V}[Y_t] = \mathsf{V}[Y_{t-1}]$, and $\mathsf{E}[\varepsilon_t Y_{t-1}] = \mathsf{E}\left[\varepsilon_t\left(\mu + \varepsilon_{t-1} + \phi\varepsilon_{t-2} + \phi^2\varepsilon_{t-3} + \ldots\right)\right] = 0$, so

$$\gamma_0 = \phi^2\gamma_0 + \mathsf{E}\left[\varepsilon_t^2\right] \Rightarrow \gamma_0 = \frac{\sigma^2}{1-\phi^2}. \tag{4.12}$$

4.3.1 The ACF

Next, to derive γ_h we can use the MA(∞) representation (4.11). Alternatively, observe that we can re-write (4.9) as

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + \varepsilon_t$$

multiply both sides by $(Y_{t-h} - \mu)$ and take expectations on both sides to deduce that

$$\begin{aligned}
\gamma_h &= \mathsf{E}[(Y_t - \mu)(Y_{t-h} - \mu)] \\
&= \mathsf{E}[(\phi(Y_{t-1} - \mu) + \varepsilon_t)(Y_{t-h} - \mu)] \\
&= \phi\underbrace{\mathsf{E}[(Y_{t-1} - \mu)(Y_{t-h} - \mu)]}_{\gamma_{h-1}} + \underbrace{\mathsf{E}[\varepsilon_t(Y_{t-h} - \mu)]}_{=0} \\
&= \phi\gamma_{h-1}
\end{aligned}$$

which is a first-order difference equation in $\{\gamma_h\}$, which can be solved for γ_h given γ_0, $\gamma_h = \phi^h\gamma_0$, and using (4.12) yields

$$\gamma_h = \frac{\phi^h}{1-\phi^2}\sigma^2. \tag{4.13}$$

The ACF ρ_h can be readily derived from (4.12) and (4.13):

$$\rho_h = \phi^h.$$

Finally, note that the single restriction $|\phi| < 1$ makes the process $\{Y_t\}$ weakly stationary and ergodic for the mean. Figure 4.2 presents simulated series from an AR(1).

4.4 The AR(p) model

A pth-order autoregressive process $\{Y_t\}$, denoted AR(p) is given by the pth-order difference equation (3.4) with w_t replaced with a constant plus white noise, i.e.,

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right). \tag{4.14}$$

Using lag operators and

$$\phi(L) = (1 - \phi_1 L - ... - \phi_p L^p) \tag{4.15}$$

we can write this as

$$\phi(L) Y_t = c + \varepsilon_t. \tag{4.16}$$

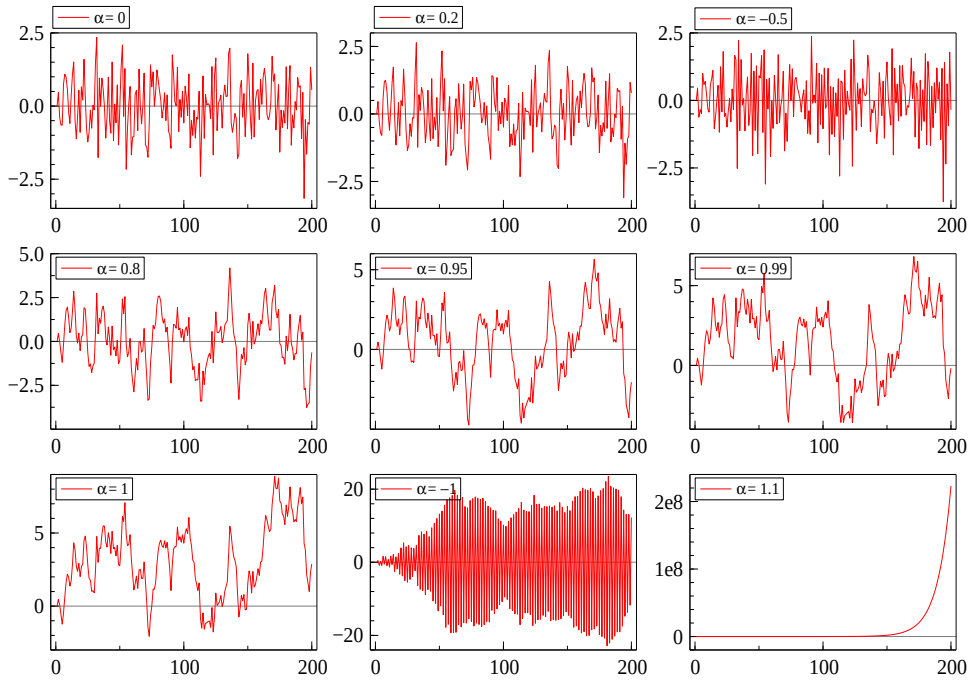

Figure 4.2: Simulated Time Series from an AR(1) $Y_t = \alpha Y_{t-1} + \varepsilon_t$ for various values of α. Notice how the time series properties seem continous when α tends to unity; in fact, the statistical properties of Y_t differ radically depending on whether $|\alpha| < 1$ (stationarity) or $|\alpha| = 1$ (nonstationarity). Notice also the explosiveness of the series for $\alpha > 1$.

We can solve for $\{Y_t\}$ in terms of $\{\varepsilon_t\}$ by inverting $\phi(L)$. The condition for this to work is that the same as the stability condition for the solution of the pth-order difference equation (3.4), namely that $\phi(z) = 0$ has all its roots $|z_i| > 1$. If that holds, then we can obtain a covariance stationary MA(∞) representation $\psi(L)$, with coefficients ψ_j that solve $\psi(L)\phi(L) = 1$. Existence of an MA(∞) representation implies that $\{Y_t\}$ is covariance stationary (and ergodic). Hence, we can derive its mean variance and autocovariances directly from equation (4.9). The mean is found by taking expectations and letting $\mu = \mathsf{E}[Y_t] = \mathsf{E}[Y_{t-1}]$

$$\phi(L)\,\mathsf{E}[Y_t] = \phi(1)\,\mu = c + \mathsf{E}[\varepsilon_t] \Rightarrow \mu = \frac{c}{\phi(1)}.$$

To find the second moments, write (4.14) in terms of

$$\widetilde{Y}_t \equiv Y_t - \mu$$

$$\widetilde{Y}_t = \phi_1 \widetilde{Y}_{t-1} + \ldots + \phi_p \widetilde{Y}_{t-p} + \varepsilon_t.$$

Then, multiply by $\widetilde{Y}_{t-h}$ for $h = 0, ..., p$, take expectations and note that $\mathsf{E}\left[\widetilde{Y}_s \widetilde{Y}_k\right] = \gamma_{s-k}$ for all s, k. Then

$$\gamma_h = \begin{cases} \phi_1\gamma_{h-1} + \ldots + \phi_p\gamma_{h-p}, & h > 0 \\ \phi_1\gamma_1 + \ldots + \phi_p\gamma_p + \sigma^2, & h = 0. \end{cases} \tag{4.17}$$

Using the fact that $\gamma_h = \gamma_{-h}$, Eq. (4.17) contains a system of $p + 1$ linear equations that can be solved uniquely for $\gamma_h, h = 0, \ldots, p$. We will do that in the special case of an AR(2) below.

Dividing (4.17) by γ_0 produces the *Yule-Walker equations*

$$\rho_h = \phi_1\rho_{h-1} + \ldots + \phi_p\rho_{h-p}, \quad h = 1, 2, ... \tag{4.18}$$

The ACF $\{\rho_h\}$ is therefore characterized by a difference equation that has the same characteristic polynomial (4.15) as the process itself.

In general, it is easier to solve for the ACF using the companion representation of the process by defining

$$\underset{p\times 1}{\mathbf{Y}_t} = \begin{pmatrix} Y_t - \mu \\ Y_{t-1} - \mu \\ \vdots \\ Y_{t-p+1} - \mu \end{pmatrix}, \; \underset{p\times 1}{\varepsilon_t} = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and the *companion matrix*

$$\underset{p\times p}{\mathbf{F}} = \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_p \\ 1 & 0 & \ldots & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{4.19}$$

so that

$$\mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \varepsilon_t. \tag{4.20}$$

Then, let $\mathbf{\Omega} = \mathsf{E}\left[\mathbf{Y}_t\mathbf{Y}_t'\right]$, $\mathbf{\Sigma} = \mathsf{E}\left[\varepsilon_t\varepsilon_t'\right]$ and note that $\mathsf{E}\left[\varepsilon_t\mathbf{Y}_{t-1}\right] = \mathbf{0}$, so that

$$\mathbf{\Omega} = \mathbf{F}\mathbf{\Omega}\mathbf{F}' + \mathbf{\Sigma}.$$

Hence

$$vec\left(\mathbf{\Omega}\right) = \left(\mathbf{F} \otimes \mathbf{F}\right) vec\left(\mathbf{\Omega}\right) + vec\left(\mathbf{\Sigma}\right)$$

or

$$vec\left(\mathbf{\Omega}\right) = \left[\mathbf{I} - \left(\mathbf{F} \otimes \mathbf{F}\right)\right]^{-1} vec\left(\mathbf{\Sigma}\right). \tag{4.21}$$

But note that

$$\mathbf{\Omega} = \mathsf{E}\left[\begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \ddots & \cdots & \vdots \\ \vdots & \ddots & \ddots & \gamma_1 \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{pmatrix}\right] \tag{4.22}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}$$

so equation (4.21) yields the first $p-1$ autocovariances of Y_t, including its variance. The remaining autocovariances can be deduced recursively from (4.17) for $h \geq p$.

4.4.1 The AR(2) process

As a special case, consider the derivation of the ACF for the AR(2) process using equations (4.17)

$$\begin{aligned} \gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma^2 \\ \gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_1 \\ \gamma_2 &= \phi_1\gamma_1 + \phi_2\gamma_0. \end{aligned}$$

Divide the bottom two equation by γ_0 to obtain the Yule-Walker equations

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2\rho_1 \Rightarrow \rho_1 = \frac{\phi_1}{1-\phi_2} \\ \rho_2 &= \phi_1\rho_1 + \phi_2 \Rightarrow \rho_2 = \frac{\phi_1^2}{1-\phi_2} + \phi_2, \end{aligned}$$

and substitute for γ_1 and γ_2 in the first equation to get

$$\gamma_0 = \phi_1\rho_1\gamma_0 + \phi_2\rho_2\gamma_0 + \sigma^2$$

$$\begin{aligned}\gamma_0 &= \frac{\sigma^2}{(1-\phi_1\rho_1-\phi_2\rho_2)} \\ &= \frac{(1-\phi_2)\,\sigma^2}{(1+\phi_2)\left[(1-\phi_2)^2-\phi_1^2\right]}.\end{aligned}$$

4.5 The partial autocorrelation function

The ACF is one way of measuring temporal dependence. Sometimes, the information present in the correlogram (the plot of ρ_h over $h = 1, 2, ...$) might be sufficient to deduce the appropriate model for a linear processes $\{Y_t\}$. For instance, the correlogram of an MA(q) is non-zero only up to lag q, while the correlogram of an AR(p) is non-zero at all lags.

However, in practice we do not know the true population correlogram, but we have to rely on an estimate of it based on our observed sample $\{y_t\}_{t=1}^T$. As we will see, the sample autocorrelations

$$\widehat{\rho}_h = \frac{\sum_{t=h+1}^T (y_t-\bar{y})(y_{t-h}-\bar{y})}{\sum_{t=1}^T (y_t-\bar{y})^2}$$

are consistent estimates of ρ_h as $T \to \infty$ for ergodic stationary processes. But clearly we cannot estimate ρ_h beyond $h = T-1$, and in fact, as h gets close to T, we will have fewer observations and less accurate estimates. So in practice, we have to live with $\widehat{\rho}_h$ for $h << T$, and bear in mind that $\widehat{\rho}_h$ will differ from 0 even when $\rho_h = 0$ due to sampling error. One can of course derive standard errors for the above estimators to judge whether they are significantly different from zero, but the fact remains that we will only have a few reliable points in the correlogram to base our decision of what model to use for the $\{Y_t\}$.

Another useful measure of serial dependence is the partial autocorrelation function (PACF). This is defined as follows. Consider the projection of $\{Y_t\}$ on $\{Y_{t-j}, j = 1, ..., m\}$, i.e. let

$$\underset{1\times(1+m)}{\mathbf{X}_t^{(m)}} = (1, Y_{t-1}, ...Y_{t-m})$$

and

$$Y_t = \mathbf{X}_t^{(m)}\alpha^{(m)} + e_t^{(m)} \tag{4.23}$$

where $e_t^{(m)}$ is defined by the moment condition $\mathsf{E}\left[\mathbf{X}_t^{(m)} e_t^{(m)}\right] = \mathbf{0}$. Note that $e_t^{(m)}$ *need not* be serially uncorrelated. The projection coefficients $\alpha^{(m)}$ are defined by

$$\alpha^{(m)} = \begin{pmatrix}\alpha_0^{(m)} \\ \vdots \\ \alpha_m^{(m)}\end{pmatrix} = \mathsf{E}\left[\mathbf{X}_t^{(m)\prime}\mathbf{X}_t^{(m)}\right]^{-1}\mathsf{E}\left[\mathbf{X}_t^{(m)} Y_t\right]$$

Note that $\mathsf{E}\left[\mathbf{X}_t^{(m)\prime}\mathbf{X}_t^{(m)}\right]$ contains the autocovariances of Y_t and so does $\mathsf{E}\left[\mathbf{X}_t^{(m)} Y_t\right]$.

The mth partial autocorrelation of Y_t is given by the coefficient $\alpha_m^{(m)}$. It can be interpreted as the correlation between Y_t and Y_{t-m} after controlling for the variation in $Y_{t-1}, ..., Y_{t-m+1}$. This measure has a natural interpretation in prediction, which we will consider later. The PACF is the sequences $\left\{\alpha_m^{(m)}\right\}_{m=1}^{\infty}$.

Now, we can look at the PACF of an AR(1), (4.9).

1. Clearly, $\alpha_1^{(1)} = \phi$, since $\mathsf{E}[\varepsilon_t Y_{t-1}] = 0$, so that (4.9) is a linear projection. All higher order autocorrelations must necessarily be zero, since it is also the case that $\mathsf{E}[Y_{t-j}\varepsilon_t] = 0$, for all $j > 1$. So, the PACF of an AR(1) looks exactly like the ACF of an MA(1). It is non-zero only up to lag 1.

2. Similarly, the PACF of the AR(p) process (4.14) will have $\alpha_m^{(m)} = 0$, for all $m > p$. Moreover, $\alpha_p^{(p)} = \phi_p$, so we see that the highest lag for which the PACF is non-zero is p. This looks like the ACF function for an MA(p).

3. By contrast, the PACF of an invertible MA(1) is non-zero at all lags. This is not surprising, given that the MA(1) admits an infinite AR representation with nonzero coefficients at all lags. By the same argument, there is no lag h such that $\alpha_m^{(m)} = 0$ for all $m > h$ for an MA(q) process.

The PACF can be estimated simply by running a sequence of linear regression of the form (4.23) and recording the last regression coefficient $\widehat{\alpha}_m^{(m)}$. Under some conditions, this will be a consistent estimate of the population PACF. The partial correlogram is the plot of $\alpha_m^{(m)}$ against m, and can be used in combination with the correlogram. to identify the structure of serial dependence in the data.

4.6 Mixed Autoregressive Moving average processes

AR and MA models can be used to capture serial dependence in a series $\{Y_t\}$. They typically imply different serial dependence patterns, as measured by the ACF and PACF. However, any autocorrelation pattern can be matched exactly by considering an infinite AR or MA representation, since, as we already discussed, AR(∞) and MA(∞) representations are observationally equivalent. In practice, we will want to use these models to study the time series properties of a process $\{Y_t\}$ using a finite sample. Consequently, we cannot hope to fit either AR(∞) or MA(∞) models with unrestricted coefficients. Once we restrict attention to AR(p) or MA(q) models for small p and q, we can no longer match all possible autocorrelation patterns in the data. Thus, it might be a good idea to combine both MA and AR terms into a single model to derive a richer class of models. Such models are known as Autoregressive Moving Average and denoted ARMA(p,q), where p is the number of autoregressive terms and q the number of MA terms.

4.6.1 The ARMA(1,1) model

Consider the process

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \quad \varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right). \tag{4.24}$$

This can be written alternatively as

$$\phi(L) Y_t = c + \theta(L) \varepsilon_t \tag{4.25}$$

where $\phi(L)$ is called the AR polynomial and $\theta(L)$ is the MA polynomial. Whether this process is covariance-stationary or not depends only on the root of $\phi(L)$. In particular, $\{Y_t\}$ is stationary if $|\phi| < 1$, in which case it admits an MA(∞) representation

$$\begin{aligned} Y_t &= \frac{c}{1-\phi L} + \frac{\varepsilon_t}{1-\phi L} + \frac{\theta \varepsilon_{t-1}}{1-\phi L} \\ &= \underbrace{\frac{c}{1-\phi}}_{=\mu} + \sum_{j=0}^{\infty} (\phi L)^j \varepsilon_t + \theta \sum_{j=0}^{\infty} (\phi L)^j \varepsilon_{t-1} \\ &= \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} + \theta \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j-1} \\ &= \mu + \varepsilon_t + \sum_{j=1}^{\infty} \underbrace{\left(\phi^j + \theta \phi^{j-1}\right)}_{\psi_j} \varepsilon_{t-j} \end{aligned} \tag{4.26}$$

Hence, the mean of the process is μ

The variance and ACF can be derived from the infinite MA representation (4.26). These are

$$\begin{aligned} \gamma_0 &= \frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2} \sigma^2 \\ \rho_1 &= \frac{(1 + \phi\theta)(\phi + \theta)}{1 + \theta^2 + 2\phi\theta} \\ \rho_h &= \phi \rho_{h-1}, \quad h = 2, 3, ... \end{aligned}$$

So, we see that the serial correlation pattern of the ARMA(1,1) is a combination of the patterns of MA(1) and AR(1). In particular, the addition of an MA(1) component affects only the variance and first autocorrelation, while higher order autocorrelations are governed by the AR part.

4.6.2 Common factors

If the AR and MA polynomials have a common root, they are said to have a common factor. For the ARMA(1,1) in Eq. (4.24), this situation occurs when $\phi = -\theta$. In that case, the model reduces to a constant plus white noise, since division of both sides by $\phi(L)$ would cancel the AR and MA polynomials. In other words, the coefficients ψ_j in the MA representation would be $\psi_j = 0$, for all $j > 0$. When this occurs, the model is *over-parameterized*, i.e., the parameters θ and ϕ are redundant (and totally unidentified).

4.6.3 The ARMA(p,q) model

If the polynomials $\phi(L)$ and $\theta(L)$ in (4.25) are of orders p and q, the model is called ARMA(p, q). This process is stationary if $\phi(L)$ has all its roots outside the unit circle. Common factors would arise if any of the roots of $\phi(z) = 0$ and $\theta(z) = 0$ are common, in which case the model becomes effectively ARMA$(p-1, q-1)$.

Chapter 5

Forecasting and Estimation of ARMA models

5.1 Forecasts

A common objective of time series analysis is to generate forecasts or predictions of future values of a sequence based on current information. For instance, we may be interested in predicting Y_{t+h}, $h \geq 1$, given the m most recent observations of the series, $X_t = Y_t, Y_{t-1}, .., Y_{t-m+1}$. In general, let X_t denote a set of variables that are used predict Y_{t+h}. A *forecast* or *predictor* of Y_{t+h} is some function of X_t, denoted $Y^*_{t+h|t}$, with associated *forecast error* $Y_t - Y^*_{t+h|t}$. $Y^*_{t+h|t}$ is also called the h-step ahead forecast of Y_{t+h} conditional on X_t.

We can compare different forecasts using some *loss function* that measures the costs of forecast errors. A common criterion is *mean square error* (MSE), namely

$$\mathsf{MSE}\left(Y^*_{t+h|t}\right) = \mathsf{E}\left[\left(Y_{t+h} - Y^*_{t+h|t}\right)^2\right] \tag{5.1}$$

The function of X_t that minimizes the MSE is the conditional expectation of Y_t given X_t, i.e., any other $g(X_t)$ would result in an MSE that is at least as high as that associated with $Y^*_{t+h|t} = \mathsf{E}[Y_{t+h}|X_t]$, also denoted simply $Y_{t+h|t}$.

This can be seen by noting that

$$Y_{t+h} - Y^*_{t+h|t} = (Y_{t+h} - \mathsf{E}[Y_{t+h}|X_t]) + \left(\mathsf{E}[Y_{t+h}|X_t] - Y^*_{t+h|t}\right).$$

Squaring both sides and taking expectations conditional on X_t,noting that the second term is known given X_t, yields

$$\mathsf{E}\left[\left(Y_{t+h} - Y^*_{t+h|t}\right)^2 |X_t\right] = \mathsf{V}(Y_{t+h}|X_t) + \left(\mathsf{E}[Y_{t+h}|X_t] - Y^*_{t+h|t}\right)^2.$$

Now, the first term on the RHS above does not depend on $Y^*_{t+h|t}$, while the second term is non-negative, and can be made exactly 0 by choosing $Y^*_{t+h|t} = \mathsf{E}[Y_{t+h}|X_t]$.

In general, $\mathsf{E}[Y_{t+h}|X_t]$ need not be linear. When we are interested in the class of linear predictors, i.e., all linear functions of X_t

$$g(X_t) = a + bX_t$$

then *the best linear unbiased predictor* (BLUP) by the MSE criterion is the linear projection $\mathsf{L}[Y_{t+h}|X_t]$, also denoted by $\widehat{Y}_{t+h|t}$ to distinguish it from the conditional *expectation* $Y_{t+h|t}$.

This chapter only presents how to compute forecasts in ARMA models. The acomparisons between various candidate forecasts, and also tests for forecast accuracy assessment will be presented in the chapter on risk.

5.2 Optimal forecasts for ARMA models

5.2.1 Autoregressive models

The AR(p) model (4.14) satisfies the orthogonality restriction $\mathsf{E}[Y_{t-j}\varepsilon_t] = 0$, for all $j = 1, 2, ...$ In other words, this model corresponds to the linear projection $\mathsf{L}[Y_t|Y_{t-1}, Y_{t-1}, \ldots]$ (notice that it is the linear projection on the *entire history* of Y_t, even though only p terms may appear with a non-zero coefficient). Hence, the optimal one-step-ahead linear forecast, or BLUP of Y_{t+1} is given by

$$\widehat{Y}_{t+1|t} = \phi_1 Y_t + \ldots + \phi_p Y_{t-p+1} \tag{5.2}$$

and the forecast error is ε_{t+1}.

The h-step ahead forecast can be derived recursively. For instance, for an AR(1), it is

$$\widehat{Y}_{t+h|t} = \phi\widehat{Y}_{t+h-1|t} = \ldots = \phi^h Y_t$$

with associated forecast error

$$Y_{t+h} - \phi^h Y_t = \varepsilon_{t+h} + \phi\varepsilon_{t+h-1} + \ldots + \phi^{h-1}\varepsilon_{t+1}.$$

For an AR(p), one can make use of the companion AR(1) representation to derive

$$\widehat{\xi}_{t+h|t} = F^h\xi_t = \left(F^h\right)_{11} Y_t + \ldots + \left(F^h\right)_{1p} Y_{t-p+1}$$

where the notation $(A)_{ij}$ denotes the (i,j)th element of a matrix A.

If we strengthen our assumption about the errors to $\varepsilon_t \sim \mathsf{GWN}\left(0, \sigma^2\right)$, then it follows that $\mathsf{E}[\varepsilon_t|Y_{t-j}] = 0$, for all j, since ε_t is serially independent, and Y_t is a (infinite) moving average of past ε_ts. Therefore, equation (4.14) is the regression $\mathsf{E}[Y_t|Y_{t-1}, Y_{t-1}, \ldots]$, which is why the model is called an auto*regression.* In that case, the forecast (5.2) becomes optimal in the class of both linear and non-linear predictors.

5.2.2 Moving average models

Consider the MA(∞) process

$$Y_t = \mu + \psi(L)\varepsilon_t, \quad \varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right) \tag{5.3}$$

where

$$\psi(L) = \sum_{j=0}^{\infty} \psi_j L^j, \psi_0 = 1, \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Suppose we observe the sequence $\{\varepsilon_s\}_{s=-\infty}^{t}$ and we are interested in forecasting Y_{t+h}. Expression (5.3) can be written as:

$$Y_{t+h} = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t+h-j} = \mu + \sum_{j=1}^{h} \psi_{h-j}\varepsilon_{t+j} + \sum_{j=0}^{\infty} \psi_{h+j}\varepsilon_{t-j}. \tag{5.4}$$

Next, the fact that ε_t is serially uncorrelated means that $\mathsf{L}[\varepsilon_{t+j}|\varepsilon_t, \varepsilon_{t-1}, ...] = 0$ for all $j > 0$. Hence, taking linear projections on both sides of (5.4) onto $\{\varepsilon_s\}_{s=-\infty}^{t}$, the second term becomes 0, and the BLUP, or optimal MSE linear forecast is given by

$$\widehat{Y}_{t+h|t} = \mathsf{L}[Y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \mu + \sum_{j=0}^{\infty} \psi_{h+j}\varepsilon_{t-j}. \tag{5.5}$$

The forecast error associated with this forecast can be deduced by subtracting (5.5) from (5.4):

$$e_{t+h} = Y_{t+h} - \widehat{Y}_{T+H|T} = \sum_{j=1}^{h} \psi_{h-j}\varepsilon_{t+j}.$$

Clearly, the forecast is unbiased ($\mathsf{E}[e_{t+h}] = 0$) and the associated MSE (or forecast error variance) is given by

$$\mathsf{MSE} = \left(1 + \psi_1^2 + \ldots + \psi_{h-1}^2\right)\sigma^2. \tag{5.6}$$

We observe that the uncertainty associated with the forecast, as measured by its variance, is non-decreasing with the forecast horizon h.

Since $\psi_j \to 0$ as $j \to \infty$, we see that the influence of current and past information on the forecast (5.5) decays as h gets large, and eventually, $\lim_{h\to\infty} \mathsf{L}[Y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \mu$.

A special case is the MA(q) process, where $\psi_j = 0$ for all $j > q$. Thus, the optimal linear forecast is given by

$$\mathsf{L}[Y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \begin{cases} \mu + \psi_h\varepsilon_t + \ldots + \psi_q\varepsilon_{t-q+h} & \text{for } h = 1, 2, ..., q \\ \mu & \text{for } h > q. \end{cases}$$

Thus, we see that beyond q periods ahead, there is no relevant past information in predicting Y_{t+h}, so the optimal conditional forecast coincides with the unconditional mean of the process.

So far, the discussion was based on the assumption that the process $\{\varepsilon_t\}$ is observable. In practice, we only observe the history of $\{Y_t\}$. So, how can we make (5.5) operational given $\{Y_s\}_{s=-\infty}^{t}$ rather than $\{\varepsilon_s\}_{s=-\infty}^{t}$?

Recall that we can solve an MA(q) model of any order for the process $\{\varepsilon_t\}$ in terms of $\{Y_t\}$, provided the MA polynomial is invertible. This is the equivalent autoregressive representation which expresses ε_t as a function of current and lagged Y_t. Let $\phi(L) = \psi^{-1}(L)$ so that (5.3) can be written as

$$\varepsilon_t = \phi(L)(Y_t - \mu)$$

and use this expression of each ε_{t-j} in (5.5). For instance, to forecast an MA(1), $Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$

$$\varepsilon_t = \frac{Y_t - \mu}{1 + \theta L} = \sum_{j=0}^{\infty} (-\theta L)^j (Y_t - \mu)$$

so that

$$\widehat{Y}_{t+1|t} = \mu + \theta\varepsilon_t = \mu + \theta \sum_{j=0}^{\infty} (-\theta L)^j (Y_t - \mu).$$

5.2.3 ARMA models

Forecasts from a covariance-stationary ARMA(p,q) model can be derived recursively as before:

$$\widehat{Y}_{t+h|t} = \mu + \phi_1 \left(\widehat{Y}_{t+h-1|t} - \mu\right) + \ldots + \phi_p \left(\widehat{Y}_{t+h-p|t} - \mu\right) + \widehat{\varepsilon}_{t+h|t} + \ldots + \widehat{\varepsilon}_{t+h-q|t}, \qquad h = 1, 2, \ldots$$

where

$$\widehat{Y}_{t+h-p|t} = Y_{t+h-p}, \quad \text{for } h \leq p \text{ and}$$

$$\widehat{\varepsilon}_{t+h-q} = \begin{cases} 0 & h > q \\ \varepsilon_{t+h-q} & h \leq q. \end{cases}$$

The MSE of the associated forecast error can be deduced from the MA(∞) representation

$$Y_t = \mu + \underbrace{\frac{\theta(L)}{\phi(L)}}_{\psi(L)} \varepsilon_t$$

and is given by (5.6).

5.3 Modeling covariance stationary processes

5.3.1 The Wold decomposition theorem

Of fundamental importance in time-series analysis is the following result, that is known as the **Wold decomposition theorem:**

Theorem 5.1 *Any covariance-stationary process Y_t can be represented in the form*

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t, \tag{5.7}$$

where $\psi_0 = 1$ and $\{\psi_j\}$ is absolutely summable. The term ε_t is white noise, and equals the forecast error associated with the optimal linear one-step-ahead forecast of Y_t namely,

$$\varepsilon_t = Y_t - \mathsf{L}\left[Y_t | Y_{t-1}, Y_{t-2}, \ldots\right].$$

The term κ_t can be predicted arbitrarily well by a linear function of past *values of Y, and is uncorrelated with ε_{t-j} for any j.*

The process Y_t is decomposed into two parts: the infinite MA is referred to as the *indeterministic* part of the process, while κ_t is termed the *deterministic* part.

Finding the Wold representation requires fitting an MA(∞) to Y_t, i.e., finding all the coefficients ψ_j in the sequence. As we already discussed, this is infeasible with a finite number of observations on Y_t, unless we impose some restrictions on the form of ψ_j, i.e., make them functions of a finite set of unknown, estimable parameters. One such possibility is to use an ARMA(p,q) model, such that

$$\psi(L) = \frac{1 + \theta_1 L + \ldots + \theta_q L^q}{1 - \phi_1 L - \ldots - \phi_p L^p}. \tag{5.8}$$

5.3.2 The Box-Jenkins methodology

How should one choose the appropriate order of an ARMA(p,q) for a given time series $\{Y_t\}$, based on a finite sample of T observations? Obviously, provided $\{Y_t\}$ is covariance-stationary, one would like to choose the orders p and q in such a way that the resulting polynomial (5.8) provides a good approximation to the coefficients ψ_j in the true Wold decomposition (5.7). If one chooses p and/or q too small, then the errors in the resulting model will generally be serially correlated, and the associated forecasts will not be optimal (e.g., suppose that the process is AR(2) and you choose to model it by an AR(1) model.) On the other hand, if one chooses p and q too generously, the model will be overparameterized, meaning that the coefficients will not be estimated as accurately as possible. In addition, overparameterization in ARMA models can induce common factors which would induce identification problems (akin to multicollinearity). Therefore, it is desirable to use a *parsimonious* representation.

In a book published in 1976 (entitled *Time Series Analysis: Forecasting and Control*), Box and Jenkins made a strong case in favor of the principle of parsimony, that is, using as few parameters as possible in the ARMA specification, and proposed the following approach:

1. Transform the data if necessary, e.g., take differences, so that the assumption of covariance stationarity is a reasonable one.

2. Plot the (possibly transformed) series together with its correlogram and partial correlogram, ACF and PACF. Use this to make an initial guess about p and q.

3. Estimate the parameters in $\phi(L)$ and $\theta(L)$ of an ARMA(p, q).

4. Perform diagnostic tests to confirm the model is consistent with the observed data, e.g., test that the fitted errors $\hat{\varepsilon}_t$ are white noise.

If diagnostic tests reveal any inconsistencies, one would repeat steps 3 and 4 with a different value of p and/or q. Alternatively, if one thinks the original choice of p and or q is too large, resulting in inefficiency and possible lack of identification, one should repeat steps 3 and 4 using a more restrictive specification.

We have already discussed how the ACF and PACF are related to the underlying parameters of the model, and hence can be used to distinguish between different patterns of serial dependence. For example, a slowly declining ACF and a PACF that is (close to) zero beyond lag 1 matches the serial dependence pattern of an AR(1). The exactly opposite picture, ACF zero beyond lag 1 and slowly declining PACF indicates an MA(1).

We will now turn to the third step of the modeling process, namely, estimation of the parameters of an ARMA model. The discussion of diagnostic tests will only be briefly touched upon.

Example 10 *Let* $u_t \sim \mathsf{GWN}\left(0, \sigma^2\right)$ *. The following* $\{y_t\}$ *stochastic processes are driven a common innovation process* $\{u_t\}$*:*

$$
\begin{aligned}
(i) \quad & y_t = 0.64 y_{t-1} + u_t \\
(ii) \quad & y_t = -0.2 y_{t-1} + 0.64 y_{t-2} + u_t \\
(iii) \quad & y_t = u_t + u_{t-1} \\
(iv) \quad & y_t = 0.64 y_{t-2} + u_t + 0.64 u_{t-1} \\
(v) \quad & y_t = y_{t-1} + u_t \\
(vi) \quad & y_t = y_{t-1} + u_t - 0.9 u_{t-1}
\end{aligned}
$$

A realization of each of these is presented in figure 5.1, together with estimated ACF and PACF. Can you tell which ARMA process generated which graph? answer: i-B, ii-D, iii-F, iv-A, v-E, vi-C.

5.4 Estimation of ARMA models

We will consider estimation of stationary and invertible *Gaussian* ARMA models:

$$
\phi(L) Y_t = c + \theta(L) \varepsilon_t, \quad \varepsilon_t \sim \mathsf{NID}\left(0, \sigma^2\right).
$$

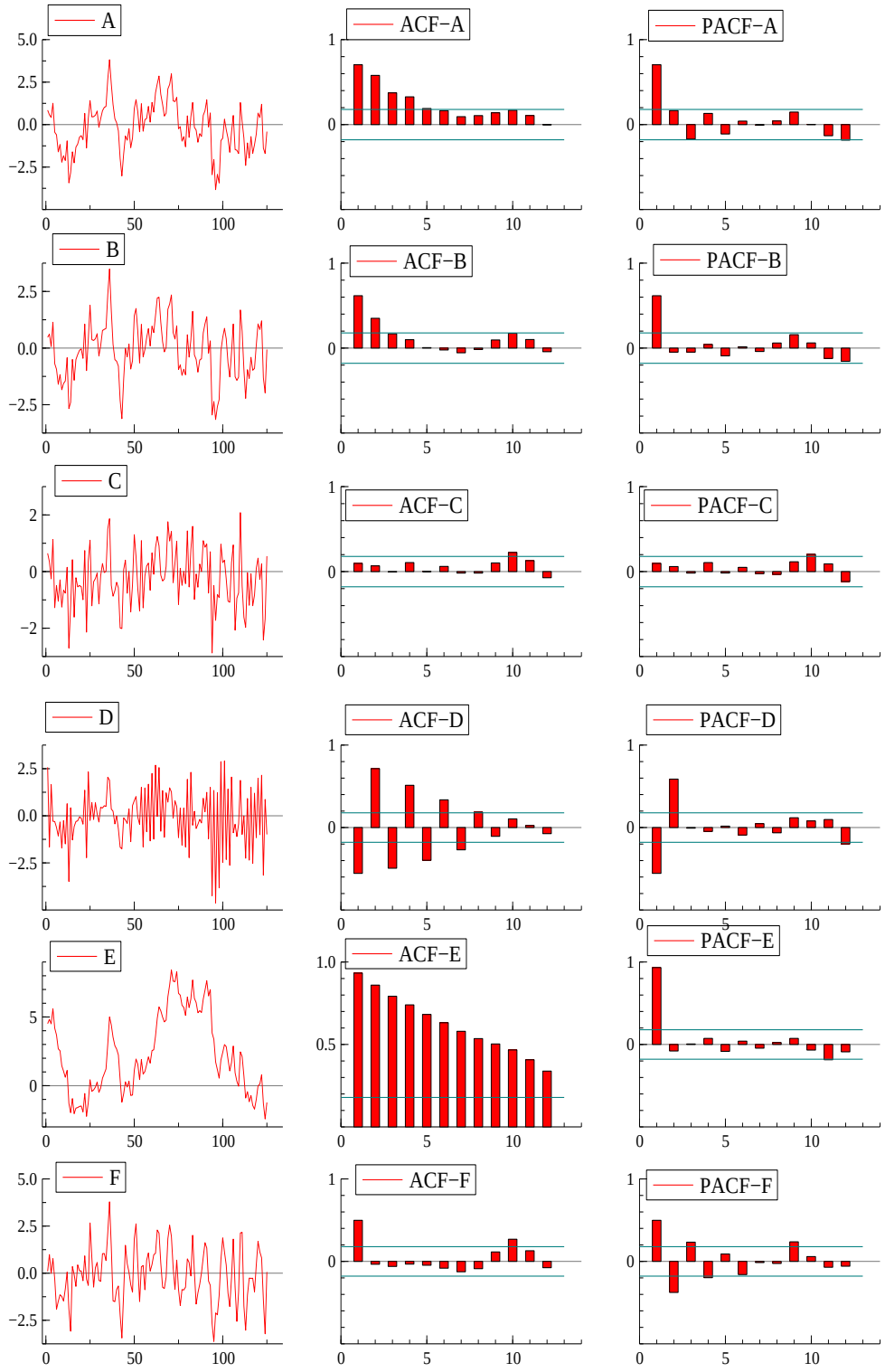

Figure 5.1: Simulated series, ACF and PACF of the ARMA processes in example 10.

The specification of the distribution of $\{\varepsilon_t\}$ effectively gives us the entire joint density of $\{Y_t\}$ (by change of variables), and can be used to derive the maximum likelihood estimator of the unknown parameters.

Collect all the unknown parameters of the model in a vector:

$$\theta = \left(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2\right)'.$$

The likelihood function given any finite sample $\{y_t, t = 1, ..., T\}$ is given by the joint density:

$$\mathcal{L}(\theta) = f_{Y_1,...,Y_T}(y_1, \ldots, y_T; \theta). \tag{5.9}$$

If the data Y_t where i.i.d., then the likelihood function (5.9) would simply be the product of the marginal densities for each observation, namely

$$\mathcal{L}(\theta) = \prod_{t=1}^{T} f_{Y_t}(y_t; \theta).$$

However, for dependent processes, this approach doesn't work. We can proceed in two ways. One way is to determine the joint density of the sample $\mathbf{Y}_T = \{Y_t, t = 1, ..., T\}$ directly from the multivariate normal density of $\{\varepsilon_t\}$. This would involve the $T \times T$ variance matrix for $\mathbf{Y}$ that can be deduced from the ACF of Y_t, and clearly depends on all of θ. An alternative approach uses the so-called *prediction error decomposition* of the joint density (5.9). This involves factorizing that density into a series of conditional densities times the density of a set of initial observations.

For instance, suppose you want the joint density of (Y_t, Y_{t-1}). This can always be factored into a conditional times a marginal density:

$$f_{Y_t,Y_{t-1}}(y_t, y_{t-1}; \theta) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta) \times f_{Y_{t-1}}(y_{t-1}; \theta).$$

For three consecutive points in the sequence, the decomposition becomes

$$\begin{aligned} &f_{Y_t,Y_{t-1},Y_{t-2}}(y_t, y_{t-1}, y_{t-2}; \theta) \\ &= f_{Y_t|Y_{t-1},Y_{t-2}}(y_t|y_{t-1}, y_{t-2}; \theta) \times f_{Y_{t-1}|Y_{t-2}}(y_{t-1}|y_{t-2}; \theta) \times f_{Y_{t-2}}(y_{t-2}; \theta). \end{aligned}$$

So, by recursive conditioning, the joint density for the entire sample can be written as

$$f_{Y_1,...,Y_T}(y_1, \ldots, y_T; \theta) = \prod_{t=p+1}^{T} f_{Y_t|\mathbf{Y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \theta) f_{\mathbf{Y}_p}(\mathbf{y}_p; \theta) \tag{5.10}$$

where $\mathbf{Y}_s = (Y_s, Y_{s-1}, \ldots, Y_1)$ denotes the possible realizations of the process from $t = 1$ to s and $\mathbf{y}_s = (y_s, y_{s-1}, \ldots, y_1)$ denotes actual realizations. Now, observe that the density

$$f_{Y_t|\mathbf{Y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \theta)$$

is effectively the density of the error in forecasting Y_t from $\mathbf{Y}_{t-1}$ using the optimal minimum MSE forecast (that is why it is called the *prediction error* decomposition). The decomposition (5.10) expresses the joint density as a product of the densities of one-step ahead prediction errors and the unconditional distribution of the first p observations in the sample. The log-likelihood can be derived by taking logs of (5.10), evaluating at the observed $\{y_t\}$, and viewing it as a function of θ

$$\ln \mathcal{L}(\theta) = l(\theta) = \sum_{t=p+1}^{T} \ln f_{Y_t|\mathbf{Y}_{t-1}}(y_t|\mathbf{y}_{t-1};\theta) + \ln f_{\mathbf{Y}_p}(\mathbf{y}_p;\theta). \tag{5.11}$$

5.4.1 Log-likelihood of an AR(1) model

Consider a stationary Gaussian AR(1) process (4.9). To derive its log-likelihood function (5.11), let us derive the density of the initial observation ($p = 1$). The process (4.9) admits the MA(∞) representation

$$Y_t = \frac{c}{1-\phi} + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}.$$

Since $\{\varepsilon_t\}$ is jointly normal and Y_t is a weighted sum of normals, it is also normally distributed with mean $c/(1-\phi)$ and variance $\sigma^2\left(1-\phi^2\right)^{-1}$, i.e.,

$$f_{Y_1}\left(y_1; c, \phi, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2/(1-\phi^2)}} \exp\left\{\frac{[y_t - c/(1-\phi)]^2}{2\sigma^2/(1-\phi^2)}\right\}. \tag{5.12}$$

Now, turn to $f_{Y_t|\mathbf{Y}_{t-1}}(y_t|\mathbf{y}_{t-1};\theta)$. Note that the Gaussian AR(1) process is Markovian, i.e.

$$f_{Y_t|\mathbf{Y}_{t-1}}(y_t|\mathbf{y}_{t-1};\theta) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\theta).$$

In fact, conditional on Y_{t-1}, that is, treating Y_{t-1} as non-random, all the randomness in Y_t is due to ε_t. So, the conditional density of Y_t is the same as that of ε_t (the prediction error)

$$f_{Y_t|Y_{t-1}}\left(y_t|y_{t-1}; c, \phi, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\varepsilon_t^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right]. \tag{5.13}$$

Putting (5.13) and (5.12) into (5.11), the log-likelihood function becomes

$$\begin{aligned}\ln \mathcal{L}\left(c, \phi, \sigma^2\right) = &-\frac{T-1}{2}\ln(2\pi) - \frac{T-1}{2}\ln\sigma^2 - \sum_{t=2}^{T}\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2} \\ &-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\frac{\sigma^2}{(1-\phi^2)} - \frac{[y_1 - c/(1-\phi)]^2}{2\sigma^2(1-\phi^2)}.\end{aligned} \tag{5.14}$$

5.4.2 Exact and conditional MLEs

Given the likelihood function, we can proceed to derive the maximum likelihood estimator (MLE) as the value of θ that maximizes the log-likelihood (5.11). This is referred to as the *exact MLE*,

$\widehat{\theta}_{emle}$ As can be seen from equation (5.14), the likelihood function is typically a highly non-linear function of the parameters, so the exact MLE is typically not analytically available, and has to be computed numerically.

There is an alternative to the exact MLE, called the *conditional MLE*, $\widehat{\theta}_{cmle}$, which is derived by maximizing only the first part of the log-likelihood function (5.11). It is called *conditional* because it treats the initial observations $\mathbf{y}_p$ as fixed. For stationary processes, both $\widehat{\theta}_{emle}$ and $\widehat{\theta}_{cmle}$ are consistent and, in fact, asymptotically equivalent, because the effect of the initial conditions becomes unimportant as the sample grows (the first term in (5.11) is of order T while the second term is of order 1). However, they can differ substantially in small samples if the process is close to being non-stationary or non-invertible (i.e., if some root of $\phi(L)$ or $\theta(L)$ is close to 1).

5.4.3 Conditional log-likelihood of an MA(1) model

Consider an invertible MA(1) process (4.1) and suppose that we know for certain that $\varepsilon_0 = 0$. This assumption implies that we can effectively observe $\{\varepsilon_t\}$ given a sample on Y and the values of μ, θ. ε_t will solve the difference equation

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}, \quad t = 1, 2, ..., T$$

or

$$\varepsilon_t = \sum_{j=0}^{t-1} (-\theta)^j (y_{t-j} - \mu).$$

Now, ε_t is the one step ahead forecast error of Y_t, so, by the prediction error decomposition, the conditional log-likelihood will be given by

$$f_{Y_t|\mathbf{Y}_{t-1}}(y_t|\mathbf{y}_{t-1};\theta) = f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1};\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\varepsilon_t^2}{2\sigma^2}\right].$$

So, the log-likelihood function is

$$\ln \mathcal{L}\left(\mu, \theta, \sigma^2\right) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \sum_{t=1}^{T} \frac{(y_t - \mu - \theta\varepsilon_{t-1})^2}{2\sigma^2}.$$

Unlike the AR(1) model, the likelihood function uses all T observations. This is due to the fact that we are assuming that ε_0 is known (the assumption that it is zero is inessential, the important thing is that it is treated as fixed in order to be able to compute ε_1 from y_1 and ε_0 and so on for ε_s,.$s \geq 1$).

The conditional log-likelihood of an MA(q) process takes the form

$$\ln \mathcal{L}\left(\mu, \theta, \sigma^2\right) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2} \tag{5.15}$$

$$\varepsilon_t = y_t - \mu - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q}$$

where we now condition on $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1}$.

5.4.4 Conditional log-likelihood for an ARMA(p,q)

Consider now the ARMA(p,q) model. The conditional log-likelihood treats the first p values of Y_t and q values of ε_t as fixed. This is effectively a combination of the AR(p) and MA(q) log-likelihoods

$$\ln \mathcal{L}(\theta) = -\frac{T-p}{2}\ln(2\pi) - \frac{T-p}{2}\ln\sigma^2 - \sum_{t=p+1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}$$

5.4.5 Computing the MLE

To derive the MLE we can compute the first derivative of the log-likelihood function, also called the *score function* which is a vector-valued function of dimension k, the number of elements of the parameter vector θ, and typical element:

$$s_i(\theta) = \frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta_i}, \quad i = 1, 2, ..., k,$$

If the score function has many zeros (multiple solutions to the first-order conditions), then the likelihood function has several points of inflexion or local optima. The MLE would be the solution of $s(\theta) = 0$ corresponding to the global maximum, at which the $k \times k$ matrix of second derivatives of the log-likelihood function (Hessian), with typical element

$$H_{ij}(\theta) = \frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta_i \partial \theta_j}, \quad i, j = 1, 2, ..., k$$

is negative definite.

Typically, the MLE is not analytically available, and has to be found by some numerical optimization procedure, such as Newton-Raphson, see Hamilton section 5.7 for details.

An exception occurs in the case of the conditional MLE for autoregressive models. The conditional log-likelihood function is the same as the log-likelihood of a Gaussian linear regression model of Y_t on $X_t = (1, Y_{t-1}, \ldots, Y_{t-p})$ over the sample $t = p+1, ..., T$. In fact, by deriving the first-order conditions of the maximization problem (the score), we can deduce that the MLE of the autoregressive coefficients ϕ_i is simply the OLS estimator of the coefficients of the linear projection of Y_t on X_t, namely

$$\begin{pmatrix} \widehat{c} & \hat{\phi}_1 & \ldots & \hat{\phi}_p \end{pmatrix} = \sum_{t=p+1}^{T} Y_t X_t \left(\sum_{t=p+1}^{T} X_t' X_t \right)^{-1}.$$

5.4.6 Asymptotic Distribution theory for serially dependent processes

Covariance-stationary processes obey both a Law of Large Numbers (LLN) and a Central Limit Theorem (CLT).

Proposition 2 *Let Y_t be a covariance stationary process with moments defined as previously and absolutely summable autocovariances. Then the sample mean satisfies:*

1. *LLN:* $\overline{Y}_T \overset{m.s.}{\to} \mu$
2. $\lim_{T\to\infty} \left\{ T \, \mathsf{E}\left[\left(\overline{Y}_T - \mu \right)^2 \right] \right\} = \sum_{j=-\infty}^{+\infty} \gamma_j$

Any ARMA process has absolutely summable autocovariances, and hence follows a LLN. Besides, it follows a CLT.

Proposition 3 *Let $Y_t - \mu \sim MA(\infty)$, then*

$$\sqrt{T}\left(\overline{Y}_T - \mu \right) \overset{L}{\to} \mathsf{N}\left(0, \sum_{j=-\infty}^{+\infty} \gamma_j \right).$$

These asymptotic results imply that OLS estimation of a covariance stationary AR(p) process is asymptotically consistent. Indeed although the estimator OLS of $\widehat{\phi}$ in

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right) \tag{5.16}$$

$$= (1, Y_{t-1}, ..., Y_{t-p})\, \phi + + \varepsilon_t \tag{5.17}$$

(4.20) is biased, yet not asymptotically so, since

$$\sqrt{T}\left(\widehat{\phi} - \phi \right) \overset{L}{\to} \mathsf{N}\left(0, \sigma^2 \mathbf{Q}^{-1}\right)$$

with

$$\mathbf{Q} = \begin{bmatrix} 1 & \mu & \mu & \cdots & \mu \\ \mu & \gamma_0 + \mu^2 & \gamma_1 + \mu^2 & \cdots & \gamma_{p-1} + \mu^2 \\ \mu & \gamma_1 + \mu^2 & \gamma_0 + \mu^2 & \cdots & \gamma_{p-2} + \mu^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mu & \gamma_{p-1} + \mu^2 & \gamma_{p-2} + \mu^2 & \cdots & \gamma_0 + \mu^2 \end{bmatrix}.$$

Also, the t-statistics are asymptotically normal.

5.5 Diagnostics checking

5.5.1 Inference

Under certain dependence conditions (such as ergodicity), we can establish the consistency and asymptotic normality of the MLE. In particular, when T is sufficiently large, the distribution of $\widehat{\theta}_{mle}$ is approximately normal with variance equal to $1/T$ times the inverse of the *information matrix* $\mathcal{I}$. The information matrix is the variance of the score vector $s(\theta)$. Given a consistent estimate of $\mathcal{I}^{-1}$, hypotheses on θ can be tested using the Wald test, as usual.

In the case of the stationary autoregressive model AR(p), the conditional MLE is OLS, and the OLS standard errors are consistent estimates of the standard errors of the MLE. Hence, t-tests and F-tests can be interpreted in the usual way. However, a cautionary note is in order. When the assumption of stationarity is suspect, for instance because a root of the autoregressive polynomial $\phi(z)$ is close to the unit circle, the normal approximation to the distribution of the MLE can be very poor, and the associated t-tests may reject too often (i.e., have an actual size that is higher than their nominal size). This situation is prevalent in macroeconomics and finance applications.

5.5.2 Hypothesis testing on the residuals

There exist many tests for the assumption that the residuals are (Gaussian) white noise. Each based on one element of the IIN hypothesis. Most in general proceed under the null of IIN residuals, and rejecting the null implies that the model is misspecified. Most of tests are the same as in the standard case of cross-section regressions.

Non-autocorrelated

The LM (Lagrange Multiplier) test for autocorrelation (or Breusch-Godfrey test) consists in regressing the residuals on their lags up until the pth lag and on the regressors entering the equation from which they are derived. Under the null hypothesis the regression is uninformative and the goodness of fit index (R^2) is close to zero. Then

$$\frac{R^2}{1-R^2} \underset{\mathsf{H}_0}{\sim} \chi^2_p \tag{5.18}$$

Homoscedasticity

- **White heteroscedasticity test:** this tests checks that the squared residuals cannot be explained from the regressors entering the equation from which they are defined and the squares thereof. Under the null, (5.18) holds.

- **ARCH Test** (AutoRegressive Conditional Heteroscedasticity): this test checks that the squared residuals are non-autocorrelated. It proceeds in the same way as the LM test but on the $\widehat{\varepsilon}^2_t$ and their lags; again (5.18) holds

$$\frac{R^2}{1-R^2} \underset{\mathsf{H}_0}{\sim} \chi^2_p$$

Normality

The Jarque-Bera statistic consists in comparing the estimated third and fourth centered moments of the residuals to those of a normal distribution, for which the skewness $\kappa_3 = \gamma_1 = 0$ and so is

the excess kurtosis $\gamma_2 = \kappa_4/\sigma^4 - 3$. With, p the number of estimated parameters in the regression that allows to compute $\widehat{\varepsilon}_t^2$

$$JB = \frac{T-p}{6}\left[\gamma_1 + \gamma_2\right] \underset{\mathsf{H}_0}{\sim} \chi_2^2$$

5.5.3 Information Criteria

When several competing models pass the diagnostic checking, an *ad hoc* criterion must be used. There are several of them. In time series, two are mainly used: choose the model that minimizes the AIC (Akaike Information Criterion) or SC (Schwarz Criterion) statistics

$$\begin{aligned}\mathsf{AIC}(p,q) &= T\log\left(\sigma_{\widehat{\varepsilon}}^2\right) + 2\left(p+q\right)\\ \mathsf{SC}\left(p,q\right) &= T\log\left(\sigma_{\widehat{\varepsilon}}^2\right) + \left(p+q\right)\log\left(T\right).\end{aligned}$$

where Y_t is assumed to follow an ARMA(p,q). SC coincides the Bayesian Information Criterion (BIC) and is often recommended for ARMA time series when both criteria contradict.

Chapter 6

Integrated processes

ARMA modeling in the Box-Jenkins approach relies on stationarity. Indeed, when the process under consideration is non-stationary, inference is non-standard, whether in univariate or multivariate settings. In particular:

1. Non-stationarity means that the parameters may have no meaning (e.g. what is the estimator of a time-varying mean?)

2. When regressing a non-stationary process on another:

$$Y_t = \beta X_t + u_t$$

   the estimator $\widehat{\beta}$ does not necessarily tend to zero even if the series are truly independent

3. The slope estimator $\widehat{\rho}$ in the regression $Y_t = \rho Y_{t-1} + \varepsilon_t$ is not Normal when $\rho = 1$ whereas it is when $|\rho| < 1$.

A convenient test of stationarity is the **unit root test**. This is based on testing for the presence of a *stochastic trend*. This is part of an extension of ARMA(p, q) models to ARIMA(p, d, q) whose formulation is:

$$\alpha(L)\,\Delta^d Y_t = \beta(L)\,\epsilon_t,$$

where $\Delta = (1 - L)$, $\alpha(L)$ and $\beta(L)$ are of order p and q and their roots satisfy the stationarity, invertibility and irreducibility conditions. $\Delta^d Y_t$ is stationary ARMA. We will show below that Y_t is then non-stationary; yet, $\Delta^d Y_t$ is stationary ARMA and Y_t is said to be **integrated of order** d, which we write

$$Y_t \sim \mathsf{I}(d).$$

A process that is integrated of order $d \geq 1$ is often simply referred to as integrated. The simplest

model of an integrated process is the *random walk,* or ARIMA$(0,1,0)$:

$$\begin{aligned}\Delta Y_t &= \epsilon_t \\ Y_t &= Y_{t-1} + \epsilon_t\end{aligned}$$

6.1 Lag-polynomials Revisited

Recall that a univariate time series $\{Y_t\}$ is called second order (or covariance or weakly) stationary if, for all t and j

$$\begin{aligned}\mathsf{E}[y_t] &= \mu, \\ \mathsf{E}[(y_t-\mu)(y_{t-j}-\mu)] &= \gamma_j\end{aligned}$$

with μ the mean of the x_t, γ_0 the variance of x_t and γ_j the covariance between x_t and x_{t-j}. We have seen that all finite order MA models generate covariance stationary time series.

By contrast for AR(p) models,

$$\phi(L)\, Y_t = \varepsilon_t$$

the roots of $\phi(L)$ must be strictly greater than unity in modulus to ensure weak stationarity. Consider the AR(1)

$$Y_t = c + \rho Y_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right)$. Recurrent substitution in the AR(1) model yields

$$\begin{aligned}Y_t &= c + \rho Y_{t-1} + \varepsilon_t \\ &= c + \rho(c + \rho Y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \rho^j Y_{t-j} + \varepsilon_t + \rho\varepsilon_{t-1} + \ldots + \rho^{j-1}\varepsilon_{t-j+1} \\ &= \rho^t Y_0 + \varepsilon_t + \rho\varepsilon_{t-1} + \ldots + \rho^{t-1}\varepsilon_1,\end{aligned}$$

and we assume that $y_0 = 0$ (This directly shows that $\rho > 1$ leads to explosive and implausible behavior). Then

$$\begin{aligned}\mathsf{E}\left[y_{t+j}|y_t\right] &= \rho^j y_t \\ \mathsf{Cor}\left[y_{t+j}, y_t\right] &= \rho^j\end{aligned}$$

When $j \to \infty$, we obtain the unconditional mean of the AR(1) model:

$$\mathsf{E}\left[Y\right] = \lim_{j\to\infty} \mathsf{E}\left[Y_{t+j}|Y_t\right] = \lim_{j\to\infty} \rho^j Y_t = 0 \text{ when } |\rho| < 1.$$

Similarly

$$\begin{aligned}\mathsf{V}\left[Y_t\right] &= \mathsf{E}\left[(\varepsilon_t + \rho\varepsilon_{t-1} + \ldots + \rho^{t-1}\varepsilon_1)^2\right] \\ &= \sigma^2 \frac{1-\rho^{2t}}{1-\rho^2}\end{aligned}$$

So, when $t \to \infty$, we obtain the unconditional variance,

$$\mathsf{V}[Y] = \lim_{t\to\infty} \mathsf{V}[Y_t] = \lim_{t\to\infty} \sigma^2 \frac{1-\rho^{2t}}{1-\rho^2} = \frac{\sigma^2}{1-\rho^2} \text{ when } |\rho| < 1.$$

For $|\rho| < 1$, the AR(1) satisfies the stationarity conditions. When $\rho = 1$, several things occur in the AR(1) model:

- Unconditional mean is not defined
- Unconditional variance becomes infinite
- $\mathsf{Cor}[Y_{t+j}, Y_t] = 1$
- $\mathsf{E}[Y_{t+j}|Y_t] = Y_t$ also when $j \to \infty$, so best predictor of future values is today's value
- In the representation of Y_{t+j}

  $$Y_{t+j} = \rho^j Y_t + \varepsilon_{t+j} + \rho\varepsilon_{t+j-1} + \ldots + \rho^{j-1}\varepsilon_{t+1},$$

  the importance of Y_t does not die out when $j \to \infty$. So, Y_t becomes like a deterministic or trend element in the representation of Y_{t+j} as it always remains present. $\rho = 1$ therefore leads to a change in the interpretation of the constant and trends in the AR(1) model, when these are present, compared to $\rho < 1$. In the **random walk** model

  $$Y_t = Y_{t-1} + \varepsilon_t$$

  then Y_t is non stationary since

  $$\begin{aligned}\mathsf{V}[Y_t] &= \mathsf{E}\left[(\varepsilon_t + \varepsilon_{t-1} + \ldots + \varepsilon_1)^2\right] \\ &= t\sigma^2\end{aligned}$$

  is non constant.

6.2 Integrated time series

When $\rho = 1$, the AR(1) model with a constant becomes

$$Y_t = \tau + Y_{t-1} + \varepsilon_t,$$

we refer to it as a **random walk with drift** model. The mean and covariance function of this process are, for all t and j

$$\begin{aligned}\mathsf{E}[Y_t|Y_0] &= \tau t + Y_0 = \mu_t, \\ \mathsf{E}[(y_t - \mu_t)(y_{t-j} - \mu_t)] &= \sigma^2(t-j)\end{aligned}$$

for which one uses that

$$\begin{aligned} Y_t &= \tau + Y_{t-1} + \varepsilon_t \\ &= \tau + (\tau + y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \tau t + Y_0 + \sum_{i=1}^{t} \varepsilon_i, \end{aligned}$$

does not satisfy the conditions for stationarity.

The first difference of the random walk with drift series:

$$\begin{aligned} \Delta Y_t &= Y_t - Y_{t-1} \\ &= \tau + Y_{t-1} + \varepsilon_t - Y_{t-1} = \tau + \varepsilon_t, \end{aligned}$$

does, however, satisfy the stationarity conditions. A random walk with drift series is therefore integrated of order 1: $I(1)$.

Definition 11 *A time series Y_t, $t = 1, \ldots$ with mean μ_t, which is a deterministic function of time, is integrated of order d, denoted by $\mathsf{I}(d)$, when $\Delta^d(Y_t - \mu_t)$ is stationary while $\Delta^{d-1}(Y_t - \mu_t)$ is non-stationary.*

Example 11 *Figure 6.1 presents French quarterly inflation (π_t) and the overnight interest rate (i_t) in logs with their first difference. Whereas inflation and interest rates are not stable around a constant mean, their differences seem more stable. Conditional on a statistical test showing that the differences in the variables are stationary whereas the levels are not, π_t and i_t seem integrated of order 1: $\pi_t \sim \mathsf{I}(1)$ and $\Delta\pi_t \sim \mathsf{I}(0)$.*

ACFs of non-stationary variables (figure 6.2) decrease slowly whereas those of stationary variables oscillate around zero. By contrast the first value of the PACF for an integrated variable is close to unity whereas the subsequent values tend to zero very fast. For stationary variables, the PACF is always different from 1.

Example 12 *Figures 6.3 and 6.4 show examples of integrated series of order 1 and 2. These graphs are representative of series of these types. $\mathsf{I}(1)$ variables are generally erratic and can take any values: these so called* stochastic trends *may seem to follow trends over very short periods but the trend changes over time. It is by contrast easy to spot $\mathsf{I}(2)$ variables: they are generally very smooth and their pattern evolves slowly over time.*

6.3 Interpretation of constants and trends in random walk model

Consider the AR(1) model with a constant term:

$$Y_t = \tau + \rho Y_{t-1} + \varepsilon_t.$$

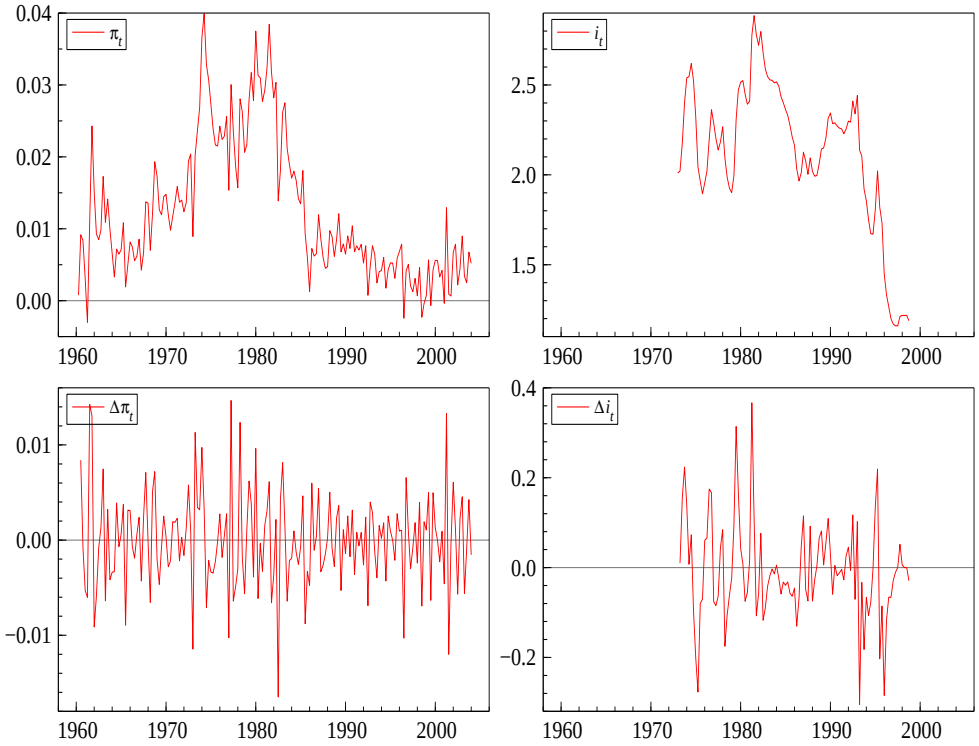

Figure 6.1: French quarterly inflation (π_t) and overnight interest rate (i_t) in logs with their first differences ($\Delta\pi_t = \pi_t - \pi_{t-1}$). Source: DataInsight.

When $\rho < 1$, we can specify the model as

$$Y_t - \mu = \rho(Y_{t-1} - \mu) + \varepsilon_t$$

by using $\tau = \mu(1-\rho)$. This model is like a standard AR(1) model in the variable $Y_t^* = Y_t - \mu$,

$$Y_t^* = \rho Y_{t-1}^* + \varepsilon_t,$$

such that as the unconditional mean of Y_t^* is zero, $\mathsf{E}[Y^*] = 0$, the unconditional mean of Y_t is equal to μ,

$$\mathsf{E}(Y) = \mu = \frac{\tau}{1-\rho},$$

and which shows that the constant represents the mean of the series.

When $\rho = 1$,

$$\begin{aligned} Y_t &= \tau + Y_{t-1} + \varepsilon_t \\ &= 2\tau + Y_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &= \tau t + \varepsilon_1 + \ldots + \varepsilon_t. \end{aligned}$$

This shows that τ represents the growth of the series and can now be interpreted as a linear trend variable.

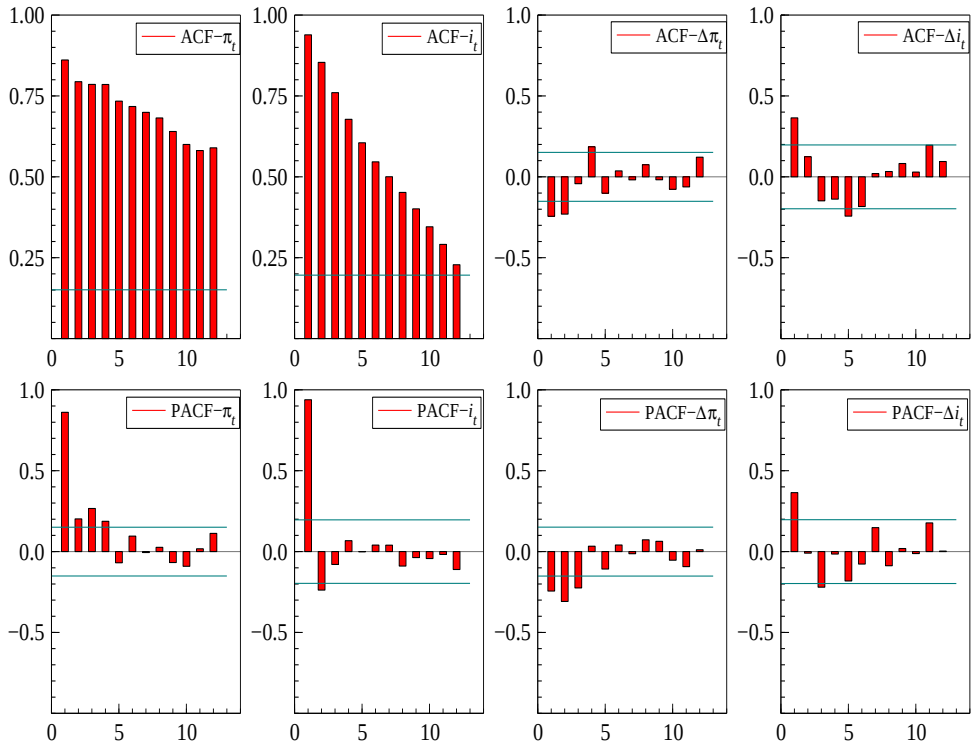

Figure 6.2: Autocorrelation Functions (ACF) and Partial Autocorrelation Functions (PACF) for the log of inflation (π_t) and overnight interest rate (i_t) and their first differences.

The interpretation of the constant term thus depends on the value of ρ.

A similar phenomenon occurs in the AR(1) model with constant and linear trend,

$$Y_t = \tau + \delta t + \rho Y_{t-1} + \varepsilon_t.$$

We can re-specify this model as:

$$(Y_t - \mu - \lambda t) = \rho(Y_{t-1} - \mu - \lambda(t-1)) + \varepsilon_t,$$

where

$$\left.\begin{array}{c} \tau = \mu\,(1-\rho) + \rho\lambda \\ \delta = \lambda(1-\rho) \end{array}\right\} \Leftrightarrow \left\{\begin{array}{c} \mu = \frac{(1-\rho)\tau - \rho\delta}{(1-\rho)^2} \\ \lambda = \frac{\delta}{1-\rho} \end{array}\right. .$$

By using $Y_t^* = Y_t - \mu - \lambda t$, this model is again a standard AR(1) such that as the unconditional mean of Y_t^* is equal to zero, the mean of Y_t is equal to $\mu + \lambda t$ when $|\rho| < 1$.

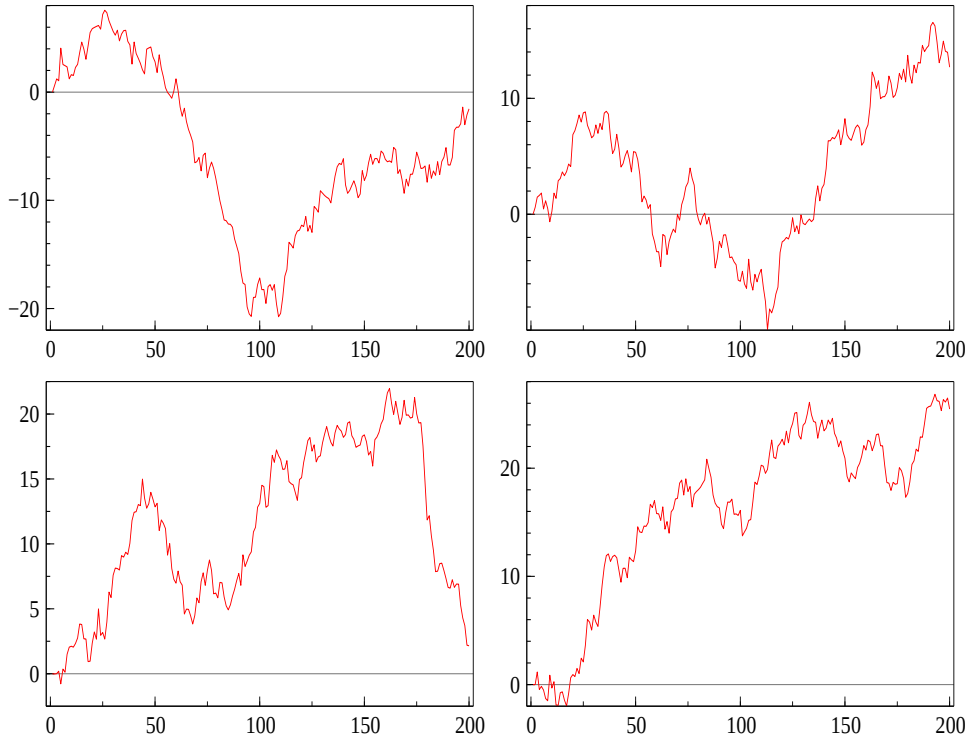

Figure 6.3: Four examples of simulated variables that are integrated of order 1 with zero initial value.

When $\rho = 1$,

$$\begin{aligned} Y_t &= \tau + \delta t + Y_{t-1} + \varepsilon_t \\ &= \tau + \delta t + (\tau + \delta(t-1) + \varepsilon_{t-1}) + \varepsilon_t \\ &= \tau t + \delta \sum_{i=1}^{t} i + \sum_{i=1}^{t} \varepsilon_i \\ &= \tau t + \frac{1}{2}\delta t(t+1) + \sum_{i=1}^{t} \varepsilon_i, \end{aligned}$$

where we used that $\sum_{i=1}^{t} i = \frac{1}{2}t(t+1)$. This shows that δ now represents a quadratic trend while it represents a linear trend when $|\rho| < 1$. So, also in the AR(1) model with a constant and a linear trend, the interpretation of these deterministic components depends on the value of ρ. This is especially important when testing for $\rho = 1$ as the model needs to be able to explain the trending pattern of the series both under $H_0 : \rho = 1$ and under $H_1 : \rho \neq 1$.

We test $H_0 : \rho = 1$ therefore either in:

- AR(1) model with constant when the series does not show deterministic trending behavior

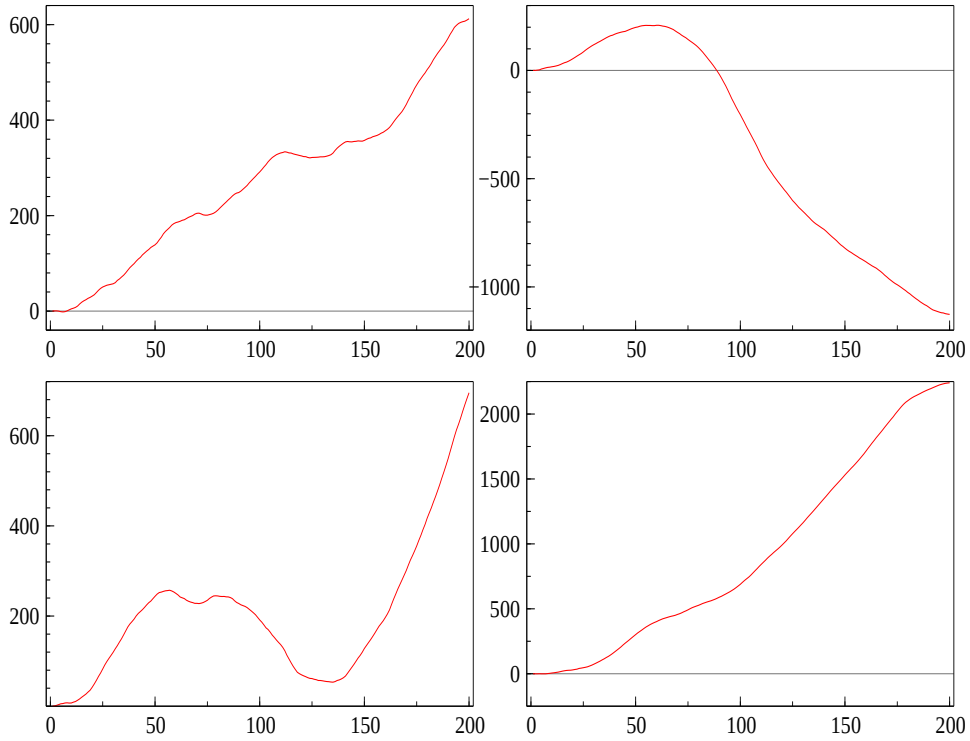

Figure 6.4: Four examples of simulated variables that are integrated of order 2 with zero initial value.

- AR(1) model with constant and linear trend when the series shows deterministic trending behavior.

The decomposition of the AR(p) polynomial

$$\begin{aligned} Y_t &= \tau + \delta t + \rho_1 Y_{t-1} + \ldots + \rho_p Y_{t-p} + \varepsilon_t \Leftrightarrow \\ \Delta y_t &= \tau + \delta t + \alpha Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \ldots + \alpha_{p-1} \Delta Y_{t-p+1} + \varepsilon_t, \end{aligned}$$

with

$$\begin{aligned} \alpha &= \rho_1 + \ldots + \rho_p - 1 = -\rho(1), \\ \alpha_i &= -(\rho_{i+1} + \ldots + \rho_p),\ i = 1, \ldots, p-1, \end{aligned}$$

allows for a direct extension of the previous results for the AR(1) model to the AR(p) since stationarity of the AR(p) polynomial implies that all roots exceed one in absolute value. The parameter α now indicates whether a unit root is present. In case of a unit root, the interpretation of constants and trends changes in an identical manner as in the AR(1), i.e.. the constant term represents the drift and the linear trend term represents quadratic growth.

6.4 Testing for a unit root (Hamilton Chap 17.1-17.4)

6.4.1 Convergence of slope estimators (*)

In the AR(1) model with $|\rho| < 1$,

$$Y_t = \rho Y_{t-1} + \varepsilon_t,$$

the limit behavior of the least squares estimator of the slope ρ,

$$\widehat{\rho} = \left(\sum_{t=1}^{T} Y_{t-1}^2 \right)^{-1} \left(\sum_{t=1}^{T} Y_{t-1} Y_t \right),$$

can be determined by using the central limit theorem. We therefore substitute the AR(1) model in the expression of the least squares estimator:

$$\begin{aligned} \widehat{\rho} = & \left(\sum_{t=1}^{T} Y_{t-1}^2 \right)^{-1} \left(\sum_{t=1}^{T} Y_{t-1} (\rho Y_{t-1} + \varepsilon_t \right) \\ = & \rho + \left(\sum_{t=1}^{T} Y_{t-1}^2 \right)^{-1} \left(\sum_{t=1}^{T} Y_{t-1} \varepsilon_t \right). \end{aligned}$$

Since the mean of Y_t equals zero and $\mathsf{V}[Y_t] = \frac{\sigma^2}{1-\rho^2}$, the law of large numbers implies that

$$\frac{1}{T} \sum_{t=1}^{T} Y_{t-1}^2 \xrightarrow{p} \frac{\sigma^2}{1-\rho^2} = \mathsf{V}\left[Y_t\right]$$

where $\xrightarrow{p}$ stands for convergence in probability.

For a series $X_t, t = 1, \ldots$ of independent random variables with mean zero and finite variance, the central limit theorem applies

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \xrightarrow{L} Z,$$

where $\xrightarrow{L}$ indicates convergence in distribution and Z is a normally distributed random variable with mean zero and variance $Q = \lim_{T \to \infty} \frac{1}{T} \mathsf{E}\left[\left(\sum_{t=1}^{T} X_t \right)^2 \right]$, $Z \sim \mathsf{N}(0, Q)$. The series $Y_{t-1}\varepsilon_t$, $t = 1, \ldots$ also consists of independent random variables with mean zero and finite variance such that the central limit theorem also applies to $\sum_{t=1}^{T} Y_{t-1}\varepsilon_t$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Y_{t-1} \varepsilon_t \xrightarrow{L} Z.$$

The random variable Z has now a normal distribution with mean zero and the variance results from

$$\mathsf{V}\left[Y_{t-1}\varepsilon_t\right] = \mathsf{V}\left[Y_{t-1}\right] \mathsf{V}\left[\varepsilon_t\right] = \frac{\sigma^2}{1-\rho^2} \sigma^2,$$

so

$$Z \sim \mathsf{N}\left(0, \frac{\sigma^2}{1-\rho^2}\sigma^2\right).$$

This implies for the least squares estimator $\widehat{\rho}$ that

$$\sqrt{T}(\widehat{\rho} - \rho) \xrightarrow{L} V,$$

with

$$V \sim \mathsf{N}\left(0, 1-\rho^2\right)$$

since $V = (\frac{\sigma^2}{1-\rho^2})^{-1}Z$ and $(\frac{\sigma^2}{1-\rho^2})^{-2}\frac{\sigma^2}{1-\rho^2}\sigma^2 = 1-\rho^2$. This explains the simple expression of the Lagrange-Multiplier statistic to test H_* : $\rho = 0$ since under H_* $\sqrt{T}\widehat{\rho} \xrightarrow{L} \mathsf{N}(0,1)$ implying that $\sqrt{T}\widehat{\rho}$ can directly be compared with ± 2 to test H_* with 95% asymptotic significance.

The construction of the central limit theorem only applies when $|\rho| < 1$. This is also revealed by the variance of V that equals $1-\rho^2$ which equals zero when ρ is one. Hence,

$$\sqrt{T}(\widehat{\rho} - 1) \xrightarrow{p} 0.$$

To obtain the limiting distribution of the least squares estimator when $\rho = 1$, we thus have to scale by a higher order than $\sqrt{T}$. We proceed by constructing this scaling factor and the limiting distribution.

When $\rho = 1$, $Y_0 = 0$ and $\varepsilon_t \sim \mathsf{N}(0, \sigma^2)$ and independent,

$$Y_t = \varepsilon_t + \ldots + \varepsilon_1,$$

so

$$Y_t \sim \mathsf{N}(0, \sigma^2 t).$$

Also

$$Y_t^2 = (Y_{t-1} + \varepsilon_t)^2 = Y_{t-1}^2 + 2Y_{t-1}\varepsilon_t + \varepsilon_t^2$$

such that

$$Y_{t-1}\varepsilon_t = \frac{1}{2}(Y_t^2 - Y_{t-1}^2 - \varepsilon_t^2).$$

This expression is summed from $t = 1$ to T in the least squares estimator:

$$\begin{aligned}\sum_{t=1}^{T} Y_{t-1}\varepsilon &= \frac{1}{2}\sum_{t=1}^{T}(Y_t^2 - Y_{t-1}^2 - \varepsilon_t^2) \\ &= \frac{1}{2}(Y_T^2 - Y_0^2) - \frac{1}{2}\sum_{t=1}^{T}\varepsilon_t^2 \\ &= \frac{1}{2}Y_T^2 - \frac{1}{2}\sum_{t=1}^{T}\varepsilon_t^2,\end{aligned}$$

since $Y_0 = 0$. We scale this expression by $(\sigma\sqrt{T})^2$:

$$\frac{1}{\sigma^2 T}\sum_{t=1}^{T} Y_{t-1}\varepsilon_t = \frac{1}{2}(\frac{Y_T}{\sigma\sqrt{T}})^2 - \frac{1}{2}\frac{1}{\sigma^2 T}\sum_{t=1}^{T}\varepsilon_t^2,$$

since $\frac{Y_T}{\sigma\sqrt{T}} \sim N(0,1)$ and, because of the law of large numbers,

$$\frac{1}{\sigma^2 T}\sum_{t=1}^{T}\varepsilon_t^2 \xrightarrow{p} 1,$$

we obtain that

$$\frac{1}{\sigma^2 T}\sum_{t=1}^{T} Y_{t-1}\varepsilon_t \xrightarrow{L} \frac{1}{2}(U-1),$$

where $U \sim \chi^2(1)$ $(= \mathsf{N}(0,1)^2)$.

Because $Y_t \sim \mathsf{N}(0,\sigma^2 t)$, $\mathsf{E}\left[Y_t^2\right] = \sigma^2 t$ and

$$\mathsf{E}\left[\sum_{t=1}^{T} Y_{t-1}^2\right] = \sum_{t=1}^{T}\mathsf{E}\left[Y_{t-1}^2\right] = \sum_{t=1}^{T}\sigma^2(t-1) = \frac{1}{2}\sigma^2(T-1)T.$$

Hence, $\sum_{t=1}^{T} Y_{t-1}^2$ needs to be scaled by T^2 and since $\sum_{t=1}^{T} Y_{t-1}\varepsilon_t$ has to be scaled by T,

$$T(\widehat{\rho}-1) \xrightarrow{L} J,$$

where J is a random variable with a non-degenerate distribution. In order to determine the distribution of J, we first define the Brownian motion.

6.4.2 Brownian motion

For the random walk with standard normal distributed innovations ($\sigma^2 = 1$),

$$Y_t = Y_{t-1} + \varepsilon_t = \varepsilon_1 + \ldots + \varepsilon_t,$$

it follows that the distribution of Y_t is:

$$Y_t \sim \mathsf{N}(0,\sigma^2 t).$$

The limit of the resulting stochastic process $T^{-1/2}Y_t$, when T converges to infinity is called a *standard Brownian motion.*

Definition 12 *The standard Brownian motion, or Wiener process, $W(.)$ is a continuous time stochastic process defined for $t \in [0,1]$ that is such that:*

1. $W(0) = 0$

2. For any dates $0 \leq t_1 \leq t_2 \leq \ldots \leq t_k < 1$, the changes $W(t_2) - W(t_1)$, $W(t_3) - W(t_2), \ldots, W(t_k) - W(t_{k-1})$ are independent normal random variables and $W(t_j) - W(t_{j-1}) \sim N(0, t_j - t_{j-1})$.

3. For any given realization, $W(t)$ is continuous in t with probability one.

By multiplying the standard Brownian motion by σ,

$$Z(t) = \sigma W(t),$$

we obtain the so-called Brownian with variance σ^2. Another stochastic process that can be considered is the square of $W(t)$,

$$Z(t) = W(t)^2,$$

which is described as t times a $\chi^2(1)$ random variable.

6.4.3 Functional central limit theorem (*)

The behavior of the random walk,

$$Y_t = Y_{t-1} + \varepsilon_t = \varepsilon_1 + \ldots + \varepsilon_t,$$

can also be analyzed using the partial sums defined by

$$X_T(r) \equiv \frac{1}{T} \sum_{t=1}^{[Tr]} \varepsilon_t,$$

with $0 \leq r \leq 1$ and $[.]$ is the integral part function; then

$$X_T(r) = \begin{cases} 0 & \text{for } 0 \leq r < \frac{1}{T} \\ \frac{1}{T}\varepsilon_1 & \text{for } \frac{1}{T} \leq r < \frac{2}{T} \\ \frac{1}{T}(\varepsilon_1 + \varepsilon_2) & \text{for } \frac{2}{T} \leq r < \frac{3}{T} \\ \vdots & \vdots \\ \frac{1}{T}(\varepsilon_1 + \ldots + \varepsilon_T) & \text{for } r = 1. \end{cases}$$

We can specify $\sqrt{T}X_T(r)$ as

$$\sqrt{T}X_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} \varepsilon_t = \left(\frac{\sqrt{Tr}}{\sqrt{T}} \right) \frac{1}{\sqrt{rT}} \sum_{t=1}^{[Tr]} \varepsilon_t$$

with

$$\frac{1}{\sqrt{rT}} \sum_{t=1}^{[Tr]} \varepsilon_t \xrightarrow{L} \mathsf{N}(0, \sigma^2).$$

Hence, since $\frac{\sqrt{Tr}}{\sqrt{T}} \to \sqrt{r}$,

$$\sqrt{T}X_T(r) \xrightarrow{L} \mathsf{N}(0, r\sigma^2)$$

or

$$\frac{\sqrt{T}}{\sigma} X_T(r) \xrightarrow{L} \mathsf{N}(0, r).$$

If we use r_2 and r_1, $r_2 > r_1$, instead of r, we can obtain:

$$\frac{\sqrt{T}}{\sigma}\left(X_T(r_2) - X_T(r_1)\right) \xrightarrow{L} \mathsf{N}(0, r_2 - r_1).$$

The sequence of stochastic functions $\frac{\sqrt{T}}{\sigma}X_T(.)$ therefore converges to a standard Brownian motion:

$$\frac{\sqrt{T}}{\sigma}X_T(.) \xrightarrow{L} W(.),$$

which is known as the *functional central limit theorem.* For $r = 1$, $X_T(1)$ is just the sample mean of the series

$$X_T(1) \equiv \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t$$

and the conventional central limit theorem is obtained since

$$\frac{\sqrt{T}}{\sigma}X_T(1) = \frac{1}{\sigma\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t \xrightarrow{L} W(1) \sim \mathsf{N}(0,1).$$

So, the conventional central limit theorem is a special case of the functional central limit theorem.

6.4.4 Continuous mapping theorem and usage for unit root processes (*)

If x_t, $t = 1, \ldots$ is a sequence of random variables and $x_T \xrightarrow{L} x$ and $g : \mathbb{R} \to \mathbb{R}$ is a continuous function, then also $g(x_T) \xrightarrow{L} g(x)$. The same reasoning applies to functionals. If $g(.)$ is a continuous functional that associates a real random variable Y with the stochastic function $S(.)$. For instance, $Y = \int_0^1 S(r)dr$ and $Y = \int_0^1 S(r)^2 dr$ represent continuous functionals. The *continuous mapping theorem* states that if $S_T(.) \xrightarrow{L} S(.)$ that $g(S_T(.)) \xrightarrow{L} g(S(.))$. An example is the function:

$$S_T(r) = [\sqrt{T}X_T(r)]^2.$$

Since $\frac{\sqrt{T}}{\sigma}X_T(.) \xrightarrow{L} W(.)$, it follows from the functional central limit theorem that

$$S_T(.) \xrightarrow{L} \sigma^2[W(.)]^2.$$

When we consider the random walk

$$Y_t = Y_{t-1} + \varepsilon_t,$$

we can denote $X_T(r)$ $(= \frac{1}{T}\sum_{t=1}^{[Tr]}\varepsilon_t)$ by

$$X_T(r) = \begin{cases} 0 & \text{for } 0 \le r < \frac{1}{T} \\ \frac{1}{T}Y_1 & \text{for } \frac{1}{T} \le r < \frac{2}{T} \\ \frac{1}{T}Y_2 & \text{for } \frac{2}{T} \le r < \frac{3}{T} \\ \vdots & \\ \frac{1}{T}Y_T & \text{for } r = 1. \end{cases}$$

and $X_T(r)$ is a step function. The area under the step function equals sum of the value at the different stepping points and the distance between the different steps:

$$\int_0^1 X_T(r)dr = \frac{1}{T}\frac{1}{T}Y_1 + \ldots + \frac{1}{T}\frac{1}{T}Y_{T-1} = \frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}.$$

When we pre-multiply both sides by $\sqrt{T}$, we obtain

$$\int_0^1 \sqrt{T}X_T(r)dr = \frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} Y_{t-1}.$$

Since $\sqrt{T}X_T(.) \xrightarrow{L} \sigma W(.)$, we obtain from the continuous mapping theorem that

$$\int_0^1 \sqrt{T}X_T(r)dr \xrightarrow{L} \sigma \int_0^1 W(r)dr$$

so

$$\frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} Y_{t-1} \xrightarrow{L} \sigma \int_0^1 W(r)dr.$$

The issue we are facing here is that whereas we have seen that $W(t) \sim \mathsf{N}(0, r)$, we do not know what the distribution of $\int_0^1 W(r)dr$ is. We show below that

$$\int_0^1 W(r)dr \sim \mathsf{N}\left(0, \frac{1}{3}\right).$$

Proof. Rewriting $T^{-\frac{3}{2}}\sum_{t=1}^{T} Y_{t-1}$, we can also obtain that

$$\begin{aligned}
\frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} Y_{t-1} &= \frac{1}{T^{\frac{3}{2}}}[\varepsilon_1 + (\varepsilon_1 + \varepsilon_2) + \ldots + (\varepsilon_1 + \ldots + \varepsilon_{T-1})] \\
&= \frac{1}{T^{\frac{3}{2}}}[(T-1)\varepsilon_1 + (T-2)\varepsilon_2 + \ldots + \varepsilon_{T-1}] \\
&= \frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T}(T-t)\varepsilon_t \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t - \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{t}{T}\varepsilon_t.
\end{aligned}$$

The random variables ε_t and $\frac{t}{T}\varepsilon_t$ have mean zero and satisfy a central limit theorem:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{pmatrix} \varepsilon_t \\ \frac{t}{T}\varepsilon_t \end{pmatrix} \xrightarrow{L} K,$$

with $K \sim N(0, Q)$ and

$$\begin{aligned}
Q = & \lim_{T\to\infty} \tfrac{1}{T} E\left[\sum_{t=1}^{T}\sum_{j=1}^{T}\begin{pmatrix} \varepsilon_t \\ \frac{t}{T}\varepsilon_t \end{pmatrix}\begin{pmatrix} \varepsilon_j \\ \frac{j}{T}\varepsilon_j \end{pmatrix}'\right] \\
= & \lim_{T\to\infty} \tfrac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} \sigma^2 & \frac{t}{T}\sigma^2 \\ \frac{t}{T}\sigma^2 & \frac{t^2}{T^2}\sigma^2 \end{pmatrix} \\
= & \lim_{T\to\infty} \sigma^2 \begin{pmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & \frac{1}{T^{\frac{3}{2}}} \end{pmatrix}\left[\sum_{t=1}^{T}\begin{pmatrix} 1 & t \\ t & t^2 \end{pmatrix}\right]\begin{pmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & \frac{1}{T^{\frac{3}{2}}} \end{pmatrix} \\
= & \lim_{T\to\infty} \sigma^2 \begin{pmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & \frac{1}{T^{\frac{3}{2}}} \end{pmatrix}\begin{pmatrix} T & \frac{1}{2}T(T+1) \\ \frac{1}{2}T(T+1) & \frac{1}{6}T(T+1)(2T+1) \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & \frac{1}{T^{\frac{3}{2}}} \end{pmatrix} \\
= & \lim_{T\to\infty} \sigma^2 \begin{pmatrix} 1 & \frac{1}{2T^2}T(T+1) \\ \frac{1}{2T^2}T(T+1) & \frac{1}{6T^3}T(T+1)(2T+1) \end{pmatrix} \\
= & \sigma^2 \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix},
\end{aligned}$$

since $\sum_{t=1}^{T} t = \frac{1}{2}T(T+1)$ and $\sum_{t=1}^{T} t^2 = \frac{1}{6}T(T+1)(2T+1)$. This implies that

$$\begin{aligned}
\tfrac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} Y_{t-1} = & \tfrac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t - \tfrac{1}{\sqrt{T}}\sum_{t=1}^{T}\tfrac{t}{T}\varepsilon_t \\
\xrightarrow{L} & \sigma\int_0^1 W(r)dr \\
\xrightarrow{L} & H,
\end{aligned}$$

with $H = \left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)' K \sim N(0, q)$ and $q = \left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)' Q \left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right) = \frac{1}{3}\sigma^2$. □

From the above expression, we also obtain that

$$\begin{aligned}
\tfrac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} t\varepsilon_t = & \tfrac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t - \tfrac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} Y_{t-1} \\
\xrightarrow{L} & \sigma W(1) - \sigma\int_0^1 W(r)dr \\
\sim & N(0, \tfrac{1}{3}\sigma^2)
\end{aligned}$$

since $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t \xrightarrow{L} \sigma W(1)$.

The limit behavior of the sum of squares of the random walk Y_t,

$$S_T(r) \equiv T[X_T(r)]^2,$$

is obtained by specifying $S_T(r)$ as

$$S_T(r) = \begin{cases} 0 & \text{for } 0 \le r < \frac{1}{T} \\ \frac{1}{T}Y_1^2 & \text{for } \frac{1}{T} \le r < \frac{2}{T} \\ \frac{1}{T}Y_2^2 & \text{for } \frac{2}{T} \le r < \frac{3}{T} \\ \vdots & \\ \frac{1}{T}Y_T^2 & \text{for } r = 1. \end{cases}$$

The integral over the functional $S_T(.)$ then reads

$$\int_0^1 S_T(r)dr = \tfrac{1}{T}\tfrac{1}{T}Y_1^2 + \ldots + \tfrac{1}{T}\tfrac{1}{T}Y_{T-1}^2 = \tfrac{1}{T^2}\sum_{t=1}^T Y_{t-1}^2$$

and the continuous mapping theorem implies that

$$\begin{aligned} \tfrac{1}{T^2}\sum_{t=1}^T Y_{t-1}^2 = & \int_0^1 S_T(r)dr \\ \xrightarrow{L} & \sigma^2 \int_0^1 W(r)^2 dr. \end{aligned}$$

In a similar manner it can be derived that:

$$\begin{aligned} \tfrac{1}{T^{\frac{5}{2}}}\sum_{t=1}^T tY_{t-1} = & \tfrac{1}{T^{\frac{3}{2}}}\sum_{t=1}^T \tfrac{t}{T}Y_{t-1} \\ = & \tfrac{1}{\sqrt{T}}\sum_{t=1}^T \tfrac{1}{T}\tfrac{t}{T}Y_{t-1} \\ = & \int_0^1 rX_T(r)dr \\ \xrightarrow{L} & \sigma\int_0^1 rW(r)dr \end{aligned}$$

and

$$\begin{aligned} \tfrac{1}{T^3}\sum_{t=1}^T tY_{t-1}^2 = & \tfrac{1}{T^2}\sum_{t=1}^T \tfrac{t}{T}Y_{t-1}^2 \\ = & \tfrac{1}{T}\sum_{t=1}^T \tfrac{1}{T}\tfrac{t}{T}Y_{t-1}^2 \\ = & \int_0^1 rS_T(r)^2 dr \\ \xrightarrow{L} & \sigma\int_0^1 rW(r)^2 dr. \end{aligned}$$

We use these expressions to obtain the limit distribution of the least squares estimator of the AR parameter when a unit root is present. But before we introduce the stochastic integral

6.4.5 Stochastic (or Itô) integral (*)

In the previous subsection, we have seen that for a functional $g(\cdot)$ and Y_t that follows a random walk

$$T^{-1}\sum_{t=1}^T g\left(\frac{Y_t}{\sqrt{T}}\right) \xrightarrow{L} \sigma\int_0^1 g(W(r))\,dr.$$

The integral to the right-hand side of the previous expression is of the usual Riemann integral type. Indeed, if we let $t_{k,T} = \frac{k}{T}$ the points used for the evaluation of

$$\begin{aligned} T^{-1}\sum_{t=1}^T g\left(\frac{Y_t}{\sqrt{T}}\right) &= \sum_{k=1}^T g\left(\frac{Y_{t_{k,T}}}{\sqrt{T}}\right)(t_{k,T} - t_{k-1,T}) \\ &= \sum_{k=1}^T g\left(\frac{Y_{t_{k,T}}}{\sqrt{T}}\right)\Delta_k \end{aligned}$$

where Δ_k is the length of the interval $[t_{k-1,T}, t_{k,T}]$. For any sequence of timing points (partition) $\tau_T = \{t_{k,T}\}$ such that

$$mesh\{\tau_T\} = \max_{k=1,\ldots,T}\Delta_k \underset{T\to\infty}{\to} 0$$

the sum converges to the same value, the ordinary Riemann integral

$$\sigma \int_0^1 g\left(W\left(r\right)\right) dr$$

The previous integral can be extended to define the **Riemann-Stieltjes** integral

$$\sum_{k=1}^{T} g\left(t_{k-1}^{*}\right)\left[h\left(t_{k,T}\right) - h\left(t_{k-1,T}\right)\right] \rightarrow \int_{r=0}^{1} g\left(r\right) dh\left(r\right)$$

where $t_{k-1}^{*} \in [t_{k-1}, t_k]$. If g and h have discontinuities at the same points, and h has bounded variation

$$\sup_{\tau} \sum_{k=1}^{T} \left|h\left(t_k\right) - h\left(t_{k-1}\right)\right| < \infty$$

then $\int_{r=0}^{1} g\left(W\left(r\right)\right) dh\left(r\right)$ exists. In general, it exists if h is differentiable, but this is very restrictive. The Riemann-Stieltjes integral allows for integration with respect to a Brownian motion

$$\sum_{k=1}^{T} g\left(r\right)\left[W\left(t_{k,T}\right) - W\left(t_{k-1,T}\right)\right] \xrightarrow{L} \int_{r=0}^{1} g\left(r\right) dW\left(r\right)$$

Now unfortunately

$$\sum_{k=1}^{T} W\left(t_{k-1}^{*}\right)\left[W\left(t_{k,T}\right) - W\left(t_{k-1,T}\right)\right]$$

does not converge for all partitions τ_k, hence $\int_0^1 W\left(r\right) dW\left(r\right)$ **cannot be defined in the Riemann-Stieltjes sense**. We need a modification: the Itô integral which is defined as the limit of

$$\sum_{k=1}^{T} W\left(t_{k-1}\right)\left[W\left(t_{k,T}\right) - W\left(t_{k-1,T}\right)\right]$$

where the integrand W is evaluated at the **left end-point of** $[t_{k-1}, t_k]$. Then

$$\sum_{k=1}^{T} W\left(t_{k-1}\right)\left[W\left(t_{k,T}\right) - W\left(t_{k-1,T}\right)\right] \xrightarrow{m.s} \int_0^1 W\left(r\right) dW\left(r\right).$$

The Itô integral does not satisfy the usual chain rule, e.g.

$$\int_0^s W\left(r\right) dW\left(r\right) = \frac{W^2\left(s\right) - s}{2}$$

whereas the Stratonovitch integral does: this is defined as the limit

$$\sum_{k=1}^{T} W\left(\frac{t_{k-1,T} + t_{k,T}}{2}\right)\left[W\left(t_{k,T}\right) - W\left(t_{k-1,T}\right)\right] \xrightarrow{m.s} \int_0^1 W\left(r\right) \circ dW\left(r\right) = \frac{1}{2} W^2\left(1\right)$$

Now, what about the limit of $\sum_{t=1}^{T} Y_{t-1}\varepsilon_t$? From the results above, we see that

$$\begin{array}{rcl} T^{-1}\sum_{t=1}^{T} Y_{t-1}\varepsilon_t & = & \sum_{t=1}^{T} \frac{Y_{t-1}}{\sqrt{T}} \Delta\left(\frac{Y_t}{\sqrt{T}}\right) \\ & \xrightarrow{L} & \sigma^2 \int_0^1 W\left(r\right) dW\left(r\right) \\ & = & \frac{\sigma^2}{2}\left(W^2\left(1\right) - 1\right) \end{array}$$

6.5 Limit behavior least squares estimator when unit root is present

6.5.1 Estimation of the unit-root

We discuss the distribution of the least squares estimator for four different specifications of the AR(1) model. For the first two specifications, the data generating process is the standard random walk model. For the latter two specifications, the random walk with drift is the data generating process.

For the random walk model:

$$Y_t = Y_{t-1} + \varepsilon_t,$$

with ε_t independent with mean zero and variance σ^2, the limit behavior of the least squares estimator in the AR(1) model,

$$Y_t = \rho Y_{t-1} + \varepsilon_t,$$

that reads

$$\begin{aligned}\widehat{\rho} = & \left(\sum_{t=1}^{T} Y_{t-1}^2\right)^{-1} \sum_{t=1}^{T} Y_{t-1}Y_t = \left(\sum_{t=1}^{T} Y_{t-1}^2\right)^{-1} \left(\sum_{t=1}^{T} Y_{t-1}(Y_{t-1}+\varepsilon_t)\right) \\ = & \ 1 + \left(\sum_{t=1}^{T} Y_{t-1}^2\right)^{-1} \left(\sum_{t=1}^{T} Y_{t-1}\varepsilon_t\right),\end{aligned}$$

follows from the limit behavior of $\sum_{t=1}^{T} Y_{t-1}^2$ and $\sum_{t=1}^{T} Y_{t-1}\varepsilon_t$. We have previously derived that

$$\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2 \xrightarrow{L} \sigma^2 \int_0^1 W(r)^2 dr$$

and

$$\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\varepsilon_t = \frac{1}{2}(\frac{Y_T}{\sqrt{T}})^2 - \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2, \xrightarrow{L} \frac{1}{2}\sigma^2(W(1)^2-1)$$

Therefore,

$$\frac{\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\varepsilon_t}{\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2} \xrightarrow{L} \frac{\frac{1}{2}(W(1)^2-1)}{\int_0^1 W(r)^2 dr}$$

and thus

$$T(\widehat{\rho}-1) \xrightarrow{L} \frac{\frac{1}{2}(W(1)^2-1)}{\int_0^1 W(r)^2 dr}.$$

which is non Gaussian. The t-statistic of $\widehat{\rho}$ that tests the hypothesis $\mathrm{H}_0 : \rho = 1$ is defined as

$$t_{\widehat{\rho}} = \frac{\widehat{\rho}-1}{\sqrt{s^2(\sum_{t=1}^{T} Y_{t-1}^2)^{-1}}},$$

with $s^2 = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \widehat{\rho}Y_{t-1})^2$. Since $s^2 \xrightarrow{p} \sigma^2$ and $\frac{1}{T^2}\sum_{t=1}^{T}Y_{t-1}^2 \xrightarrow{L} \sigma^2 \int_0^1 W(r)^2 dr$, the limit behavior of the t-statistic under H_0 is

$$\begin{aligned} t_{\widehat{\rho}} = & \frac{T(\widehat{\rho}-1)}{\sqrt{s^2(\frac{1}{T^2}\sum_{t=1}^{T}Y_{t-1}^2)^{-1}}} \\ \xrightarrow{L} & \frac{\frac{1}{2}(W(1)^2-1)}{\sqrt{\int_0^1 W(r)^2 dr}} = \frac{\frac{1}{2}(W(1)^2-1)}{\sqrt{\int_0^1 W(r)^2 dr}}. \end{aligned}$$

The critical values of this distribution have been tabulated by Dickey and Fuller which explains why we refer to this t-statistic as the Dickey-Fuller statistic. Since we rule out explosive behavior a priori, we typically do only an one sided test so we test $H_0 : \rho = 1$ against $H_1 : \rho < 1$, the 95% critical value of this statistic is -1.95 and the 99% critical value is -2.58.

6.5.2 Estimating a constant (*)

The model without a constant that is estimated in the previous example is not the appropriate model to compare with the random walk model since it has a mean equal to zero while the random walk model has a stochastic mean. The limit behavior of the least squares estimator in the model:

$$Y_t = \tau + \rho Y_{t-1} + \varepsilon_t,$$

while the data is generated by the random walk:

$$Y_t = Y_{t-1} + \varepsilon_t,$$

is therefore of more practical importance. The limit behavior of the least squares estimator of $(\tau \; \rho)'$

$$\begin{aligned} \begin{bmatrix} \widehat{\tau} \\ \widehat{\rho} \end{bmatrix} &= \left(\sum_{t=1}^{T}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}'\right)^{-1}\left(\sum_{t=1}^{T}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}Y_t\right) \\ &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \left(\sum_{t=1}^{T}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}'\right)^{-1}\left(\sum_{t=1}^{T}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}\varepsilon_t\right), \end{aligned}$$

is determined by the limit behavior of $\sum_{t=1}^{T}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}'$ and $\sum_{t=1}^{T}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}\varepsilon_t$. The limit behavior of

$$\sum_{t=1}^{T}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}\begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}' = \sum_{t=1}^{T}\begin{bmatrix} 1 & Y_{t-1} \\ Y_{t-1} & Y_{t-1}^2 \end{bmatrix}$$

results from the limit behavior of each of its elements: $\frac{1}{T}\sum_{t=1}^{T} 1 = 1$, and

$$\frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} Y_{t-1} \xrightarrow{L} \sigma \int_0^1 W(r)dr,$$
$$\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2 \xrightarrow{L} \sigma^2 \int_0^1 W(r)^2 dr,$$

so

$$\begin{bmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & \frac{1}{T} \end{bmatrix} \sum_{t=1}^{T} \begin{bmatrix} 1 & Y_{t-1} \\ Y_{t-1} & Y_{t-1}^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & \frac{1}{T} \end{bmatrix}$$
$$\xrightarrow{L} \begin{bmatrix} 1 & \sigma \int_0^1 W(r)dr \\ \sigma \int_0^1 W(r)dr & \sigma^2 \int_0^1 W(r)^2 dr \end{bmatrix}.$$

The limit behavior of $\sum_{t=1}^{T} \begin{pmatrix} 1 \\ Y_{t-1} \end{pmatrix} \varepsilon_t$ results from the limit behavior of $\sum_{t=1}^{T} \varepsilon_t$ and $\sum_{t=1}^{T} Y_{t-1}\varepsilon_t$:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \varepsilon_t \xrightarrow{L} \sigma W(1), \qquad \frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\varepsilon \xrightarrow{L} \frac{1}{2}\sigma^2(W(1)^2 - 1),$$

so

$$\begin{bmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & \frac{1}{T} \end{bmatrix} \sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \varepsilon_t \xrightarrow{L} \begin{bmatrix} \sigma W(1) \\ \frac{1}{2}\sigma^2(W(1)^2 - 1) \end{bmatrix}.$$

Combining all results, the limit behavior of the least squares estimator $(\widehat{\tau}, \widehat{\rho})'$ becomes:

$$\begin{bmatrix} \sqrt{T} & 0 \\ 0 & T \end{bmatrix} \left(\begin{bmatrix} \widehat{\tau} \\ \widehat{\rho} \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)$$
$$\xrightarrow{L} \begin{bmatrix} 1 & \sigma \int_0^1 W(r)dr \\ \sigma \int_0^1 W(r)dr & \sigma^2 \int_0^1 W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} \sigma W(1) \\ \frac{1}{2}\sigma^2(W(1)^2 - 1) \end{bmatrix}$$
$$= \begin{bmatrix} \sigma & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \int_0^1 W(r)dr \\ \int_0^1 W(r)dr & \int_0^1 W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \frac{1}{2}(W(1)^2 - 1) \end{bmatrix}.$$

Using properties of the partitioned inverse, we can derive the expression for the inverse:

$$\begin{bmatrix} 1 & \int_0^1 W(r)dr \\ \int_0^1 W(r)dr & \int_0^1 W(r)^2 dr \end{bmatrix}^{-1}$$
$$= \frac{1}{\int_0^1 W(r)^2 dr - (\int_0^1 W(r)dr)^2} \begin{bmatrix} \int_0^1 W(r)^2 dr & -\int_0^1 W(r)dr \\ -\int_0^1 W(r)dr & 1 \end{bmatrix}$$

and therefore

$$\sqrt{T}\widehat{\tau} \xrightarrow{L} \sigma \frac{W(1)\int_0^1 W(r)^2 dr - \frac{1}{2}(W(1)^2 - 1)\int_0^1 W(r)dr}{\int_0^1 W(r)^2 dr - (\int_0^1 W(r)dr)^2}$$
$$T(\widehat{\rho} - 1) \xrightarrow{L} \frac{\frac{1}{2}(W(1)^2 - 1) - W(1)\int_0^1 W(r)dr}{\int_0^1 W(r)^2 dr - (\int_0^1 W(r)dr)^2}.$$

The limiting distribution of the t-statistic that test $H_0 : \rho = 1$ results in a similar manner:

$$\begin{aligned} t_{\widehat{\rho}} &= \frac{(\widehat{\rho}-1)}{\sqrt{s^2(\sum_{t=1}^{T} Y_{t-1}^2)^{-1}}} = \frac{T(\widehat{\rho}-1)}{\sqrt{s^2(\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2)^{-1}}} \\ &\xrightarrow{L} \frac{\frac{\frac{1}{2}(W(1)^2-1)-W(1)\int_0^1 W(r)dr}{\int_0^1 W(r)^2dr-(\int_0^1 W(r)dr)^2}}{\sqrt{\sigma^2(\sigma^2[\int_0^1 W(r)^2dr-(\int_0^1 W(r)dr)^2])^{-1}}} \\ &= \frac{\frac{1}{2}(W(1)^2-1)-W(1)\int_0^1 W(r)dr}{\sqrt{\int_0^1 W(r)^2dr-(\int_0^1 W(r)dr)^2}}, \end{aligned}$$

where $Y_t = Y_t - \frac{1}{T}\sum_{t=1}^{T} Y_t$ and $\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2 = \frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2 - \left[\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\right]^2 \xrightarrow{L} \int_0^1 W(r)^2dr - (\int_0^1 W(r)dr)^2$.

The critical values of this distribution have been tabulated by Dickey and Fuller. Again since we rule out explosive behavior a priori, we typically do only a one sided test so we test $H_0 : \rho = 1$ against $H_1 : \rho < 1$, the 95% critical value of this statistic is -2.86 and the 99% critical value is -3.43.

6.5.3 Estimating a random walk with drift (*)

When the random walk with drift:

$$Y_t = \tau + Y_{t-1} + \varepsilon_t,$$

is the data generating process, the limit behavior of Y_t changes because recurrent substitution gives

$$Y_t = \tau t + \xi_t + Y_0$$

with $\xi_t = \xi_{t-1} + \varepsilon_t = \sum_{t=1}^{T} \varepsilon_t$, so

$$\sum_{t=1}^{T} Y_{t-1} = \sum_{t=1}^{T} \tau(t-1) + \xi_t + Y_0.$$

Since $\frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} \xi_{t-1} \xrightarrow{L} \sigma\int_0^1 W(r)dr$ and

$$\begin{aligned} \frac{1}{T^2}\sum_{t=1}^{T} \tau(t-1) = & \quad \frac{1}{T^2}\tau\frac{1}{2}(T-1)T \\ \xrightarrow{p} & \quad \frac{1}{2}\tau, \end{aligned}$$

we obtain that

$$\frac{1}{T^2}\sum_{t=1}^{T} Y_t = \frac{1}{T^2}\sum_{t=1}^{T} \xi_{t-1} + \frac{1}{T^2}\sum_{t=1}^{T} \tau(t-1) \xrightarrow{p} \frac{1}{2}\tau,$$

which differs from the expression that we obtained for the standard random walk model. Similarly, for the square of Y_t,

$$Y_t^2 = (\tau t + \xi_t + Y_0)^2 = \tau^2 t^2 + \xi_t^2 + Y_0^2 + 2\tau t(\xi_t + Y_0) + 2\xi_t Y_0,$$

we obtain that

$$\sum_{t=1}^{T} Y_{t-1}^2 = \sum_{t=1}^{T} \tau^2 (t-1)^2 + \xi_{t-1}^2 + Y_0^2 + 2\tau(t-1)(\xi_{t-1} + Y_0) + 2\xi_{t-1} Y_0.$$

The first term in this expression is of the order T^3, since $\sum_{t=1}^{T}(t-1)^2 = \sum_{t=1}^{T-1} t^2 = \frac{1}{6}(T-1)T(2T-1)$. All other terms are at most of the order $T^{\frac{5}{2}}$ and therefore cancel out when we normalize by $\frac{1}{T^3}$. Hence,

$$\frac{1}{T^3} \sum_{t=1}^{T} Y_{t-1}^2 \xrightarrow{p} \frac{1}{3}\tau^2.$$

A similar result can be derived for $Y_{t-1}\varepsilon_t$:

$$Y_{t-1}\varepsilon_t = (\tau(t-1) + \xi_{t-1} + Y_0)\varepsilon_t$$

such that

$$\sum_{t=1}^{T} Y_{t-1}\varepsilon_t = \sum_{t=1}^{T} \tau(t-1)\varepsilon_t + \xi_{t-1}\varepsilon_t + Y_0\varepsilon_t.$$

Also here the first term needs to be normalized by $\frac{1}{T^{\frac{3}{2}}}$ while the normalization factors of the others are at most $\frac{1}{T}$. So

$$\frac{1}{T^{\frac{3}{2}}} \sum_{t=1}^{T} Y_{t-1}\varepsilon_t \xrightarrow{L} N(0, \frac{\sigma^2}{3}\tau^2).$$

Combining all these elements, we can construct the limit behavior of the least squares estimator of $(\tau\ \rho)'$ in the AR(1) model

$$Y_t = \tau + \rho Y_{t-1} + \varepsilon_t,$$

that reads

$$\begin{aligned}
\begin{bmatrix} \widehat{\tau} \\ \widehat{\rho} \end{bmatrix} &= \left(\sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}' \right)^{-1} \sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} Y_t \\
&= \begin{bmatrix} \tau \\ 1 \end{bmatrix} + \left(\sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}' \right)^{-1} \sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \varepsilon_t.
\end{aligned}$$

This limit behavior is determined by $\sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}'$ and $\sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \varepsilon_t$. The limit behavior of

$$\sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}' = \sum_{t=1}^{T} \begin{bmatrix} t & Y_{t-1} \\ Y_{t-1} & Y_{t-1}^2 \end{bmatrix}$$

can be described using the above results by

$$\begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T^{\frac{3}{2}} \end{bmatrix} \sum_{t=1}^{T} \begin{bmatrix} t & Y_{t-1} \\ Y_{t-1} & Y_{t-1}^2 \end{bmatrix} \begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T^{\frac{3}{2}} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} 1 & \frac{1}{2}\tau \\ \frac{1}{2}\tau & \frac{1}{3}\tau^2 \end{bmatrix}$$

and the limit behavior of $\sum_{t=1}^{T} \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix} \varepsilon_t$ results as

$$\begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T^{\frac{3}{2}} \end{bmatrix} \sum_{t=1}^{T} \begin{bmatrix} \varepsilon_t \\ Y_{t-1}\varepsilon_t \end{bmatrix} \xrightarrow{L} \mathsf{N}(0, \sigma^2 P),$$

with

$$\begin{aligned} P &= \lim_{T\to\infty} \begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T^{\frac{3}{2}} \end{bmatrix} \sum_{t=1}^{T} \begin{bmatrix} 1 & \tau(t-1) \\ \tau(t-1) & \tau^2(t-1)^2 \end{bmatrix} \begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T^{\frac{3}{2}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{1}{2}\tau \\ \frac{1}{2}\tau & \frac{1}{3}\tau^2 \end{bmatrix}. \end{aligned}$$

The expression of P is identical to the limit behavior of the normalized expression for $\sum_{t=1}^{T} \begin{bmatrix} t & Y_{t-1} \\ Y_{t-1} & Y_{t-1}^2 \end{bmatrix}$ and we therefore obtain the limit behavior of the least squares estimator $\begin{bmatrix} \widehat{\tau} \\ \widehat{\rho} \end{bmatrix}$,

$$\begin{bmatrix} T^{\frac{1}{2}} & 0 \\ 0 & T^{\frac{3}{2}} \end{bmatrix} \left(\begin{bmatrix} \widehat{\tau} \\ \widehat{\rho} \end{bmatrix} - \begin{bmatrix} \tau \\ 1 \end{bmatrix} \right) \xrightarrow{L} \mathsf{N}(0, \sigma^2 P^{-1}).$$

Unlike the distribution of the least squares estimator that we constructed for the other specifications, the limiting distribution of the least squares estimator is normal.

6.5.4 Estimating with a linear trend (*)

When a random walk with drift:

$$Y_t = \tau + Y_{t-1} + \varepsilon_t,$$

is the data generating process, a linear trend is present in the series and it is therefore more natural to estimate the model

$$Y_t = \tau + \delta t + \rho Y_{t-1} + \varepsilon_t.$$

since a linear trend is present in the data. To analyze the limiting distribution of the least squares estimator in the AR model with constant and trend, it is convenient to re-specify this model towards

$$Y_t = \tau^* + \delta^* t + \rho^* \xi_{t-1} + \varepsilon_t,$$

with $\tau^* = (1-\rho)\tau$, $\rho^* = \rho$, $\lambda^* = \lambda + \rho\tau$ and $\xi_t = Y_t - \tau t$. When the random walk with drift is the data generating process, ξ_t is a standard random walk, $\xi_t = \xi_{t-1} + \varepsilon_t = \sum_{i=1}^{t} \varepsilon_i + Y_0$. We construct the limiting distribution of the least squares estimator of τ^*, δ^*, ρ^* under the assumption that $\tau = \tau_0$ such that we can construct ξ_t directly. This would correspond to the hypothesis H_0 : $\rho = 0, \tau = \tau_0$.

$$\begin{aligned}\begin{bmatrix} \widehat{\tau}^* \\ \widehat{\delta}^* \\ \widehat{\rho}^* \end{bmatrix} &= \left(\sum_{t=1}^{T} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix}' \right)^{-1} \sum_{t=1}^{T} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix} Y_t \\ &= \begin{bmatrix} 0 \\ \tau_0 \\ 1 \end{bmatrix} + \left(\sum_{t=1}^{T} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix}' \right)^{-1} \sum_{t=1}^{T} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix} \varepsilon_t .\end{aligned}$$

The limiting behavior of

$$\sum_{t=1}^{T} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix}' = \sum_{t=1}^{T} \begin{bmatrix} 1 & t & \xi_{t-1} \\ t & t^2 & t\xi_{t-1} \\ \xi_{t-1} & t\xi_{t-1} & \xi_{t-1}^2 \end{bmatrix}$$

results from the limiting behavior of its different elements:

$$\frac{1}{T}\sum_{t=1}^{T} 1 = 1, \qquad \frac{1}{T^2}\sum_{t=1}^{T} t \xrightarrow{p} \frac{1}{2},$$

$$\frac{1}{T^3}\sum_{t=1}^{T} t^2 \xrightarrow{L} \frac{1}{3}, \qquad \frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} \xi_{t-1} \xrightarrow{L} \sigma \int_0^1 W(r)dr$$

$$\frac{1}{T^{\frac{5}{2}}}\sum_{t=1}^{T} t\xi_{t-1} \xrightarrow{L} \sigma \int_0^1 rW(r)dr, \qquad \frac{1}{T^2}\sum_{t=1}^{T} \xi_{t-1}^2 \xrightarrow{L} \sigma^2 \int_0^1 W(r)^2 dr.$$

Similarly, the limiting behavior of

$$\sum_{t=1}^{T} \begin{bmatrix} 1 \\ t \\ \xi_{t-1} \end{bmatrix} \varepsilon = \sum_{t=1}^{T} \begin{bmatrix} \varepsilon_t \\ t\varepsilon_t \\ \xi_{t-1}\varepsilon_t \end{bmatrix},$$

is determined by the limiting behavior of its elements:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \varepsilon_t \xrightarrow{L} \sigma W(1), \qquad \frac{1}{T^{\frac{3}{2}}}\sum_{t=1}^{T} t\varepsilon_t \xrightarrow{L} \sigma[W(1) - \int_0^1 W(r)dr],$$

$$\frac{1}{T}\sum_{t=1}^{T} \xi_{t-1}\varepsilon_t \xrightarrow{L} \frac{1}{2}\sigma^2(W(1)^2 - 1).$$

So,

$$\begin{bmatrix} \frac{1}{\sqrt{T}} & 0 & 0 \\ 0 & \frac{1}{T^{\frac{3}{2}}} & 0 \\ 0 & 0 & \frac{1}{\sigma T} \end{bmatrix} \sum_{t=1}^{T} \begin{bmatrix} 1 & t & \xi_{t-1} \\ t & t^2 & t\xi_{t-1} \\ \xi_{t-1} & t\xi_{t-1} & \xi_{t-1}^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{T}} & 0 & 0 \\ 0 & \frac{1}{T^{\frac{3}{2}}} & 0 \\ 0 & 0 & \frac{1}{\sigma T} \end{bmatrix}$$

$$\xrightarrow{L} \begin{bmatrix} 1 & \frac{1}{2} & \int_0^1 W(r)dr \\ \frac{1}{2} & \frac{1}{3} & \int_0^1 rW(r)dr \\ \int_0^1 W(r)dr & \int_0^1 rW(r)dr & \int_0^1 W(r)^2dr \end{bmatrix}$$

and

$$\begin{bmatrix} \frac{1}{\sigma\sqrt{T}} & 0 & 0 \\ 0 & \frac{1}{\sigma T^{\frac{3}{2}}} & 0 \\ 0 & 0 & \frac{1}{\sigma^2 T} \end{bmatrix} \sum_{t=1}^{T} \begin{bmatrix} \varepsilon_t \\ t\varepsilon_t \\ \xi_{t-1}\varepsilon_t \end{bmatrix} \xrightarrow{L} \begin{bmatrix} W(1) \\ W(1) - \int_0^1 W(r)dr \\ \frac{1}{2}(W(1)^2 - 1) \end{bmatrix}.$$

The limiting behavior of the least squares estimator then is given by

$$\begin{pmatrix} \sqrt{T} & 0 & 0 \\ 0 & T^{\frac{3}{2}} & 0 \\ 0 & 0 & T \end{pmatrix} \left[\begin{pmatrix} \widehat{\tau}^* \\ \widehat{\delta}^* \\ \widehat{\rho}^* \end{pmatrix} - \begin{pmatrix} 0 \\ \tau_0 \\ 1 \end{pmatrix} \right] \xrightarrow{L} \begin{pmatrix} \sigma & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\times \begin{pmatrix} 1 & \frac{1}{2} & \int_0^1 W(r)dr \\ \frac{1}{2} & \frac{1}{3} & \int_0^1 rW(r)dr \\ \int_0^1 W(r)dr & \int_0^1 rW(r)dr & \int_0^1 W(r)^2dr \end{pmatrix}^{-1} \begin{pmatrix} W(1) \\ W(1) - \int_0^1 W(r)dr \\ \frac{1}{2}(W(1)^2 - 1) \end{pmatrix}.$$

The limiting distribution of the t-statistic of $\widehat{\rho}^*$ that tests the hypothesis $\mathrm{H}_0 : \rho^* = 1$, that is

$$t_{\widehat{\rho}} = \frac{(\widehat{\rho}^* - 1)}{\sqrt{s^2(\sum_{t=1}^{T} \widehat{Y}_{t-1}^2)^{-1}}},$$

where s^2 is the least squares estimator of σ^2, $s^2 = \frac{1}{T-2}\sum_{t=1}^{T}(Y_t - \widehat{\tau}^* - \widehat{\delta}^* t - \widehat{\rho}^* \xi_{t-1})^2$ and $\widehat{Y}_{t-1} = Y_t - \widehat{a} - \widehat{b}t$, with $\binom{\widehat{a}}{\widehat{b}} = \left(\sum_{t=1}^{T} \binom{1}{t}\binom{1}{t}'\right)^{-1} \sum_{t=1}^{T} \binom{1}{t} Y_t$. Since

$$\frac{1}{T^2}\sum_{t=1}^{T} \widehat{Y}_{t-1}^2$$

$$\xrightarrow{L} \sigma^2 \int_0^1 W(r)^2 dr - \sigma^2 \begin{bmatrix} \int_0^1 W(r)dr \\ \int_0^1 rW(r)dr \end{bmatrix}' \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}^{-1} \begin{bmatrix} \int_0^1 W(r)dr \\ \int_0^1 rW(r)dr \end{bmatrix}$$

the limit behavior of the t-statistic results as the product of the limit behavior of $T(\widehat{\rho}^* - 1)$ and the square root of the limit behavior of $\sum_{t=1}^{T} \hat{y}_{t-1}^2$ divided by σ. The critical values of this distribution have been tabulated by Dickey and Fuller, we typically do only a one sided test so we test $\mathrm{H}_0 : \rho = 1$ against $\mathrm{H}_1 : \rho < 1$, the 95% critical value of this statistic is -3.41 and the 99% critical value is -3.96.

6.5.5 Summary of the different cases

Summarizing, we distinguish four different cases with two possible data generating processes:

1. Data generating process is random walk:

$$Y_t = Y_{t-1} + \varepsilon_t.$$

   (a) Estimated model is:

$$Y_t = \rho Y_{t-1} + \varepsilon_t.$$

   The critical values that result from the t-statistic that tests $\rho = 1$ equal: -1.95 (95%) and -2.58 (99%).

   (b) Estimated model is:

$$Y_t = \tau + \rho Y_{t-1} + \varepsilon_t.$$

   The critical values that result from the t-statistic that tests $\rho = 1$ equal: -2.86 (95%) and -3.43 (99%).

2. Data generating process is random walk with drift:

$$Y_t = \tau + Y_{t-1} + \varepsilon_t.$$

   (a) Estimated model is:

$$Y_t = \tau + \rho Y_{t-1} + \varepsilon_t.$$

   The critical values that result from the t-statistic that tests $\rho = 1$ equal: -1.645 (95%) and -2.325 (99%).

   (b) Estimated model is:

$$Y_t = \tau + \delta t + \rho Y_{t-1} + \varepsilon_t.$$

   The critical values that result from the t-statistic that tests $\rho = 1$ equal: -3.41 (95%) and -3.96 (99%).

For practical purposes, only 1b and 2b are relevant. 1b is used when the data do not contain a trend while 2b is used when the data contains a trend.

6.6 Augmented Dickey-Fuller test

In the AR(1) models that we use to test for a unit root value, for instance:

$$Y_t = \tau + \delta t + \rho Y_{t-1} + \varepsilon_t,$$

it is convenient to re-specify them towards a so-called error correction specification:

$$\Delta Y_t = \tau + \delta t + \phi Y_{t-1} + \varepsilon_t,$$

with $\Delta Y_t = Y_t - Y_{t-1}$ and $\phi = \rho - 1$. Testing for a unit root value of ρ is then identical to testing for a zero value of ϕ. The standard t-statistic can then be used for this purpose: this is called the Dickey-Fuller test.

The re-specification of the AR model towards the error correction specification is especially convenient for higher order AR models:

$$\begin{aligned} Y_t &= \tau + \delta t + \rho_1 Y_{t-1} + \ldots + \rho_p Y_{t-p} + \varepsilon_t \Leftrightarrow \\ \Delta Y_t &= \tau + \delta t + \phi Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \ldots + \alpha_{p-1} \Delta Y_{t-p+1} + \varepsilon_t, \end{aligned} \tag{6.1}$$

with

$$\begin{aligned} \phi &= \rho_1 + \ldots + \rho_p - 1 = -\rho(1), \\ \alpha_i &= -(\rho_{i+1} + \ldots + \rho_p),\ i = 1, \ldots, p-1. \end{aligned}$$

A unit root value then implies a zero value of $\rho(1)$ which is identical to ϕ. Least squares regression can therefore directly be applied to obtain the unit root parameter ϕ. The standard t-statistic for testing for a zero value of ϕ can thus be used to test for a unit root in higher order AR models. These tests are referred to as **augmented Dickey-Fuller** (**ADF**) statistics and their limiting distribution is identical to the limiting distribution of the t-statistic in the AR(1) model. Hence, we can use the same critical values as for the AR(1) model.

In the expression (6.1) with $\rho\,(1) = 0$, we regress ΔY_t on Y_{t-1}, i.e. an $\mathsf{I}(0)$ variable on an $\mathsf{I}(1)$: this is an *unbalanced* regression and the estimate should be close to zero since we try to explain a stationary variable using a non-stationary one.

The ADF is a **parametric** test for the presence of a unit root since it relies on estimating the parameters of model for Y_t which is fitted to the data by specifying the lag order p and/or the presence of nonzero drift/linear trend. It suffers from various drawbacks:

1. expression (6.1) is assumed to hold: the ε_t Normal and iid. This necessitates that p is not underestimated.

2. the deterministic parameters do not play the same role depending on whether $\phi = 0$ or not.

The solution to (1) is to start from a large enough p so ε_t passes all the diagnostic tests and then reduce p in the approach of Box-Jenkins. Many econometric packages choose p using the Schwarz criterion.

The solution to the issue of constants and trends is either to follow a precise sequential testing algorithm (starting from the more general and verifying in turn each hypothesis precisely) or to

Table 6.1: Distribution of the F statistic for the test $(\tau, \delta, \phi) = (\tau, 0, 0)$ in $\Delta Y_t = \tau + \delta t + \phi Y_{t-1} + \varepsilon_t$.

	Probability of a value less than							
Sample Size	.01	.025	.05	.10	.90	.95	.975	.99
25	.74	.90	1.08	1.33	5.91	7.24	8.65	10.61
20	.76	.93	1.11	1.37	5.61	6.73	7.81	9.31
100	.76	.94	1.12	1.38	5.47	6.49	7.44	8.73
250	.76	.94	1.13	1.39	5.39	6.34	7.25	8.43
500	.76	.94	1.13	1.39	5.36	6.30	7.20	8.34
∞	.77	.94	1.13	1.39	5.34	6.25	7.16	8.27

Source: Dickey & Fuller (1976), Table VI

resort to alternative tests. The latter solution is generally preferred since the ADF does not perform necessarily well in finite samples. In the model above, depending on the values of (τ, δ, ρ) the following behaviors for Y_t result:

(τ, δ, ρ)	$\lvert\rho\rvert < 1$	$\lvert\rho\rvert = 1$
$\delta \neq 0$	stationary around a linear trend	integrated and exhibits a quadratic trend
$\tau \neq 0, \delta = 0$	stationary with a nonzero mean	integrated and exhibits a linear trend
$\tau = 0, \delta = 0$	stationary with zero mean	integrated without deterministic trend

It is often recommended to start with the more general model with nonzero (τ, δ). Test $\phi = 1$, then test for $\delta = 0$ using the F test for the joint hypothesis $(\delta, \phi) = (0, 0)$. This means computing the Fisher statistic

$$F = \frac{ESS_R - ESS_{UR}}{(N - k)\, q},$$

where N is the number of observations used in the regression ($T - 1$ since we use ΔY_t, not Y_t), k the number of estimated parameters in the unrestricted regression (i.e. estimating δ and ρ), q the number of restrictions (two here), ESS_R is the sum of squares of the modeled variables (i.e. $\sum_{i=1}^{T} \Delta \widehat{Y}_i^2$) under the null hypothesis $(\delta, \phi) = (0, 0)$ (i.e. excluding the corresponding regressors from the regression) and ESS_{UR} the sum $\sum_{i=1}^{T} \Delta \widehat{Y}_i^2$ in the unrestricted regression. The critical values are given table 6.1.

If the hypothesis is accepted then impose it and retest for $\phi = 0$ using the appropriate Dickey-Fuller distribution, and so on. Hopefully no contradiction appears.

6.7 Alternative unit root tests

There exist many tests for the presence of a unit root. We present below the most commonly present in econometric packages.

6.7.1 Phillips-Perron (1988) test (PP)

Peter Phillips and Pierre Perron suggested a *nonparametric* correction to the DF test that replaces the ADF. Here the autocorrelation of ΔY_t is not modeled. Instead a simple DF model ($p = 1$) is fitted to the data. The residuals are assumed to be stationary (this is true if Y_t is either stationary or integrated of order 1) so they admit a Wold representation. The distribution of the t statistic for ϕ depends on the autocorrelation structure of the residuals. Phillips-Perron proposed a correction for t using the difference between the estimated sum of all covariances of the residuals (estimated *spectral density at frequency zero, or long-run variance*) and their estimated variance. This difference is zero if the residuals are iid.

The PP still needs to specify the deterministic terms in the regression. This test is robust to misspecification of the correlation structure of ΔY_t. As with nonparametric corrections, it is less precise when the ADF equation correctly represents the reality. Both PP and ADF are asymptotically equivalent.

6.7.2 Schmidt-Phillips (1992) test (SP)

This test aims to solve the issue of whether to include a deterministic trend in the DF equation or not. It consists in arbitrarily detrending Y_t in $Y_t - \widehat{a} - \widehat{b}t$ using least-squares (OLS for Schmidt-Phillips but GLS is also feasible).

6.7.3 Eliott-Rothenberg-Stock (1996) test (ERS)

This test used the fact that in finite samples, values of ρ that are close to, yet strictly different from, unity are indistinguishable. Instead of differencing Y_t, ERS suggest quasi-differencing Y_t as $\widetilde{Y}_t = Y_t - \rho_T Y_{t-1}$, with $\rho_T = 1 - \frac{c}{T}$, and $c = 7$ or 13.5 respectively if only a constant, or a constant and a linear trend are estimated. The ERS statistic is

$$P_T = \frac{SSR(\rho_T) - \rho_T SSR(1)}{f_0}$$

where $SSR(x)$ is the sum of the squared residuals from the Dickey-Fuller equation using quasi-difference parameter x. f_0 is as usual computed under the null. This test is optimal against the point alternative ρ_T and is quite robust in general. This test can be combined with GLS detrending.

6.7.4 Kwiatkowski, Phillips, Schmidt, and Shin (1992) test (KPSS)

This test is based on a different null hypothesis: that of stationarity (the others test for the presence of a unit root). Here the equation that is involved is

$$Y_t = X_t\beta + u_t$$

where X_t is an *exogenous* variable (a constant and a deterministic trend here, but it could be any other variable, you will see this in the context of cointegration testing). The LM statistic is then defined as

$$LM = \frac{\sum_t S_t^2}{T^2 f_0}$$

where $S_t = \sum_{i=1}^{t} \widehat{u}_t$ and f_0 is an estimate of the spectral density of u_t at frequency zero. Under the null of stationarity, KPSS tabulated the distribution of the LM test. Unfortunately this test is not very robust and tends to under reject stationarity.

6.8 Summary and application

Summarizing, two important phenomena occur when the AR(1) parameter becomes equal to the random walk value:

1. Interpretation of constants and trends changes
2. Critical values in t-test are different from the standard ones and depend on whether no constant (-1.96), a constant (-2.89) or a constant and trend (-3.46) are present in the AR(1) model.

Both phenomena are important for practitioners. The first one is important as it shows that we can mistakenly end up with random walk values of the AR parameters when we do not include enough deterministic components.

Example 13 *Consider the series Y_t which is generated by the model*

$$Y_t = 0.1t + 0.9Y_{t-1} + \varepsilon_t,$$

*where $\varepsilon_t \sim N(0, \sigma^2)$. A graph of this series looks like figure 6.5 We test for a random walk by estimating the (**wrong!**) model:*

$$\Delta Y_t = \tau + \phi Y_{t-1} + \varepsilon_t,$$

and obtain the value of the DF test statistic 1.169 with critical values:

$$\begin{aligned} 1\% \textit{ Critical Value:} \quad & -3.497 \\ 5\% \textit{ Critical Value:} \quad & -2.891 \\ 10\% \textit{ Critical Value:} \quad & -2.583 \end{aligned}$$

As we estimated the wrong model, the results are misleading. A statistical model always tries to explain the main features of the data. In this case the main feature of the data is the trending behavior. The only parameter configuration for which the model

$$Y_t = \tau + \rho Y_{t-1} + \varepsilon_t,$$

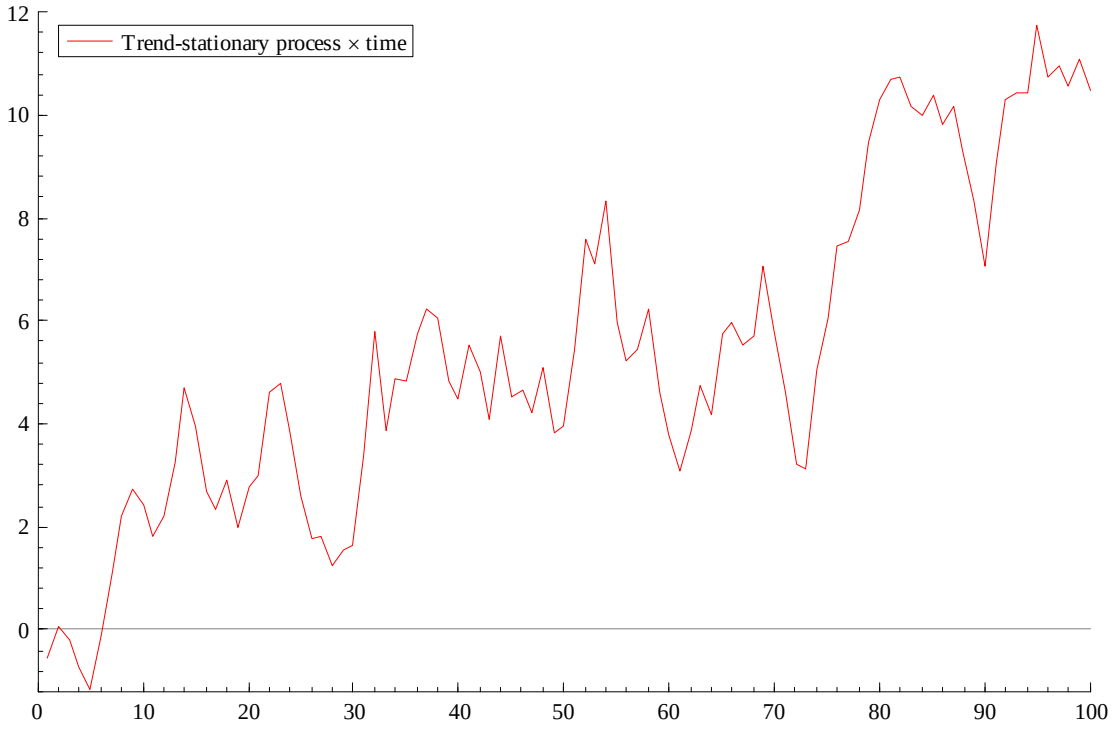

Figure 6.5: Example of a time series generated by $Y_t = 0.1t + 0.9Y_{t-1} + \varepsilon_t$, with ε_t a GWN.

contains a trend is $\rho = 1$. *As a consequence, since the generated series has a trend, we find a random walk model when we only incorporate a constant term. When we estimate the correct model,*

$$\Delta Y_t = \tau + \delta t + \phi Y_{t-1} + \varepsilon_t,$$

we obtain the E-Views output, using the command y.uroot(t) *on figure 6.6. This correctly estimated model leads to a rejection of the random walk hypothesis. This shows that the specification of the constants and trends is crucial in random walk testing. A rule of thumb is that we should always include a trend when the series is upward or downward sloping.*

In practice, we only use the AR models:

$$\Delta Y_t = \tau + \phi Y_{t-1} + \varepsilon_t$$

in case that the series is not trending (interest and exchange rates) for which we reject the hypothesis of a random walk with 95% significance when the t-value of ϕ is less then -2.89, and

$$2.\ \Delta Y_t = \tau + \delta t + \phi Y_{t-1} + \varepsilon_t,$$

when the series is trending (stock prices, gnp), for which we reject the hypothesis of a random walk with 95% significance when the t-value of ϕ is less then -3.46.

In the case of series with constant growth rates, we typically work with the log instead of the level of the data.

$$\begin{aligned} Y_t &= (1+g)Y_{t-1} \\ &= (1+g)^t Y_0 \end{aligned}$$

such that the logarithm is represented by:

$$\log(Y_t) = t\log(1+g) + \log(Y_0)$$

which is linear in t. Since we are estimating linear models, we want to express the dependent variable as a linear function of the explanatory variables. This explains why analyze the logarithm instead of the level of the series in case of dependent variables with a "fixed" growth rate. Logs are often written using lower case letters

$$y_t = \log(Y_t)$$

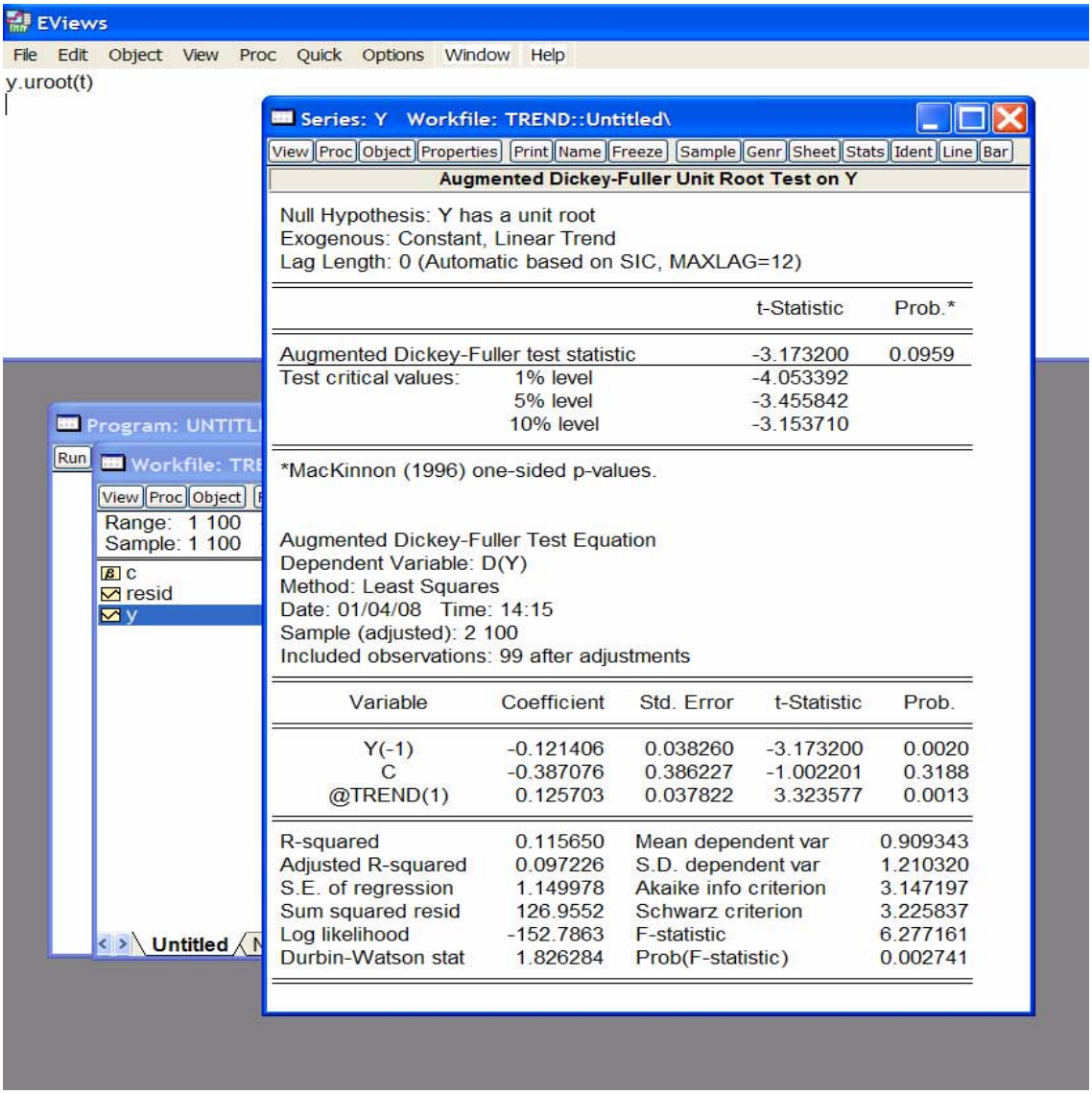

Figure 6.6: E-Views output for an ADF unit-root test with a trend for a sample a 100 observations generated by $Y_t = 0.1t + 0.9Y_{t-1} + \varepsilon_t$, with ε_t a GWN.

Chapter 7

Multivariate Time Series Analysis

In many situations, analyzing a time series in isolation is reasonable. However, in some cases, univariate analysis can be limiting. For example, Campbell (1996) links financially interesting variables, including stock returns and the default premium, together in a multivariate system allowing shocks to one variable to propagate to the others. A model which allow dependence between state variables seems reasonable; investors constantly observe shocks in one asset which result in changed assessments of others.

The vector autoregression is the mechanism that is used to link multiple stationary time-series variables together. When variables contain unit roots, a different type of analysis, cointegration, is needed. This chapter covers these two topics relying on many results from the analysis of univariate time series.

7.1 Vector Autoregressions

Vector autoregressions are remarkably similar to univariate autoregressions; so similar that the intuition behind most results carries over by simply replacing scalars with a matrices and scalar operations with matrix operations. The important new concepts of VAR analysis are Granger causality and impulse response functions.

7.1.1 Definition

A pth order vector autoregression (VAR(p)) is defined as process with dynamics governed by

$$\mathbf{y}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + ... + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \epsilon_t$$

where $\mathbf{y}_t$ is a K by 1 vector stochastic process, $\mathbf{\Phi}_0$ is a K by 1 vector of intercepts, $\mathbf{\Phi}_j, j = 1, ..., p$ are K by K matrices and ϵ_t is a vector white noise process, satisfying

$$\begin{aligned} \mathsf{E}\left[\epsilon_t\right] &= \mathbf{0} \\ \mathsf{E}\left[\epsilon_t \epsilon_{t-s}'\right] &= \mathbf{0} \\ \mathsf{E}\left[\epsilon_t \epsilon_t'\right] &= \mathbf{\Sigma} \end{aligned}$$

where $\mathbf{\Sigma}$is a positive definite finite matrix. Simply replacing the vectors and matrices with scalars will produce the definition of an AR(p).

7.1.2 Properties of a VAR(1)

The properties of the VAR(1) are fairly simple to study. More importantly, the next section shows that all VAR(p)can be rewritten as a VAR(1). In other words, more general cases require no additional effort.

Stationarity

A VAR(1), given by,

$$\mathbf{y}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \epsilon_t$$

is covariance stationary if the eigenvalues of $\mathbf{\Phi}_1$ are less than 1 in modulus.[1] In the univariate case, this is equivalent to the condition $|\phi_1| < 1$. Backward substitution can be used on the VAR(1) to show that

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \mathbf{\Phi}_0 + \sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \epsilon_{t-i}$$

which is equivalent to

$$\mathbf{y}_t = \left(\mathbf{I}_K - \mathbf{\Phi}_1\right)^{-1} \mathbf{\Phi}_0 + \sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \epsilon_{t-i}$$

The eigenvalue condition ensures that $\mathbf{\Phi}_1^i$ will converge to zero as i grows large.

[1] The definition of an eigenvalue is:

Definition (**Eigenvalue**) λ is an eigenvalue of a square matrix $\mathbf{A}$ if and only if $|\mathbf{A} - \lambda \mathbf{I}_n| = 0$.

The crucial properties of eigenvalues for applications to VARs are given in the following theorem:

Theorem (**Matrix Power**) Let $\mathbf{A}$ be an n by n matrix. Then the following statement are equivalent:

- $\mathbf{A}^m \to 0$ as $m \to \infty$
- All eigenvalues of $\mathbf{A}$, $\lambda_i, i = 1, 2, ..., n$, are less than 1 in modulus ($|\lambda_i| < 1$).
- The series $\sum_{i=0}^{m} \mathbf{A}^i \to \left(\mathbf{I}_n - \mathbf{A}\right)^{-1}$ as $m \to \infty$

Mean

Taking expectations of $\mathbf{y}_t$ using the backward substitution form yields

$$\mathsf{E}[\mathbf{y}_t] = (\mathbf{I}_K - \mathbf{\Phi}_1)^{-1} \mathbf{\Phi}_0$$

Again, this result is fundamentally similar to that of a univariate AR(1) which has a mean of $(1-\phi_1)^{-1}\phi_0$. The eigenvalues play an important role in determining the mean. If an eigenvalue of $\mathbf{\Phi}_1$ is close to one, $(\mathbf{I}_K - \mathbf{\Phi}_1)^{-1}$ will contain large values and the unconditional mean will be large. Similarly, if $\mathbf{\Phi}_1 = \mathbf{0}$, then the mean is simply $\mathbf{\Phi}_0$ and the process is simply white noise plus a constant.

Variance

Again, using the backward substitution for of the VAR(1), the long run variance can be shown to be

$$\begin{aligned}
\mathsf{E}\left[(\mathbf{y}_t - \mu)(\mathbf{y}_t - \mu)'\right] &= \mathsf{E}\left[\left(\sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \epsilon_{t-i}\right)'\right] \\
&= \mathsf{E}\left[\left(\sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty} \epsilon_{t-i}' \mathbf{\Phi}_1^{i\prime}\right)\right] \\
&= \sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \mathsf{E}\left[\epsilon_{t-i}\epsilon_{t-i}'\right] \mathbf{\Phi}_1^{i\prime} \text{ (covariances are zero)} \\
&= \sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \mathbf{\Sigma} \mathbf{\Phi}_1^{i\prime}
\end{aligned}$$

Using the operator vec which transforms a matrix into a vector by stacking the columns on top of one another

$$vec\left(\mathsf{E}\left[(\mathbf{y}_t - \mu)(\mathbf{y}_t - \mu)'\right]\right) = \sum_{i=0}^{\infty} vec\left(\mathbf{\Phi}_1^i \mathbf{\Sigma} \mathbf{\Phi}_1^{i\prime}\right)$$

and $vec(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\, vec(\mathbf{B})$ hence

$$\begin{aligned}
vec\left(\mathsf{E}\left[(\mathbf{y}_t - \mu)(\mathbf{y}_t - \mu)'\right]\right) &= \sum_{i=0}^{\infty} \left(\mathbf{\Phi}_1^i \otimes \mathbf{\Phi}_1^i\right) vec(\mathbf{\Sigma}) \\
&= \sum_{i=0}^{\infty} \left(\mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1\right)^i vec(\mathbf{\Sigma}) \\
&= \left(\mathbf{I}_{K^2} - \mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1\right)^{-1} vec(\mathbf{\Sigma})
\end{aligned}$$

where $\mu = (\mathbf{I}_K - \mathbf{\Phi}_1)^{-1}\mathbf{\Phi}_0$. Compared to the long run variance of a univariate autoregression, $\sigma^2/\left(1-\phi_1^2\right)$, the similarities are less obvious. The differences arise from the noncommutative nature of matrices ($\mathbf{AB} \neq \mathbf{BA}$ in general). The final lines makes use of the vec operator to

"simplify" the expression. The vec operator and a Kronecker product stacks the elements of a matrix in the a single column (don't worry if you aren't familiar with these mathematical tools, you don't have to be). Once again the eigenvalues of $\mathbf{\Phi}_1$ play an important role. If any are close to 1, the variance will be large since the eigenvalues fundamentally determine the persistence of shocks: as was the case in scalar autoregressions, higher persistence leads to higher variance.

Autocovariance

The autocovariances of a stationary vector valued stochastic process are defined as

$$\mathbf{\Gamma}_s = \mathsf{E}\left[(\mathbf{y}_t - \mu)(\mathbf{y}_{t-s} - \mu)'\right]$$

and

$$\mathbf{\Gamma}_{-s} = \mathsf{E}\left[(\mathbf{y}_t - \mu)(\mathbf{y}_{t+s} - \mu)'\right]$$

and they present the first significant deviation from univariate time series analysis. Instead of being symmetric around t, they are symmetric in their transpose. Specifically,

$$\mathbf{\Gamma}_s = \mathbf{\Gamma}'_{-s}.$$

In contrast, the autocovariances of a stationary scalar processes satisfy $\gamma_s = \gamma_{-s}$. Computing the autocovariances is also easily accomplished using the backward substitution form,

$$\begin{aligned}\mathsf{E}\left[\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\epsilon_{t-s-i}\right)'\right] &= \mathsf{E}\left[\left(\sum_{i=0}^{s-1}\mathbf{\Phi}_1^i\epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\epsilon_{t-s-i}\right)'\right] \\ &\quad + \mathbf{\Phi}_1^s\mathsf{E}\left[\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\epsilon_{t-s-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\epsilon_{t-s-i}\right)'\right] \\ &= \mathbf{\Phi}_1^s\mathsf{V}\left[\mathbf{y}_t\right]\end{aligned}$$

where $\mathsf{V}\left[\mathbf{y}_t\right]$ is the covariance of the VAR(1). Like most properties of a VAR, this result is fundamentally similar to the autocovariance function of an AR(1): $\gamma_s = \phi_1\mathsf{V}\left[y_t\right]$. It is also worth noting that

$$\mathbf{\Gamma}_{-s} = \mathsf{V}\left[\mathbf{y}_t\right]\mathbf{\Phi}_1^{s\prime}.$$

VAR(p) is really a VAR(1)

Once the properties of a VAR(1) have been studied, one surprising and useful result is that any VAR(p) can be rewritten as a VAR(1). Suppose $\{\mathbf{y}_t\}$ follows a VAR(p) process,

$$\mathbf{y}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + ... + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \epsilon_t$$

By subtracting the mean and stacking p of $\mathbf{y}_t$ into a large column vector denoted $\mathbf{z}_t$, this VAR(p) can be transformed into an VAR(1) by

$$\mathbf{z}_t = \boldsymbol{\Upsilon}\mathbf{z}_{t-1} + \xi_t$$

where

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{y}_t - \mu \\ \mathbf{y}_{t-1} - \mu \\ \vdots \\ \mathbf{y}_{t-p+1} - \mu \end{bmatrix}, \boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \boldsymbol{\Phi}_3 & \cdots & \boldsymbol{\Phi}_p \\ \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & \mathbf{I}_K & \mathbf{0} \end{bmatrix}, \xi_t = \begin{bmatrix} \epsilon_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

$$\mu = \left(\mathbf{I} - \sum_{k=1}^{p} \boldsymbol{\Phi}_k\right)^{-1} \boldsymbol{\Phi}_0$$

The previous properties can be directly applied noting that

$$\mathsf{E}\left[\xi_t \xi_t\right] = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

This is known as the **companion form** and allows the statistical properties of any VAR(p) process to be directly computed from previous results. From these results, it can be determined that in a stationary VAR(p),

$$\mathsf{E}\left[\mathbf{y}_t\right] = \left(\mathbf{I} - \sum_{k=1}^{p} \boldsymbol{\Phi}_k\right)^{-1} \boldsymbol{\Phi}_0$$

and that the pth order VAR is stationary if all of the eigenvalues of $\sum_{k=1}^{p} \boldsymbol{\Phi}_k$ are less than one.

Example: The interaction of stock and bond returns

Stocks and long term bonds are often thought to hedge one another. VARs provide a simple method to determine whether their returns are linked through time. Consider the VAR(1)

$$\begin{bmatrix} VWM_t \\ 10YR_t \end{bmatrix} = \begin{bmatrix} \phi_{01} \\ \phi_{02} \end{bmatrix} + \begin{bmatrix} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{bmatrix} \begin{bmatrix} VWM_{t-1} \\ 10YR_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

which implies models for stock returns

$$VWM_t = \phi_{01} + \phi_{11,1} VWM_{t-1} + \phi_{12,1} 10YR_{t-1} + \epsilon_{1t}$$

and for long bond returns

$$10YR_t = \phi_{02} + \phi_{21,1} VWM_{t-1} + \phi_{22,1} 10YR_{t-1} + \epsilon_{2t}$$

Since these models do not share any parameters, they can be separately estimated using OLS (or in Eviews using Quick/Estimate VAR). Using data on the VWM from CRSP and the 10 constant maturity treasury yield from FRED[2] from May 1953 until December 2004, a VAR(1) was estimated.

$$\begin{bmatrix} VWM_t \\ 10YR_t \end{bmatrix} = \begin{bmatrix} \underset{(0.00)}{0.996} \\ \underset{(0.68)}{0.046} \end{bmatrix} + \begin{bmatrix} \underset{(0.76)}{0.012} & \underset{(0.00)}{0.239} \\ \underset{(0.03)}{-0.058} & \underset{(0.00)}{0.334} \end{bmatrix} \begin{bmatrix} VWM_{t-1} \\ 10YR_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

where the $p$value is in parenthesis below each coefficient. A few things are worth noting. Stock returns are not predictable with their own lags but do appear to be predictable using lagged bond returns: positive bond returns lead to positive future returns in stocks. In contrast, positive returns in equities result in negative returns for future bond holdings. The long run mean can be computed as

$$\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.012 & 0.239 \\ -0.058 & 0.334 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0.996 \\ 0.046 \end{bmatrix} = \begin{bmatrix} 1.004 \\ -0.018 \end{bmatrix}$$

These values are similar to the sample means of 1.003 and -0.022.

Example: Campbell's VAR

Campbell (1996) builds a theoretical model for asset prices where economically meaningful variables evolve according to an VAR, including stock returns, real labor income growth, the term premium, the relative t-bill rate and the dividend yield. The VWM series from CRSP is used for equity returns and real labor income is the log change in income from labor minus the log change in core inflation where both series are from FRED. The term premium is the difference between a 10 year constant maturity yield and a 3-month t-bill rate. Both series are from FRED. The relative t-bill rate is the current yield on a 1-month t-bill minus the average yield over the past 12 months and the data is available on Ken French's web site. The dividend yield was computed as the difference in the VWM with and without dividends; both series are available from CRSP.

Using a VAR(1) specification, the model can be described

$$\begin{bmatrix} VWM_t \\ LBR_t \\ RTB_t \\ TERM_t \\ DIV_t \end{bmatrix} = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \begin{bmatrix} VWM_{t-1} \\ LBR_{t-1} \\ RTB_{t-1} \\ TERM_{t-1} \\ DIV_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

Parameter estimated are in table 7.1. Two sets of parameters are presented. The top panel contains estimates using non-scaled data. This produces some very large (in magnitude, not statistical significance) estimates which are the result of two variables having different scales. The

[2]The yield is first converted to prices and then returns are computed as log differences in the prices.

Raw Data

	VWM_{t-1}	LBR_{t-1}	RTB_{t-1}	$TERM_{t-1}$	DIV_{t-1}
VWM_t	0.045	91.040	-0.265	0.444	-26.881
	(0.29)	(0.01)	(0.90)	(0.03)	(0.80)
LBR_t	0.000	-0.134	-0.002	0.000	-0.175
	(0.18)	(0.00)	(0.56)	(0.34)	(0.21)
RTB_t	-0.001	0.668	0.628	-0.020	1.936
	(0.09)	(0.17)	(0.00)	(0.00)	(0.22)
$TERM_t$	-0.010	-6.972	0.176	0.983	13.639
	(0.00)	(0.00)	(0.21)	(0.00)	(0.05)
DIV_t	0.000	-0.011	0.000	-0.000	-0.130
	(0.52)	(0.43)	(0.96)	(0.01)	(0.00)

Standardized Series

	VWM_{t-1}	LBR_{t-1}	RTB_{t-1}	$TERM_{t-1}$	DIV_{t-1}
VWM_t	0.045	0.117	-0.006	0.111	-0.011
	(0.29)	(0.01)	(0.90)	(0.03)	(0.80)
LBR_t	0.057	-0.134	-0.029	0.048	-0.054
	(0.18)	(0.00)	(0.56)	(0.34)	(0.21)
RTB_t	-0.047	0.038	0.628	-0.223	0.034
	(0.09)	(0.17)	(0.00)	(0.00)	(0.22)
$TERM_t$	-0.040	-0.036	0.016	0.983	0.022
	(0.00)	(0.00)	(0.21)	(0.00)	(0.05)
DIV_t	0.028	-0.034	0.003	-0.137	-0.130
	(0.52)	(0.43)	(0.96)	(0.01)	(0.00)

Table 7.1: Parameter estimates from Campbell's VAR. The top panel contains estimates using unscaled data while the bottom panel contains estimates from data which have been standardized to have unit variance. While the magnitudes of many coefficients change, the $p$values and the eigenvalues of these two parameter matrices are identical. The standardized series have one slight advantage in that the parameters are roughly comparable since the series have approximately the same variance.

bottom panel contains estimated from data which have been transformed by dividing each series through by its standard deviation. This puts the coefficient on a roughly level playing field. You should also notice that the pvalues are unchanged by the choice. This shouldn't be surprising: OLS is closed to scalings of this type. One less obvious feature of the two sets of estimates is that the eigenvalues of the two parameter matrices are identical; both estimates indicate have the same persistence.

VAR forecasting

Once again, VAR(p) behavior is essentially identical to that of an AR(p). Recall that the h-step ahead forecast form an AR(1) is given by

$$\mathsf{E}\left[y_{t+h}\right] = \left(\sum_{k=0}^{h-1} \phi^k\right) \phi_0 + \phi_1^h y_t$$

The h-step ahead forecast of an VAR(1) is essentially identical,

$$\mathsf{E}\left[\mathbf{y}_{t+h}\right] = \left(\sum_{k=0}^{h-1} \mathbf{\Phi}^k\right) \mathbf{\Phi}_0 + \mathbf{\Phi}_1^h \mathbf{y}_t$$

and forecasts from higher order can be constructed by simple forward recursions beginning at $h = 1$.

Example: The interaction of stock and bond returns

One important feature of VAR occurs when two series are related in time. Univariate forecasts cannot adequately capture the feed back between two series and are misspecified if the data follow a VAR. To illustrate the differences, recursively estimated 1-step ahead forecasts were produced from both the stock-bond VAR(1)

$$\begin{bmatrix} VWM_t \\ 10YR_t \end{bmatrix} = \begin{bmatrix} 0.996 \\ 0.046 \end{bmatrix} + \begin{bmatrix} 0.012 & 0.239 \\ -0.058 & 0.334 \end{bmatrix} \begin{bmatrix} VWM_{t-1} \\ 10YR_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

and simple AR(1)'s for each series. The data set contains a total of 620 observations. Beginning at observation 381 and continuing until observation 620, the models (the VAR and the two ARs) were estimated using an expending window of data and 1-step ahead forecasts were computed. Figure 7.1 contains a graphical representation of the differences between the AR(1)s and the VAR(1). The forecasts for the market are substantially different while the forecasts for the 10-year bond return are not. The changes (or lack thereof) are simply a function of the model specification: the return on the 10-year bond has predictive power for both so the VAR(1) is a much better model for stock returns than an AR(1) yet it not much better for bond returns.

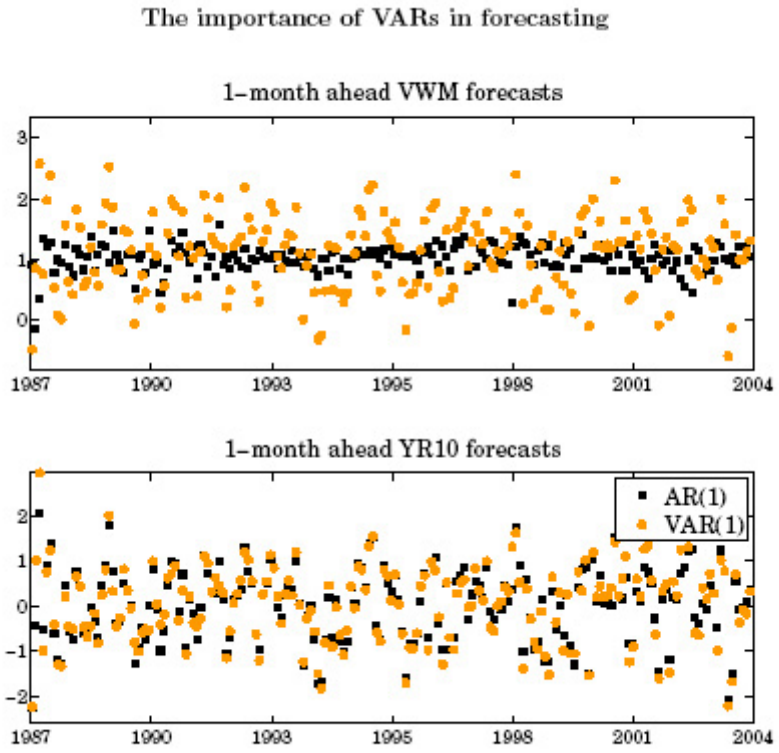

Figure 7.1: The figure contains 1-step ahead forecasts from both a VAR(1) and an AR(1) for both the value-weighted market returns and the return on a 10 year bond. These two pictures indicate that the return on the long bond has substantial predictive power for equity returns and the opposite is not true.

Estimation and Identification

Estimation and identification is the first significant break from directly applying the lessons learned about univariate models to multivariate models. In addition to ACFs and PACFs, vector stochastic processes also have cross-correlation functions (CCFs) and partial cross-correlation functions (PCCFs). The crosscorrelations between two series x and y are defined as

$$\rho_{xy,s} = \frac{\mathsf{E}\left[\left(x_t - \mu_x\right)\left(y_{t-s} - \mu_y\right)\right]}{\sqrt{\mathsf{V}\left[x_t\right]\mathsf{V}\left[y_t\right]}}$$

It should be obvious that, unlike autocorrelations, cross-correlation are not symmetric so the order xy or yx matters. Partial cross-correlation are defined in a similar manner; the correlation between x_t and y_{t-s} controlling for $y_{t-1}, ..., y_{t-s+1}$. To get a feel for the value of these two functions, figure 7.2 contains the ACF and CCF of two VAR(1)with identical persistence. The top panel contains the functions for

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} .5 & .4 \\ .4 & .5 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

while the bottom contains the functions for a trivial VAR

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} .9 & 0 \\ 0 & .9 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

which is actually two AR(1)in disguise. The nontrivial VAR(1) demonstrates dependence with respect to both series while the AR-in-disguise shows no dependence between y_t and x_{t-j} , $j > 0$.

With the new tools, it would seem that Box-Jenkins could be directly applied to vector processes. However, while this is true, it is extraordinarily difficult to examine the ACF, PACF, CCF and PCCF and determine what type of model is needed or the appropriate coefficients. There are just too much interaction and many possible models for a set of ACFs and CCFs.

The solution is to take a hands off approach as advocated by Sims. The initial VAR specification should include all variables which theory indicate are relevant to the problem at hand and a lag length should be chosen which has a high likelihood of capturing all of the dynamics. Once this value has been set, either a general-to-specific search can be conducted over the lag length or an information criteria should be used. In the VAR case, the AIC and SIC are given by

$$\begin{aligned} \mathsf{AIC} &: \ln\left|\mathbf{\Sigma}\left(p\right)\right| + \frac{2K^2 p}{T} \\ \mathsf{SC} &: \ln\left|\mathbf{\Sigma}\left(p\right)\right| + \frac{K^2 p \ln T}{T} \end{aligned}$$

where $\mathbf{\Sigma}\left(p\right)$ is the covariance of the residuals using p lags and $|\cdot||$ indicates the determinant. The lag length should be chosen to minimize one of these criteria, and the SIC will always choose a (weakly) smaller model than the AIC.

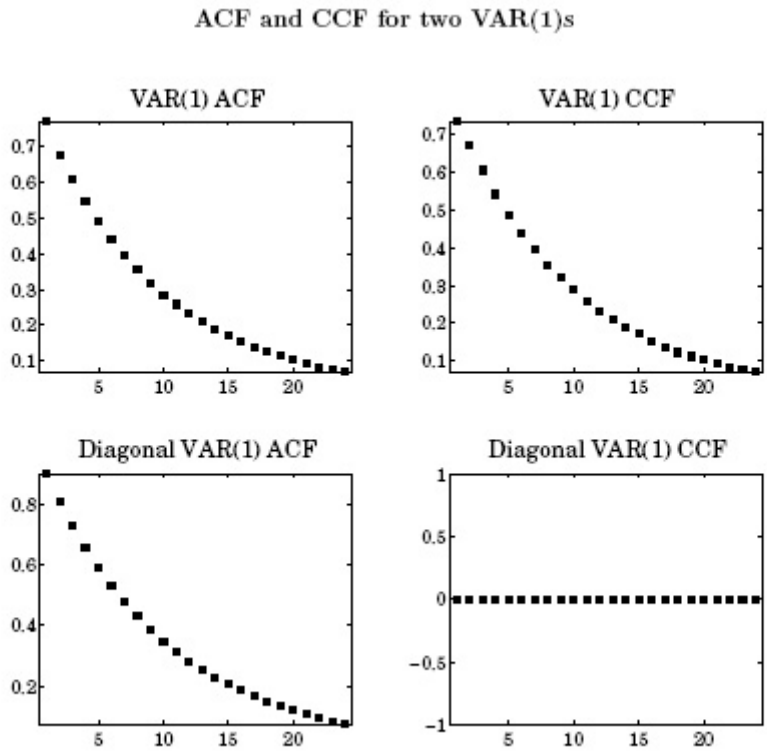

Figure 7.2: The top panel contains the ACF and CCF for a nontrivial VAR process where contemporaneous values depend on both series. The bottom contains the ACF and CCF for a trivial VAR which is simply composed to two AR(1)s.

Lags	AIC	SIC	LR	P-val
0	6.78	5.91	14924	0.00
1	3.42	2.74	161.9	0.00
2	3.18	2.70	1428	0.00
3	0.28	0.00	35.75	0.07
4	0.29	0.20	24.92	0.46
5	0.32	0.43	120.9	0.00
6	0.11	0.41	23.92	0.52
7	0.14	0.64	30.21	0.21
8	0.15	0.84	22.16	0.62
9	0.17	1.06	26.47	0.38
10	0.17	1.25	23.39	0.55
11	0.18	1.45	68.83	0.00
12	0.00	1.47	N/A	N/A

Table 7.2: Normalized values for the AIC and SIC for Campbell's VAR. The AIC chooses 12 lags while the SIC chooses only 3. A general-to-specific search would stop at 12 lags since the likelihood ratio test of 12 lags against 11 rejects with a p-value of 0. If the initial number of lags was less than 12, the GeTS procedure would choose 6 lags. Note that the LR and $p$value corresponding to lag l is a test of the null of l lags against an alternative of $l+1$ lags.

To use a general-to-specific approach, a simple likelihood ratio test can be computed

$$\left(T - p_2K^2\right)\left(\ln\left|\mathbf{\Sigma}\left(p_1\right)\right| - \ln\left|\mathbf{\Sigma}\left(p_2\right)\right|\right) \rightarrow \chi^2_{(p_2-p_1)K^2}$$

where p_1 is the number of lags in the restricted (smaller) model, p_2 is the number of lags in the unrestricted (larger) model and K is the dimension of $\mathbf{y}_t$. Since model 1 is a restricted version of model 2, its variance is larger which ensures this statistic is positive (a good thing since it has a χ^2 distribution).

Example: Campbell's VAR

A lag length selection procedure was conducted using Campbell's VAR. The results are contained in table 7.2. This table contains both the AIC and SIC values for lags 0 through 12 as well as likelihood ratio test results for testing l lags against $l+1$. Note that the LR and $p$value corresponding to lag l is a test of the null of l lags against an alternative of $l+1$ lags. Using the AIC, 12 lags would be selected since it produces the smallest value. If the initial lag length was less than 12, 6 lags would be selected. The SIC chooses 3 lags in an unambiguous manner. A general-to-specific procedure would choose 12 lags while a specific-to-general procedure would choose 4. The test statistic for a null $\mathsf{H}_0 : p = 11$ against an alternative $\mathsf{H}_1 : p = 12$ has a value of 68 and a $p$value of 0. One final specification search was conducted. Rather than begin at the largest lag and work

down one by one, a large specification search which evaluates models with every combination of lags up to 12 was computed. This required fitting 4096 regressions which fortunately only requires 9 seconds. For each possible combination of lags, the AIC and the BIC were computed. Using this methodology, the AIC search selected lags 1-4, 6, 10 and 12 while the BIC selected a smaller model with only lags 1, 3 and 12. Search procedures of this type are computationally viable for checking up to about 20 lags.

Granger Causality

Granger causality is the first concept exclusive to vector analysis. GC is the standard method to determine whether one variable is useful in predicting another and it is a good indicator of whether a VAR is needed.

Definition 13 *Granger causality is generally defined in the negative. A scalar random variable* $\{x_t\}$ *is said not to Granger cause* $\{y_t\}$ *if*[3]

$$\mathsf{E}\left[y_t|x_{t-1}, y_{t-1}, x_{t-2}, y_{t-2}, ...\right] = \mathsf{E}\left[y_t|y_{t-1}, y_{t-2}, ...\right]$$

That is, $\{x_t\}$ *does not Granger cause* $\{y_t\}$ *if the forecast of* y_t *is the same whether conditioned on past values of* x_t *or not.*

Granger causality can be simply illustrated in a bivariate VAR.

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \phi_{11,2} & \phi_{12,2} \\ \phi_{21,2} & \phi_{22,2} \end{bmatrix} \begin{bmatrix} x_{t-2} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

In this model, if $\phi_{21,1} = \phi_{21,2} = 0$ then x_t does not Granger cause y_t. If this is the case, one may be tempted to think that

$$y_t = \phi_{22,1} y_{t-1} + \phi_{22,2} y_{t-2} + \epsilon_{2t}$$

is a correct specification of y_t. However, it is not: ϵ_{1t} and ϵ_{2t} can be contemporaneously correlation. If it happens to be the case that x_t does not Granger cause y_t and ϵ_{1t} and ϵ_{2t} have no contemporaneous correlation, then y_t is said to be weakly exogenous and y_t can be modeled completely independently of x_t.

Finally it is worth noting that $\{x_t\}$ not Granger causing $\{y_t\}$ says nothing about wether $\{y_t\}$ Granger causes $\{x_t\}$.

One limitation of GC is that it doesn't account for indirect effects. For example, suppose x_t and y_t are both Granger Caused by z_t. When this is the case, x_t will usually Granger Cause y_t even when it has no effect once z_t has been conditioned on. Specifically,

[3]Technically, this definition is for Granger Causality in the mean. Other definition exist for Granger causality in the variance (replace conditional expectation with conditional variance) and distribution (replace conditional expectation with conditional distribution).

$$\mathsf{E}\left[y_t|x_{t-1},y_{t-1},z_{t-1},...\right]=\mathsf{E}\left[y_t|y_{t-1},z_{t-1},...\right]$$

but

$$\mathsf{E}\left[y_t|x_{t-1},y_{t-1},...\right]\neq\mathsf{E}\left[y_t|y_{t-1},...\right]$$

Testing Testing for Granger causality in a VAR(p) is usually conducted using a likelihood ratio test. In this specification,

$$\mathbf{y}_t=\mathbf{\Phi}_0+\mathbf{\Phi}_1\mathbf{y}_{t-1}+\mathbf{\Phi}_2\mathbf{y}_{t-2}+...+\mathbf{\Phi}_p\mathbf{y}_{t-p}+\epsilon_t$$

$\{y_{jt}\}$ does not Granger cause $\{y_{it}\}$ if $\phi_{ij,1}=\phi_{ij,2}=...=\phi_{ij,p}=0$. The likelihood ratio test can be computed as

$$\left(T-pK^2+K\right)\left(\ln|\mathbf{\Sigma}_r|-\ln|\mathbf{\Sigma}_u|\right)\rightarrow\chi^2_p$$

where $\mathbf{\Sigma}_r$ is the estimated residual covariance when the null of no Granger causation is imposed ($\mathsf{H}_0:\phi_{ij,1}=\phi_{ij,2}=...=\phi_{ij,p}=0$) and u is the estimated covariance in the unrestricted VAR(p). If there is no Granger causation in your VAR, it's probably not a good idea to use one.

Example: Campbell's VAR

Campbell's VAR will be used to illustrate testing for Granger causality. Table 7.3 contains the results of Granger causality tests from a VAR which included lags 1, 3 and 12 (as chosen by the BIC) for the 5 series in Campbell's VAR. Tests of y_t causing y_t have been omitted as these aren't particularly informative in the multivariate context. The table tests whether the variables in the left hand column Granger cause the variables across the top. Remember that the null is no causality so that rejection (large test statistics and small $p$values) means there is a relationship. From the table, it can be seen that every variable causes at least one other variable since each row contains a $p$value indicating significance using standard test sizes (5 or 10%). It can also be seen that every variable is caused by another by examining the $p$values column by column.

7.1.3 Impulse Response Function

The second concept exclusive to vector analysis is the impulse response function. In the univariate world, the ACF was sufficient to understand shocks decay. When analyzing vector data, this is not longer the case. A shock to one series has an immediate effect but can also affect the other variables in a system which, in turn, can feed back into the original variable. After a few iterations of this cycle, it can be difficult to determine how a shock propagates even in a simple VAR(1).

Definition 14 *The impulse response function of $y_i$with respect to a shock in ϵ_j , for any j and i, is defined as the change in y_{it+s}, $s\geq 0$ for a unit shock in ϵ_{jt}. This definition is somewhat hard*

	VWM		LBR		RTB		TERM		DIV	
Exclusion	Test	P-val	Test	P-val	Test	P-val	Test	P-val	Test	P-val
VWM	-	-	3.08	0.38	2.07	0.56	15.2	0.00	103.6	0.00
LBR	12.3	0.01	-	-	4.3	0.23	14.4	0.00	0.678	0.88
RTB	2.81	0.42	10.1	0.02	-	-	15	0.00	7.22	0.07
TERM	12.4	0.01	3.26	0.35	288.3	0.00	-	-	0.54	0.91
DIV	2.63	0.45	3.43	0.33	16.3	0.00	8.9	0.03	-	-
All	31.5	0.00	27.1	0.01	351.9	0.00	51.9	0.00	135.4	0.00

Table 7.3: Tests of Granger causality. This table contains tests where the variable on the left hand side is excluded from the regression for the variable along the top. Since the null is no GC, rejection indicates a relationship between past values of the variable on the left and contemporaneous values of variables on the top.

to parse and the impulse response function can be clearly illustrated through a vector moving average (VMA). As long as y_t is covariance stationarity it must have a VMA representation,

$$\mathbf{y}_t = \mu + \epsilon_t + \mathbf{\Theta}_1\epsilon_{t-1} + \mathbf{\Theta}_2\epsilon_{t-2} + ...$$

Using this VMA, the impulse response y_i with respect to a shock in ϵ_j is simply $\left\{1, \mathbf{\Theta}_{1[ii]}, \mathbf{\Theta}_{2[ii]}, \mathbf{\Theta}_{3[ii]}, ...\right\}$ if $i = j$ and $\left\{1, \mathbf{\Theta}_{1[ij]}, \mathbf{\Theta}_{2[ij]}, \mathbf{\Theta}_{3[ij]}, ...\right\}$ otherwise. The difficult part is finding the $\{\mathbf{\Theta}_l\}$,$l \geq 1$. In the simple VAR(1) model this is easy since

$$\mathbf{y}_t = \left(\mathbf{I}_K - \mathbf{\Phi}_1\right)^{-1}\mathbf{\Phi}_0 + \epsilon_t + \mathbf{\Phi}_1\epsilon_{t-1} + \mathbf{\Phi}_2\epsilon_{t-2} + ...$$

However, in more complicated models, whether higher order VARs or VARMAs, determining the MA(∞) form can be tedious. One surprisingly simply, but completely correct, method to compute the elements of $\{\mathbf{\Theta}_j\}$ is to simulate the effect of a unit shock of ϵ_{jt} directly. Suppose the model is a demeaned VAR(p),

$$\mathbf{y}_t = \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + ... + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \epsilon_t.$$

The impulse responses can be computed by 'shocking' ϵ_t by 1 unit and stepping the process forward. To use this procedure, set $\mathbf{y}_{t-1} = \mathbf{y}_{t-2} = \mathbf{y}_{t-p} = \mathbf{0}$ and then begin the simulation by setting $\epsilon_{jt} = 1$. The 0^{th} impulse will obviously be $\mathbf{e}_j$, a vector with a 1 in the j^{th} position and zeros everywhere else. The second impulse will be,

$$\mathbf{\Theta}_1 = \mathbf{\Phi}_1\mathbf{e}_j$$

while the second will be

$$\mathbf{\Theta}_2 = \mathbf{\Phi}_1^2\mathbf{e}_j + \mathbf{\Phi}_2\mathbf{e}_j$$

and the third is

$$\mathbf{\Theta}_3 = \mathbf{\Phi}_1^3\mathbf{e}_j + \mathbf{\Phi}_1\mathbf{\Phi}_2\mathbf{e}_j + \mathbf{\Phi}_2\mathbf{\Phi}_1\mathbf{e}_j + \mathbf{\Phi}_3\mathbf{e}_j$$

The p^{th} lag contains the original shock while the other coefficients are capturing the complicated dynamics of the changes in $\mathbf{y}_{t-s}, s < p$. While manual computation of an IR is tedious, it is trivial in computer packages such as Matlab or Eviews.

Correlated Shocks and non-unit Variance The previous discussion has made use of unit shocks, $\mathbf{e}_j$ which represent a change of 1 in j^{th} error. This presents two problems: actual errors do not have unit variances and are often correlated. The solution to these problems is to use standardized residuals and/or correlated residuals. Suppose that the residuals in a VAR have a covariance of $\mathbf{\Sigma}$. To simulate the effect of a shock to element j, $\mathbf{e}_j$ can be replaced with $\widetilde{\mathbf{e}}_j = \mathbf{\Sigma}^{1/2}\mathbf{e}_j$ and the impulses can be computed using the procedure previously outlined.

The effect of this is two fold. First, every series will generally have an instantaneous reaction to any shock when the errors are correlated. Second, the choice of matrix square root, $\mathbf{\Sigma}^{1/2}$, matters. The two matrix square roots are the Choleski and the spectral decomposition. The Choleski square root is a lower triangular matrix which imposes a natural order to the shocks. Shocking element j (using $\mathbf{e}_j$) has an effect of every series $1, 2, ..., j$but not on $j+1, ..., K$. In contrast the spectral is symmetric and a shock to the j^{th} error will generally effect every series instantaneously. Unfortunately there is no right choice. If there is a natural ordering in a VAR where shocks to one series can be reasoned to have no contemporaneous effect on the other series, then the Choleski is the correct choice. However, in many situations there is little theoretical guidance and the spectral decomposition is the natural choice.

7.1.4 Example: Impulse Response in Campbell's VAR

Campbell's VAR will be used to illustrate impulse response functions. Figure 7.3 contains the impulse responses of the term premium to shocks in the four other variables: equity returns, labor income growth, the relative t-bill rate and the dividend rate. The dotted lines represent 2 standard deviation confidence intervals. The term premium decreases subsequent to positive shocks in the market or in labor income. Presumably this indicates that the economy is improving and there are inflationary pressures driving up the short end of the yield curve. Increases in the RTB lead to increases in the term premium as do shocks to the dividend yield.

Confidence Intervals Impulse response functions, like the parameters of the VAR, are estimated quantities and subject to statistical variation. Hence, it is a good practice to place confidence bands around impulse response functions to allow anyone digesting your work to know whether an impulse response is large in a statistically meaningful way. Since the parameters of the VAR

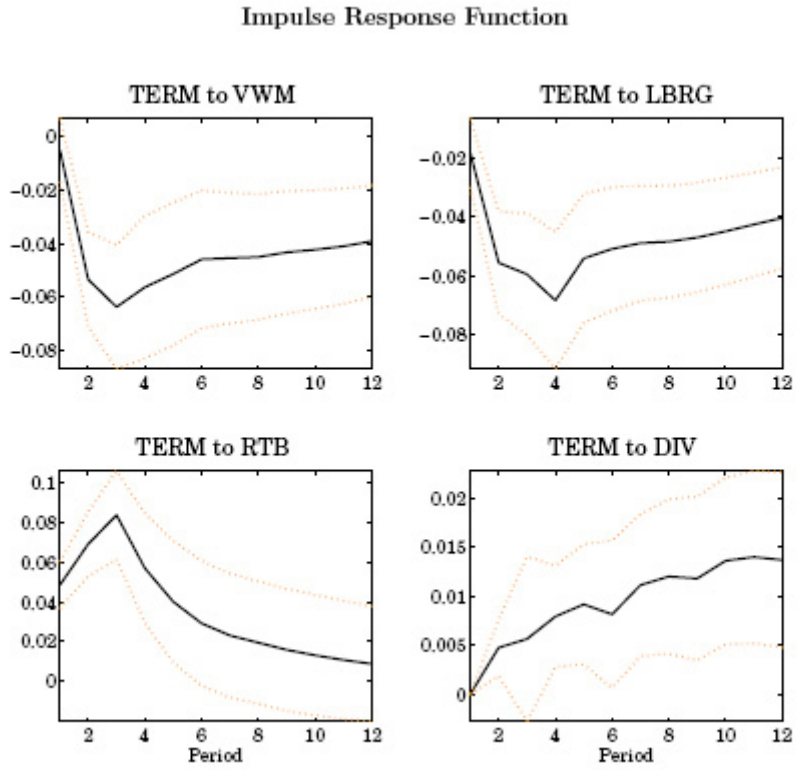

Figure 7.3: Impulse response function for 12 steps of the response of the term premium to equity returns, labor income growth, the relative T-bill rate and the dividend yield. The dotted lines represent 2 standard deviation (in each direction) confidence intervals.

are asymptotically normal (as long as it is stationary and the innovations are white noise), the impulse responses will also be asymptotically normal by applying a technique known as the delta method (covered in the GMM notes). Unfortunately, the derivation is extremely tedious and has essentially no intuitive value. Interested readers can refer to 11.7 in Hamilton (1994). Instead, two computational methods to construct confidence bands for impulse response functions will be described:

Monte Carlo and using a procedure known as the *bootstrap*, see Hamilton for details.

7.2 Cointegration

Two features of economic time series make cointegration an important tool: many contain unit roots and most equilibrium models require that deviations between key variables are transitory. Before formally defining cointegration, imagine a scenario where two important economics variables that contain unit roots, consumption and income, had no long-run relationship. If this were true, the values of these variables would grow arbitrarily far apart given enough time. Clearly this cannot occur so there must be some long run relationship between these two time series. Alternatively, consider the relationship between the spot and future price of Oil. Standard finance theory dictates that the future's price, f_t, is a conditionally unbiased estimate of the spot price in period $t+1$, s_{t+1} ($\mathsf{E}[s_{t+1}] = f_t$). Additionally, today's spot price is also an unbiased estimate of tomorrow's spot price ($\mathsf{E}[s_{t+1}] = s_t$. However, both of these price series contain unit roots. Combining these two identities reveals a cointegrating relationship: $s_t - f_t$ should be stationary even if the spot and future prices contain unit roots.

It is also important to note how cointegration is different from stationary VAR analysis. In stationary time series, whether scalar or when the multiple processes are linked through a VAR, the process is self-equilibrating; given enough time, a process will always mean revert to its unconditional level. In a VAR, both the series and linear combinations of the series are stationary. Cointegrated processes behavior is meaningfully different. Treated in isolation, each process contains a unit root and has shocks with permanent impact. However, when combined with another series, a cointegrated pair will show a tendency to revert to one another. In other words, a cointegrated pair is mean reverting to a stochastic trend.

Cointegration and error correction provide the tools to analyze temporary deviations from long run equilibria. In a nutshell, cointegrated time series may show temporary deviations from a long run trend but are ultimately mean reverting to this trend. It may also be useful to relate cointegration to what has been studied thus far: cointegration is to VARs as unit roots are to stationary time series.

7.2.1 Definition

Recall that an integrated process is defined as a process which is not stationary in levels but is stationary in differences. When this is the case, $\mathbf{y}_t$ is said to be I(1) and $\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ is described as I(0). Cointegration builds on this structure by defining relationships across series which can transform I(1) series into I(0) series.

Consider two series, x_t and y_t which are both I(1). These series are said to be cointegrated if there exists a vector β with both elements non-zero such that

$$\beta_1 x_t - \beta_2 y_t \sim \mathrm{I}(0)$$

Put another way, there exists a nontrivial linear combination of x_t and y_t which is stationary. This feature, when present, is a powerful link in the analysis of nonstationary data. While the data are extremely persistent and can ultimately take any value, there is a combination of the data which is always well behaved. Moreover, in many cases this relationship takes a meaningful form such as $y_t - x_t$. You should be aware that cointegrating relationships are only defined up to a constant. For example if $x_t - \beta y_t$ is a cointegrating relationship, then $2x_t - 2\beta y_t$ is also a cointegrating relationship. The standard practice is to chose one variable to normalize. For example, if $\beta_1 x_t - \beta_2 y_t$ was a cointegrating relationship, one normalized version would be $x_t - \beta_2/\beta_1 y_t = x_t - \widetilde{\beta} y_t$.

The definition in the general case is similar, albeit slightly more intimidating. A set of K I(1) variables $\mathbf{y}_t$ are said to be cointegrated if there exists a non-zero, reduced rank K by K matrix $\mathbf{\Pi}$ such that

$$\mathbf{\Pi}\mathbf{y}_t \sim \mathrm{I}(0).$$

The non-zero requirement is obvious: if $\mathbf{\Pi} = \mathbf{0}$ then $\mathbf{\Pi}\mathbf{y} = \mathbf{0}$ and is trivially I(0). The second requirement, that $\mathbf{\Pi}$ is of reduced rank, is not so obvious. This technical requirement is necessary since whenever $\mathbf{\Pi}$ is full rank and $\mathbf{\Pi}\mathbf{y}_t \sim \mathrm{I}(0)$, it must be the case that $\mathbf{y}_t$ is also I(0). However, in order for variables to be cointegrated they must also be integrated. Thus, if the matrix is full rank, there is no possibility for the common unit roots to cancel and it must have the same "integratedness" before and after the multiplication by $\mathbf{\Pi}$.

For example, suppose x_t and y_t are cointegrated and $x_t - \beta y_t$ is stationary. One choice for $\mathbf{\Pi}$ is

$$\mathbf{\Pi} = \begin{bmatrix} 1 & -\beta \\ 1 & -\beta \end{bmatrix}$$

To begin developing a feel for cointegration, examine the plots in figure 7.4. These four plots correspond to two nonstationary processes and two stationary processes all beginning at the same point and all using the same shocks. These plots contain data from a simulated VAR(1) with

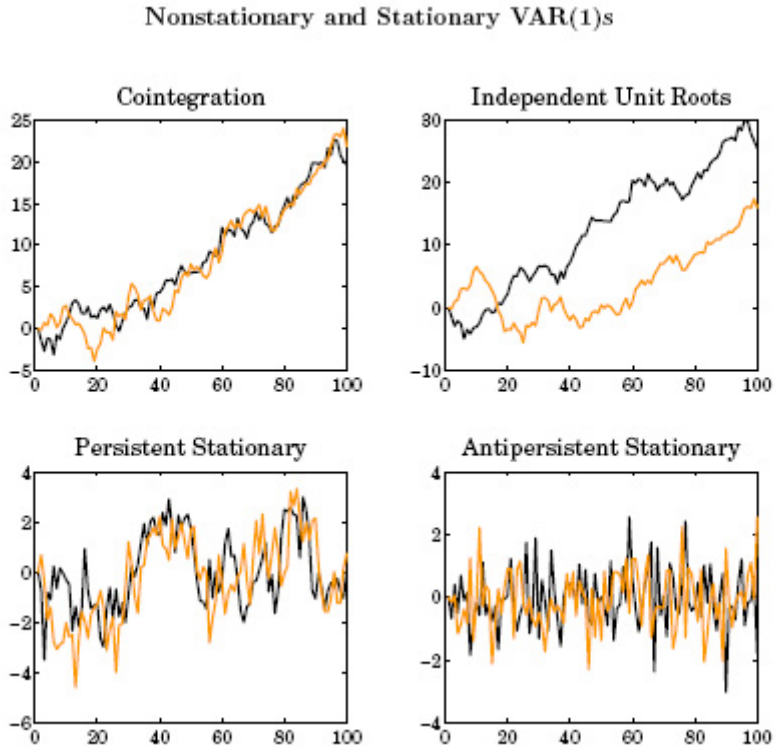

Figure 7.4: A plot of four time series that all begin at the same point in time and use the same shocks. All data were generated by $\mathbf{y}_t = \mathbf{\Phi}_{ij}\mathbf{y}_{t-1} + \epsilon_t$ where $\mathbf{\Phi}_{ij}$ varies depending on the process.

different parameters.

$$\mathbf{y}_t = \mathbf{\Phi}_{ij}\mathbf{y}_{t-1} + \epsilon_t$$

$$\mathbf{\Phi}_{11} = \begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix} \quad \mathbf{\Phi}_{12} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\lambda_i = 1, 0.6 \qquad \lambda_i = 1, 1$$

$$\mathbf{\Phi}_{21} = \begin{bmatrix} .7 & .2 \\ .2 & .7 \end{bmatrix} \quad \mathbf{\Phi}_{12} = \begin{bmatrix} -.3 & .3 \\ .1 & -.2 \end{bmatrix}$$

$$\lambda_i = 0.9, 0.5 \qquad \lambda_i = -0.43, -0.06$$

where λ_i are the eigenvalues of the parameter matrices. Note that the eigenvalues of the nonstationary processes contain the value 1 while the eigenvalues for the stationary processes are all less then 1 (in absolute value). Also, note that the cointegrated process has only one eigenvalue which is unity while the independent unit root process has two. Higher dimension cointegrated systems may contain between 1 and $K-1$ unit eigenvalues. The picture presents evidence of another issue in cointegration analysis: it can be very difficult to tell when two series are cointegrated, a feature

shared with unit root testing of scalar processes.

7.2.2 Error Correction Models (ECM)

The Granger representation theorem provides a key insight to understanding cointegrating relationships. Granger demonstrated that if a system if cointegrated then there exists an error correction model and if there is an error correction model then the system must be cointegrated. The error correction model is a form which governs short deviations from the trend (a stochastic trend or unit root). The simplest ECM is given by

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

which states that changes in x_t and y_t are related to the levels of x_t and y_t through the cointegrating matrix ($\mathbf{\Pi}$). However, since x_t and y_t are cointegrated, there exists β such that $x_t - \beta y_t$ is I(0). Substituting this into this equation, it can be rewritten as

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \begin{bmatrix} 1 & -\beta \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

so each the short-run dynamics of each variable take the form

$$\Delta x_t = \alpha_1 \left(x_{t-1} - \beta y_{t-1} \right) + \epsilon_t$$

The pieces of this ECM can be clearly labeled: $(x_{t-1} - \beta y_{t-1})$ is the deviation from the long run trend and α is the speed of adjustment parameter. ECM imposes one restriction of the α's: they cannot both be 0 (if they were, $\mathbf{\Pi}$ would also be 0). In its general form, an ECM can be augmented to allow past short run deviations to also influence present short run deviations or to include deterministic trends. In vector form, the generalized ECM is

$$\Delta \mathbf{y}_t = \mathbf{\Pi}_0 + \mathbf{\Pi} \mathbf{y}_{t-1} + \mathbf{\Pi}_1 \Delta \mathbf{y}_{t-1} + \mathbf{\Pi}_2 \Delta \mathbf{y}_{t-2} + ... + \mathbf{\Pi}_p \Delta \mathbf{y}_{t-p} + \epsilon_t$$

The Mechanics of the ECM It may not be obvious how a cointegrated VAR is transformed into an ECM. Consider a simple cointegrated bivariate VAR(1)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

To transform this VAR to an ECM, begin by subtracting $[x_{t-1} \quad y_{t-1}]'$ from both sides

$$\begin{aligned}
\begin{bmatrix} x_t \\ y_t \end{bmatrix} - \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} &= \left(\begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \\
&= \begin{bmatrix} -.2 & .2 \\ .2 & -.2 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \\
&= \begin{bmatrix} -.2 \\ .2 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}
\end{aligned}$$

In this example, the speed of adjustment parameters are $-.2$ for x_t and $.2$ for y_t and the normalized (on x_t) cointegrating relationship is $[1 \quad -1]$. In the general multivariate case, a cointegrated VAR(p) can be turned into an ECM by recursive substitution. Consider a cointegrated VAR(3),

$$\mathbf{y}_t = \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \mathbf{\Phi}_3 \mathbf{y}_{t-3} + \epsilon_t.$$

This system will be cointegrated if at least one but fewer than K eigenvalues of $\mathbf{\Pi} = \mathbf{\Phi}_1 + \mathbf{\Phi}_2 + \mathbf{\Phi}_3 - \mathbf{I}_K$ are not zero. To begin the transformation, add and subtract $\mathbf{\Phi}_3 \mathbf{y}_{t-2}$ to the right hand side

$$\begin{aligned}
\mathbf{y}_t &= \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \mathbf{\Phi}_3 \mathbf{y}_{t-2} - \mathbf{\Phi}_3 \mathbf{y}_{t-2} + \mathbf{\Phi}_3 \mathbf{y}_{t-3} + \epsilon_t \\
&= \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \mathbf{\Phi}_3 \mathbf{y}_{t-2} - \mathbf{\Phi}_3 \Delta \mathbf{y}_{t-2} + \epsilon_t \\
&= \mathbf{\Phi}_1 \mathbf{y}_{t-1} + (\mathbf{\Phi}_2 + \mathbf{\Phi}_3) \mathbf{y}_{t-2} - \mathbf{\Phi}_3 \Delta \mathbf{y}_{t-2} + \epsilon_t
\end{aligned}$$

then add and subtract $(\mathbf{\Phi}_2 + \mathbf{\Phi}_3) \mathbf{y}_{t-1}$ to the right side,

$$\begin{aligned}
\mathbf{y}_t &= \mathbf{\Phi}_1 \mathbf{y}_{t-1} + (\mathbf{\Phi}_2 + \mathbf{\Phi}_3) \mathbf{y}_{t-1} - (\mathbf{\Phi}_2 + \mathbf{\Phi}_3) \mathbf{y}_{t-1} + (\mathbf{\Phi}_2 + \mathbf{\Phi}_3) \mathbf{y}_{t-2} - \mathbf{\Phi}_3 \Delta \mathbf{y}_{t-2} + \epsilon_t \\
&= (\mathbf{\Phi}_1 + \mathbf{\Phi}_2 + \mathbf{\Phi}_3) \mathbf{y}_{t-1} - (\mathbf{\Phi}_2 + \mathbf{\Phi}_3) \Delta \mathbf{y}_{t-1} - \mathbf{\Phi}_3 \Delta \mathbf{y}_{t-2} + \epsilon_t
\end{aligned}$$

Finally, subtract $\mathbf{y}_{t-1}$ from both sides,

$$\begin{aligned}
\Delta \mathbf{y}_t &= (\mathbf{\Phi}_1 + \mathbf{\Phi}_2 + \mathbf{\Phi}_3 - \mathbf{I}_K) \mathbf{y}_{t-1} - (\mathbf{\Phi}_2 + \mathbf{\Phi}_3) \Delta \mathbf{y}_{t-1} - \mathbf{\Phi}_3 \Delta \mathbf{y}_{t-2} + \epsilon_t \\
&= \mathbf{\Pi} \mathbf{y}_{t-1} + \mathbf{\Pi}_1 \Delta \mathbf{y}_{t-1} + \mathbf{\Pi}_2 \Delta \mathbf{y}_{t-2} + \epsilon_t
\end{aligned}$$

which is equivalent to

$$\Delta \mathbf{y}_t = \alpha \beta' \mathbf{y}_{t-1} + \mathbf{\Pi}_1 \Delta \mathbf{y}_{t-1} + \mathbf{\Pi}_2 \Delta \mathbf{y}_{t-2} + \epsilon_t$$

where α contains the speed of adjustment parameters and β contains the cointegrating vectors. This recursion can be used to transform any cointegrated VAR(p)

$$\mathbf{y}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + ... + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \epsilon_t$$

into its ECM from

$$\Delta \mathbf{y}_t = \mathbf{\Pi} \mathbf{y}_{t-1} + \mathbf{\Pi}_1 \Delta \mathbf{y}_{t-1} + \mathbf{\Pi}_2 \Delta \mathbf{y}_{t-2} + ... + \mathbf{\Pi}_{p-1} \Delta \mathbf{y}_{t-p+1} + \epsilon_t$$

using the identities $\mathbf{\Pi} = -\mathbf{I}_K + \sum_{i=1}^{p} \mathbf{\Phi}_i$ and $\mathbf{\Pi}_j = -\sum_{i=j+1}^{p} \mathbf{\Phi}_i$

Cointegrating Vectors

The key to understanding cointegration in systems with 3 or more variables is to note that the matrix which governs the cointegrating relationship, $\mathbf{\Pi}$, can always be decomposed into two matrices,

$$\mathbf{\Pi} = \alpha\beta'$$

where α and β are both K by r matrices where r is the number of cointegrating relationships and $K - r$ the number of uncorrelated stochastic trends. If $r = 0$ the vector process is composed of K stochastic trends, if $r = K$, then the process is stationary.

Relationship to Common Features and common trends

Cointegration is special case of a broader concept known as common features. In the case of cointegration, both series have a common stochastic trend (or common unit root). Other examples of common features which have been examined are common heteroskedasticity, defined as x_t and y_t are heteroskedastic but there exists a combination $x_t - \beta y_t$ which is not, and common nonlinearities which are defined in an analogous manner (replacing heteroskedasticity with nonlinearity). When modeling multiple time series, you should always consider whether the aspects you are interested in may be common.

7.2.3 Testing

Testing for cointegration shares one important feature with its scalar counterpart (unit root testing): it can be complicated. Two methods will be presented, the original Engle-Granger 2-step procedure and the more sophisticated Johansen method. The Engle-Granger method is generally only applicable if there are two variables or the cointegrating relationship is known (e.g. an accounting identity where the left-hand side has to add up to the right-hand side). The Johansen method is substantially more general and can be used to examine complex systems with many variables and numerous cointegrating relationships.

Engle-Granger The Engle-Granger method exploits the key feature of any cointegrated pair: if data are cointegrated, a linear combination of the two series can be constructed that is stationary. If not, any linear combination will remain I(1). The Engle-Granger methodology begins by specifying the cross-section regression

$$y_t = \beta x_t + \epsilon_t$$

and estimating $\widehat{\beta}$ using OLS. It may be necessary to include a constant and

$$y_t = \tau + \beta x_t + \epsilon_t$$

can be estimated instead if the residuals form the first regression are not mean 0. Once the coefficients have been estimated, the model residuals, $\widehat{\epsilon}_t$, can be tested for the presence of a unit root. If x_t and y_t were both I(1) and $\widehat{\epsilon}_t$ is I(0), the series appear to be cointegrated. The procedure concludes by using $\widehat{\epsilon}_t$ to estimate the error correction form of the model to examine parameters which may be of interest (e.g. the speed of convergence parameters).

Step 1: Begin by analyzing x_t and y_t in isolation to ensure that they are both integrated. You should plot the data and perform ADF tests. Remember, variables can only be cointegrated if they are integrated.

Step 2: Estimate the long run relationship by estimating

$$y_t = \tau + \beta x_t + \epsilon_t$$

using OLS and computing the estimated residuals $\{\widehat{\epsilon}_t\}$. Use an ADF test (or DF-GLS for more power) and test $\mathsf{H}_0 : \varrho = 0$ against $\mathsf{H}_0 : \varrho < 0$ in the regression

$$\Delta\widehat{\epsilon}_t = \varrho\widehat{\epsilon}_t + \delta_1\Delta\widehat{\epsilon}_{t-1} + ... + \delta_p\Delta\widehat{\epsilon}_{t-p} + \eta_t$$

It may be necessary to include deterministic trends. Fortunately, the classic procedure for examining time series for unit roots can be used to examine if this series contains a unit root. If the null is rejected and $\widehat{\epsilon}_t$ is stationary, then x_T and y_t appear to be cointegrated. Alternatively, if $\widehat{\epsilon}_t$ still contains a unit root, the series are not cointegrated.

Step 3: If a cointegrating relationship is found, using the estimated parameters, specify and estimate the error correction form

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \pi_{01} \\ \pi_{02} \end{bmatrix} + \begin{bmatrix} \alpha_1 (y_t - \beta x_t) \\ \alpha_2 (y_t - \beta x_t) \end{bmatrix} + \mathbf{\Pi}_1 \begin{bmatrix} \Delta x_{t-1} \\ \Delta y_{t-1} \end{bmatrix} + ... + \mathbf{\Pi}_p \begin{bmatrix} \Delta x_{t-p} \\ \Delta y_{t-p} \end{bmatrix} + \begin{bmatrix} \eta_{1t} \\ \eta_{2t} \end{bmatrix}$$

You should spot something funny in these regressions: they are not linear in the parameters. They both have interactions between the α and β parameters so OLS cannot be used. Engle and Granger noted that $(y_t - \beta x_t)$ can be replaced with $\widehat{\epsilon}_t$,

The parameters of these equation can be estimated by applying OLS to each equation.

Step 4: The final step is to assess the model adequacy and test hypotheses about α_1 and α_2. Standard diagnostic checks including plotting the residuals and examining the ACF should be used to examine model adequacy. Impulse response functions for the short run deviations can be examined to assess the effect of a shock on the deviation of the series from the long term trend.

Cointegration in Consumption, Asset Prices and Income

To illustrate cointegration and error correction, three series which have played an important role in the revival of the CCAPM in recent years will be examined. These three series are consumption (c), asset prices (a) and labor income (y). The data were made available by Martin Lettau

Series	T-stat	P-val	Lags
c	-1.162	0.68	1
a	-0.022	0.95	0
y	-1.276	0.64	0
$\hat{\epsilon}_t$	-4.071	0.00	0

Table 7.4: Unit root test results. The top three lines contain the results of ADF tests for unit roots in the three components of cay: Consumption, Asset Prices and Aggregate Wealth. None of these series can reject the null of a unit root. The final line contains the results of a unit root test on the estimated residuals where the null is strongly rejected indicating that there is a cointegrating relationship between the three. The lags column reports the number of lags used in the ADF procedure as automatically selected using the AIC.

on his website[4] `http://pages.stern.nyu.edu/~mlettau/data/cay_q_06Q4.txt` and contain quarterly data from 1951:4 until 2006:4.

To begin the Engle-Granger procedure it is necessary to perform unit root testing and examine the data.

Table 7.4 and figure 7.5 contain these tests and graphs. All three series cannot reject a unit root and have time-detrended errors which appear to be nonstationary.

The next step is to specify the cointegrating regression

$$c_t = \beta_1 + \beta_2 a_t + \beta_3 y_t + \epsilon_t$$

and to estimate the long run relationship using OLS. The estimated cointegrating vector from the first stage OLS was $[1 \quad -0.207 \quad -0.678]$ which corresponds to a long run relationship of $c_t - 0.207a_t - 0.678y_t$. Finally, the residuals were tested for the presence of a unit root. The results of this test are in the final line of table 7.4 and indicate a strong rejection of a unit root in the errors. Based on the Engle-Granger procedure, these three series appear to be cointegrated.

Spurious Regression and Balance

In light of cointegration, it may seem reasonable to regress any I(1) variable on any other I(1) variable irrespective of any economic theory motivating a relationship. You should not do this. When two related I(1) variables are regressed on one another, the cointegrating relationship dominates and the regression coefficients can be directly interpreted as the cointegrating vectors. However, then two unrelated I(1) variables are regressed on one another, the regression coefficient is no

[4] More recent data is available at his new website
`http://faculty.haas.berkeley.edu/lettau/data_cay.html`

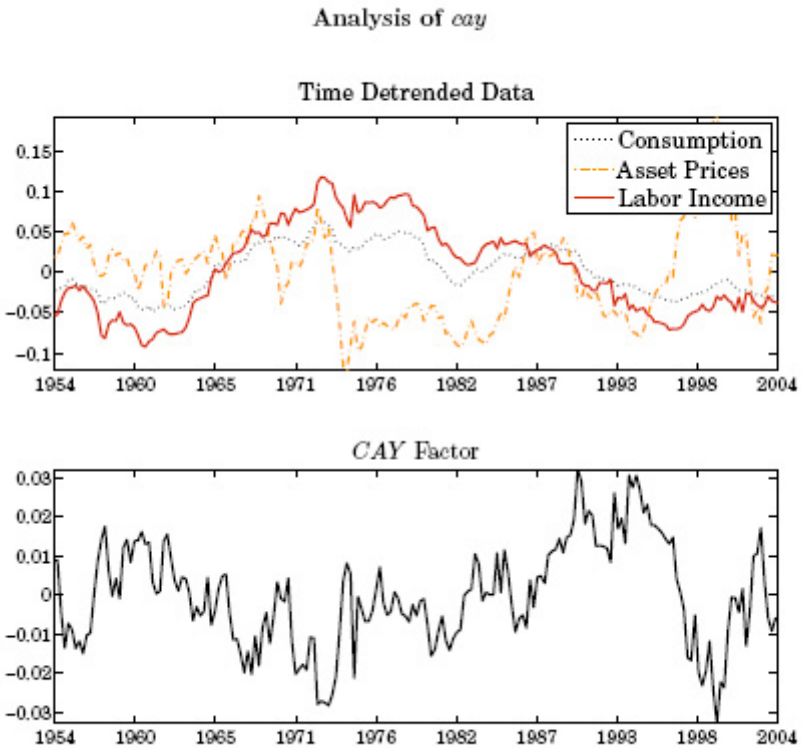

Figure 7.5: The top panel contains plots of time detrended residuals from regressions of consumption, asset prices and labor income on a linear time trend. While it is difficult to say, they may be unit roots and are clearly highly persistent. The bottom panel contains a plot of $\epsilon_t = c_t - 0.207a_t - 0.678y_t$ which is commonly known as the cay scaling factor (pronounced consumption-aggregate wealth). The null of a unit root is strongly rejected for this series indicating that the three original series are cointegrated.

longer consistent for anything. For example, let x_t and y_t be independent random walk processes.

$$x_t = x_{t-1} + \eta_t$$
$$y_t = y_{t-1} + \nu_t$$

In the regression

$$x_t = \beta y_t + \epsilon_t$$

$\widehat{\beta}$ is not consistent for 0 despite the independence of x_t and y_t.

Models that include independent I(1) processes are known as spurious regressions. When the regressions are spurious, the estimated $\widehat{\beta}$ can take any value and typically have t-stats which indicates significance at conventional levels. The solution to this problems is simple: whenever regressing one I(1) variable on another, always check to be sure that the regression residuals are I(0) and not I(1). In other words, use the Engle-Granger procedure.

Balance is another important concept when data which contain both stationary and integrated data. An equation is said to be balanced if all variables have the same order of integration. The usual case occurs when a stationary variable (I(0)) is related to other stationary variables. However, other situation may arise and it is useful to consider the four possibilities:

- I(0) on I(0): The usual case. Standard asymptotic arguments apply.
- I(1) on I(0): This regression is unbalanced. An I(0) variable can never explain the long-run variation in an I(1) variable. The usual solution is to difference the I(1) and the examine whether the short run dynamics in the I(1) can be explained by the I(0).
- I(1) on I(1): One of two things are possible: cointegration or spurious regression.
- I(0) on I(1): This regression is unbalanced. An I(1) variable can never explain the variation in an I(0) variable. Unlike spurious regressions, the t-stat still has a standard asymptotic distribution although caution is needed as small sample properties can be very poor. As a rule, imbalanced regressions are not meaningful in explaining economic phenomena.

7.2.4 Johansen Methodology

Recall that one of the requirements for an ECM to indicate that a set of integrated series are cointegrated is that $\mathbf{\Pi}$ has reduced rank.

$$\Delta \mathbf{y}_t = \mathbf{\Pi} \mathbf{y}_{t-1} + \mathbf{\Pi}_1 \Delta \mathbf{y}_{t-1} + \mathbf{\Pi}_2 \Delta \mathbf{y}_{t-2} + ... + \mathbf{\Pi}_{p-1} \Delta \mathbf{y}_{t-p+1} + \epsilon_t$$

This means that the number of non-zero eigenvalues is between 1 and $K - 1$. The Johansen cointegration framework uses the magnitude of these eigenvalue to directly test for cointegration. Additionally, the Johansen methodology extends the Engle-Granger procedure to allow the number

of cointegrating relationships to be determined, a key feature missing from the Engle-Granger two-step procedure.

The Johansen methodology makes use of two statistics, the trace statistic (λ_{trace}) and the maximum eigenvalue statistic (λ_{max}). Both statistics test functions of the estimated eigenvalues of $\mathbf{\Pi}$ but employ different null hypotheses and have different alternatives. The trace statistic tests the null that the number of cointegrating relationships is less than or equal to r against an alternative that the number if greater than r. Define $\widehat{\lambda}_i$ to be the eigenvalues of $\widehat{\mathbf{\Pi}}_1$ and let them be ordered such that $\lambda_1 > \lambda_2 > ... > \lambda_K$. The trace statistic is defined

$$\lambda_{\text{trace}}(r) = -T \sum_{i=r+1}^{K} \ln\left(1 - \widehat{\lambda}_i\right).$$

There are K possible trace statistics. The first, $\lambda_{\text{trace}}(0) = -T \sum_{i=1}^{K} \ln\left(1 - \widehat{\lambda}_i\right)$, tests that null of no cointegrating relationships against an alternative that the number of cointegrating relationships is 1 or more. For example, if the were no cointegrating relationships, each of the eigenvalues would be close to zero and $\lambda_{\text{trace}}(0) \approx 0$. Every unit root in the ECM corresponds to a zero eigenvalue in $\mathbf{\Pi}$. When the series are cointegrated, $\mathbf{\Pi}$ will have one or more non-zero eigenvalues. Like unit root tests, cointegration tests have nonstandard distributions. Fortunately, most software packages, including Eviews, return the appropriate critical values for the length of data and any included deterministic regressors.

The maximum eigenvalue test examines the null that the number of cointegrating relationships is r against the alternative that the number is $r + 1$. The maximum eigenvalue statistic is defined $\lambda_{\text{max}}(r, r+1) = -T \ln\left(1 - \widehat{\lambda}_{r+1}\right)$

Intuitively, if there are $r + 1$ cointegrating relationships, then the $r + 1^{th}$ eigenvalue should be different from zero and the value of $\lambda_{\text{max}}(r, r+1)$ should be large. On the other hand, if there are only r cointegrating relationships, the $(r+1)^{th}$ eigenvalue should be different from zero and the statistic will be small. Again, the distribution is nonstandard but most statistical packages provide appropriate critical values for the number of observations and the included deterministic regressors.

The steps to implement the Johansen procedure are:

Step 1: Plot the data series being analyzed and perform univariate unit root testing. A set of variables can only be cointegrated if that are all integrated. If the series are trending, either linearly or quadratically, make note of this and remember to include deterministic terms when estimating the ECM.

Step 2: The second stage is to lag length selection. Select the lag length Using one of the procedures outlined in the VAR lag length selection section (General-to-Specific, AIC or SIC). For example, to use the General-to-Specific approach, first select a maximum lag length l and then test

l lags against $l-1$ using a likelihood ratio test,

$$LR = \left(T - l \cdot K^2\right)\left(\ln\left|\mathbf{\Sigma}_{l-1}\right| - \ln\left|\mathbf{\Sigma}_{l}\right|\right) \sim \chi^2_K.$$

Repeat the test decreasing the number of lags by one each iteration until the LR rejects the null that the smaller model is appropriate.

Step 3: Estimate the selected model

$$\Delta\mathbf{y}_t = \mathbf{\Pi}\mathbf{y}_{t-1} + \mathbf{\Pi}_1\Delta\mathbf{y}_{t-1} + \mathbf{\Pi}_2\Delta\mathbf{y}_{t-2} + ... + \mathbf{\Pi}_{p-1}\Delta\mathbf{y}_{t-p+1} + \epsilon_t$$

and determine the rank of $\mathbf{\Pi}$ where p is the lag length previously determined. If the series appear to be trending, then the model should include a deterministic term and

$$\Delta\mathbf{y}_t = \Pi_0 + \mathbf{\Pi}\mathbf{y}_{t-1} + \mathbf{\Pi}_1\Delta\mathbf{y}_{t-1} + \mathbf{\Pi}_2\Delta\mathbf{y}_{t-2} + ... + \mathbf{\Pi}_{p-1}\Delta\mathbf{y}_{t-p+1} + \epsilon_t$$

should be estimated. Using the λ_{trace} and λ_{max} tests, determine the cointegrating rank of the system. For an example of using these statistics, see next section.

Step 4: Analyze the normalized cointegrating vectors to determine whether these conform to implications in economic or financial theory. Hypothesis tests on the cointegrating vector can also be performed to examine whether the long run relationships conform to a prior beliefs.

Step 5: The final step of the procedure is to assess the adequacy of the model by plotting and analyzing the residuals. This step should be the final task in any analysis of time series data, not just the Johansen methodology. If the residuals do not resemble white noise, the model should be reconsidered. If the residuals are stationary, then more lags may be necessary. If they are I(1), then the system may not be cointegrated.

Example: Consumption Aggregate Wealth

The Johansen methodology begins by examining the original data for unit roots. Since it has been clearly established that all series have unit roots, this will be skipped. The next step tests eigenvalues of $\mathbf{\Pi}$ in the error correction model

$$\Delta\mathbf{y}_t = \Pi_0 + \mathbf{\Pi}\mathbf{y}_{t-1} + \mathbf{\Pi}_1\Delta\mathbf{y}_{t-1} + \mathbf{\Pi}_2\Delta\mathbf{y}_{t-2} + ... + \mathbf{\Pi}_{p-1}\Delta\mathbf{y}_{t-p+1} + \epsilon_t$$

using λ_{trace} and λ_{max} tests. Table 7.5 contains the results of the two tests. You should notice that all of the $p$values indicate no significance at conventional levels (5-10%) and the Johansen methodology leads to a different conclusion than the Engle-Granger methodology: there is no evidence these three series are cointegrated. This seem counter intuitive, but testing alone cannot provide a reason why this has occurred; only theory can.

Trace Test

Null	Alternative	λ_{trace}	Crit. Val.	P-val
$r = 0$	$r \geq 1$	20.73	29.79	0.37
$r = 1$	$r \geq 2$	4.330	15.49	0.87
$r = 2$	$r = 3$	0.328	3.841	0.56

Max Test

Null	Alternative	λ_{max}	Crit. Val.	P-val
$r = 0$	$r = 1$	16.40	21.13	0.20
$r = 1$	$r = 2$	4.002	14.26	0.85
$r = 2$	$r = 3$	0.328	3.841	0.56

Table 7.5: Results of testing using the Johansen methodology. Unlike the Engle-Granger procedure, no evidence of cointegration is found using either test statistic.

Chapter 8

Univariate Volatility Modeling

Note: The primary references for these notes are chapters 10 and 11 in Taylor (2005). Alternative, but less comprehensive, treatments can be found in chapter 21 of Hamilton (1994) or chapter 4 of Enders (2004). Many of the original articles can be found in Engle (1995).

Volatility measurement and modeling is the most important contribution of financial econometrics to date. This chapter begins by introducing volatility as a meaningful concept and then described a widely employed model: the ARCH model. The chapter describes various forms ARCH family models, some of their important properties and how the unknown parameters can be estimated. Attention then turns to a new tool in the measurement and modeling of financial econometrics, realized volatility, before concluding with a discussion of volatility forecasting.

8.1 Why does volatility change?

Time-varying volatility is a pervasive fact of modern empirical finance. Time variation in volatility is so pervasive that it can be difficult to find an asset return series which does not have time-varying volatility. However, statistical descriptions of the time-variation, the theme of this chapter, cannot provide an explanation for why the volatility of many asset returns is time-varying. A number of explanations have been proffered, although none are completely satisfactory.

- **News Announcements:** The arrival of new news forces agents to update beliefs. These new beliefs cause portfolio rebalancing and high periods of volatility correspond to agents dynamically solve for the new asset prices. Unfortunately, there is a surprising lack of evidence that macroeconomic news, defined as announcement surprises, has an impact on volatility.

- **Leverage:** Suppose firms are financed using debt bonds and equity and that the firms profit stream has a volatility $\overline{\sigma}$. Since the portion of the firm that is levered cannot express the volatility of the firm, the non-levered portion (equity) must reflect all of the variation in

cash flows. Thus, decreases in equity prices should lead to increased volatility. While this phenomena is pervasive in financial asset returns, the amount if variation in leverage ratios is insufficient to explain the amount of variation in volatility.

- **Volatility Feedback:** Volatility feedback is motivated by a model where the volatility of an asset is priced. When the price of an asset falls, the volatility must increase to reflect the increased expected return (in the future) of this asset and vice-versa. There is evidence that this explanation is empirically supported although it cannot explain the totality of time-variation of volatility.

- **Illiquidity:** Short run spells of illiquidity can cause time varying volatility even when shocks are IID. Intuitively, if the market is over sold (bought), a small negative (positive) shock will cause a small decrease (increase) in demand. However, since there are few participants willing to buy (sell), this shock has a large effect of prices. This liquidity runs tend to last from 20 minutes to a few days and this explanation cannot explain the long cycles in present volatility.

- **State Uncertainty:** Asset prices are an important tool that agents express beliefs about the state of the economy. When the state is uncertain, slight changes in beliefs cause a large shift in portfolio holdings which then feeds back into beliefs about the state. This feedback loop can generate time-varying volatility and should have the biggest effect when the economy is transitioning from an expansion to a recession or a recession to an expansion.

The actual cause of the time-variation in volatility is likely a combination of all of these and some not present on this list.

8.1.1 What is volatility?

Volatility comes in many shapes and forms. To be precise, it is important to be clear what is meant when the term volatility used.

Volatility The volatility is simply the standard deviation. Volatility is often preferred to variance as it is measured in the same units as the data it corresponds to. For example, when using returns, the volatility is also in returns, and a volatility of 5% indicates that $\pm 5\%$ is a meaningful quantity.

Realized Volatility Realized volatility has historically been used to denote a measure of the volatility over some arbitrary period of time,

$$\widehat{\sigma} = \sqrt{T^{-1} \sum_{t=1}^{T} (r_t - \widehat{\mu})^2}$$

but is now used to describe a volatility measure constructed using high-frequency data.

Conditional Volatility Conditional volatility is the expected volatility at some future point in time $t+h$ based on all available information up to time t ($\mathcal{F}_t$). The one-period ahead conditional volatility is denoted $\mathsf{E}_t\left[\sigma_{t+1}\right]$

Implied Volatility Implied volatility is the average volatility which would be required to correctly price an option. For example, if the Black-Scholes pricing formula were valid, the price of a European call option depends on the current price of the underlying, the strike, the risk-free rate, the time-to-maturity and the *volatility*,

$$BS\left(S_t, K, r, t, \sigma\right) = C$$

where C is the price of the call. The implied volatility is the value which solves the Black-Scholes taking the option price, the strike, the risk-free and the time-to-maturity as given,

$$\sigma\left(S_t, K, r, t, C\right)$$

Annualized Volatility When volatility is measured over an interval other than a year, such as a day, week or month, it can always be scaled to reflect what the volatility of this asset would be if held for a year.

For example, if σ denotes the daily volatility of an equity and there are 252 trading days in a year, the annualized volatility is $\sqrt{252}\sigma$. Annualized volatility is a useful tool to remove the sampling interval from reported volatilities.

Variance All of the uses of volatility can be replaced with variance and most of this chapter is dedicated to modeling the *conditional variance*, denoted $\mathsf{E}_t\left[\sigma_{t+h}^2\right]$

8.2 ARCH Models

When it comes to financial econometrics, an arch is not an architectural feature of a building; it is a fundamental tool for analyzing the time-variation in conditional variance. The **ARCH** (**A**uto**R**egressive **C**onditional **H**eteroskedasticity) family of models have become pervasive in modeling of volatility for two reasons: they are essentially ARMA models which provides a natural tie-in to standard time-series analysis and ARCH-family models are generally easy to estimate.

In the following, we mostly work with Gaussian distributions to facilitate computations. But this is clearly a limitation. t-distributions constitute a standard choice of fat-tailed distributions. The lower the number of degrees of freedom, the higher the kurtosis.

In this chapter as before, you will need to keep in mind the Law of Iterated Expectations:

$$\begin{aligned}\mathsf{E}\left[Y\right] &= \mathsf{E}_X\left[\mathsf{E}\left[Y|X\right]\right] \\ \mathsf{Var}\left[Y\right] &= \mathsf{E}_X\left[\mathsf{Var}\left[Y|X\right]\right] \\ &\quad + \mathsf{Var}_X\left[\mathsf{E}\left[Y|X\right]\right]\end{aligned}$$

where the notation E_X is here to remind you that the only random element with respect to which the expectation is computed is X.

8.2.1 The ARCH model

The complete ARCH(P) model can be described by

$$\begin{aligned}
r_t &= \mu_t + \epsilon_t \\
\mu_t &= \phi_0 + \phi_1 r_{t-1} + ... + \phi_S r_{t-S} \\
\sigma_t^2 &= \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + ... + \alpha_P \epsilon_{t-P}^2 \\
\epsilon_t &= \sigma_t e_t \\
e_t &\sim \mathsf{WN}(0,1)
\end{aligned}$$

The key aspect of this model is that the variance of the shock, ϵ_t, is time varying and depends on the past P shocks, ϵ_{t-1}, ϵ_{t-2}, ..., ϵ_{t-P} through their squares. σ_t^2 is the time $t-1$ conditional variance and is completely determined at time $t-1$. This can be verified by noting that all of the RHS (right-hand side) variables in the σ_t^2 equation are measurable at time $t-1$ (check the time indices). A crucial requirement of ARCH models is that the parameters of the variance evolution, α_1, α_2, ... must all be positive. The intuition behind the result is that if one were negative, eventually a shock would be sufficiently large to force the variance negative resulting in a nonsensical model for the conditional variance. An alternative way to describe an ARCH(P) model is

$$\begin{aligned}
r_t|\mathcal{F}_t &\sim \mathsf{N}\left(\mu_t, \sigma_t^2\right) \\
\mu_t &= \phi_0 + \phi_1 r_{t-1} + ... + \phi_S r_{t-S} \\
\sigma_t^2 &= \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + ... + \alpha_P \epsilon_{t-P}^2 \\
\epsilon_t &= r_t - \mu_t
\end{aligned}$$

which is read "r_t given the information set at time $t-1$ is conditionally normal with mean μ_t and variance σ_t^2. In these two examples, I have chosen an AR(S) to model the conditional mean, although this choice is arbitrary. The conditional mean of r_t can depend on any time $t-1$ measurable variable including its own lags, shocks (for a MA model) or exogenous variables such as the default or term premium.

The conditional variance is denoted σ_t^2

$$\mathsf{E}_{t-1}\left[\epsilon_t^2\right] = \mathsf{E}_{t-1}\left[e_t^2 \sigma_t^2\right] = \sigma_t^2 \mathsf{E}_{t-1}\left[\epsilon_t^2\right] = \sigma_t^2$$

while the *unconditional* variance is denoted using $\overline{\sigma}$,

$$\mathsf{E}\left[\epsilon_t^2\right] = \overline{\sigma}^2$$

It is crucial that the nature of the expectation, whether conditional or unconditional, be clearly denoted. The type of the expectation makes a large difference in the value the expectation takes. The first thing to note is that the conditional mean in these models is completely uninteresting since ϵ_t is a white noise process as long as $\overline{\sigma}^2$ is finite. Thus, standard tools of time-series analysis can be used to model the conditional mean. When using daily stock returns, the mean is often chosen to be constant ($\mu_t = \mu$ for all t) and sometimes chosen to be 0.[1]

The first interesting property of the ARCH(P) model is the unconditional variance. Assuming the unconditional variance exists, $\mathsf{E}\left[\sigma_t^2\right]$ can be derived from

$$\begin{aligned}
\mathsf{E}\left[\sigma_t^2\right] &= \mathsf{E}\left[\omega + \alpha_1\epsilon_{t-1}^2 + \alpha_2\epsilon_{t-2}^2 + ... + \alpha_P\epsilon_{t-P}^2\right] \\
&= \omega + \alpha_1\mathsf{E}\left[\epsilon_{t-1}^2\right] + \alpha_2\mathsf{E}\left[\epsilon_{t-2}^2\right] + ... + \alpha_P\mathsf{E}\left[\epsilon_{t-P}^2\right] \\
&= \omega + \alpha_1\mathsf{E}\left[\sigma_{t-1}^2\right]\mathsf{E}\left[e_{t-1}^2\right] + \alpha_2\mathsf{E}\left[\sigma_{t-2}^2\right]\mathsf{E}\left[e_{t-2}^2\right] + ... + \alpha_P\mathsf{E}\left[\sigma_{t-P}^2\right]\mathsf{E}\left[e_{t-P}^2\right] \\
&= \omega + \alpha_1\mathsf{E}\left[\sigma_{t-1}^2\right] + \alpha_2\mathsf{E}\left[\sigma_{t-2}^2\right] + ... + \alpha_P\mathsf{E}\left[\sigma_{t-P}^2\right]
\end{aligned}$$

hence

$$\begin{aligned}
\mathsf{E}\left[\sigma_t^2\right] - \alpha_1\mathsf{E}\left[\sigma_{t-1}^2\right] - \alpha_2\mathsf{E}\left[\sigma_{t-2}^2\right] - ... - \alpha_P\mathsf{E}\left[\sigma_{t-P}^2\right] &= \omega \\
\mathsf{E}\left[\sigma_t^2\right](1 - \alpha_1 - \alpha_2 - ... - \alpha_P) &= \omega \\
\overline{\sigma}^2 &= \frac{\omega}{1 - \alpha_1 - \alpha_2 - ... - \alpha_P}
\end{aligned}$$

This derivation makes use of a number of properties of ARCH family models. First, $\epsilon_t^2 = \sigma_t^2 e_t^2$ which simply follows from the definition of ϵ_t and e_t. What is less obvious is that e_t and σ_t^2 are independent. However, σ_t^2 depends on ϵ_{t-1}, ϵ_{t-2}, . . . while e_t is drawn at time t and thus must be independent. With these two properties in hand, the derivation is simple as $\mathsf{E}\left[\sigma_t^2\right] = \mathsf{E}\left[\sigma_{t-P}^2\right] = \overline{\sigma}^2$ since the expectation is unconditional and assumed to exist. We also use

$$\mathsf{E}\left[\epsilon_t^2\right] = \mathsf{E}\left[e_t^2\sigma_t^2\right] = \mathsf{E}\left[\sigma_t^2\right]\mathsf{E}\left[\epsilon_t^2\right] = \mathsf{E}\left[\sigma_t^2\right] \cdot 1 = \mathsf{E}\left[\sigma_t^2\right] = \overline{\sigma}^2$$

Inspection of the final line in the derivation reveals the condition required for the expectation for be finite: $1 - \alpha_1 - \alpha_2 - ... - \alpha_P > 0$. As was the case in an AR model, as the persistence (as measured by α_1, α_2, . . .) increases towards a unit root, the unconditional variance explodes.

Stationarity

An ARCH(P) model is covariance stationary as long as the mean model corresponds to a stationary model and $1 - \alpha_1 - \alpha_2 - ... - \alpha_P > 0$.

[1] It is not plausible for the unconditional mean to be zero. However, when using daily equity data, the mean squared is typically less than 1% of the variance and there are few ramifications for setting this value to 0.

Since the mean is an ARCH model is uninteresting and including more lags in the variance complicates the analysis without providing any deeper insight, consider a simple ARCH(1) model

$$\begin{aligned} r_t &= \epsilon_t \\ \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 \\ \epsilon_t &= \sigma_t e_t \\ e_t &\sim \mathsf{WN}(0,1) \end{aligned}$$

The next step is to derive the time-series properties of the squared innovation, $\left\{\epsilon_t^2\right\}$. While the ARCH(1) model is superficially different from any studied thus far, it is actually an AR(1) for ϵ_t^2 in disguise. By adding $\epsilon_t^2 - \sigma_t^2$ to both sides of the volatility equation,

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 \\ \sigma_t^2 + \epsilon_t^2 - \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \epsilon_t^2 - \sigma_t^2 \\ \epsilon_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \epsilon_t^2 - \sigma_t^2 \\ &= \omega + \alpha\epsilon_{t-1}^2 + \sigma_t^2\left(e_t^2 - 1\right) \\ &= \omega + \alpha\epsilon_{t-1}^2 + \nu_t \end{aligned}$$

the ARCH model can be transformed into an AR(1). The error term, ν_t is the volatility surprise, $\epsilon_t^2 - \sigma_t^2$ which can be rewritten $\sigma_t^2\left(e_t^2 - 1\right)$ and is a mean 0 white noise process since $\mathsf{E}\left[e_t^2 - 1\right] = 0$. Thus, the autocovariances of ϵ_t^2 are simple to derive. First note that $\epsilon_t^2 - \overline{\sigma}^2 = \sum_{i=0}^{\infty}\alpha^i\nu_{t-i}$. The autocovariance can be expressed

$$\begin{aligned} \mathsf{E}\left[\left(\epsilon_t^2 - \overline{\sigma}^2\right)\left(\epsilon_{t-1}^2 - \overline{\sigma}^2\right)\right] &= \mathsf{E}\left[\left(\alpha\left(\epsilon_t^2 - \overline{\sigma}^2\right) + \nu_t\right)\left(\epsilon_{t-1}^2 - \overline{\sigma}^2\right)\right] \\ &= \alpha\mathsf{V}\left[\epsilon_t^2 - \overline{\sigma}^2\right] + \mathsf{Cov}\left[\nu_t, \epsilon_{t-1}^2 - \overline{\sigma}^2\right] \\ &= \alpha\mathsf{V}\left[\epsilon_t^2\right] \end{aligned}$$

where $\mathsf{V}\left[\epsilon_t^2\right]$ is the variance of the squared innovations.[2] By repeated substitution,

$$\mathsf{E}\left[\left(\epsilon_t^2 - \overline{\sigma}^2\right)\left(\epsilon_{t-k}^2 - \overline{\sigma}^2\right)\right] = \alpha^k\mathsf{V}\left[\epsilon_t^2\right].$$

You should note that this is exactly the autocovariance you would get form an AR(1).

Autocorrelations

From the autocovariances, the autocorrelations can be shown to be

$$\mathsf{Corr}\left[\epsilon_t^2, \epsilon_{t-k}^2\right] = \frac{\alpha^k\mathsf{V}\left[\epsilon_t^2\right]}{\mathsf{V}\left[\epsilon_t^2\right]} = \alpha^k$$

[2] For now, assume this is finite.

Further, the relationship between the kth autocorrelation of an ARCH(1) and an AR(1) holds for an ARCH(P) process. The autocorrelations are identical to those of an AR(P) process with parameters $\alpha_1, \alpha_2, ..., \alpha_p$.

One interesting aspect of ARCH(P) processes (and any covariance stationary ARCH-family model) is that the autocorrelations must be positive. It is also necessary that $\omega > 0$ to ensure covariance stationarity. For the same reason that the parameters of the ARCH(P) must be positive, the autocorrelations must be positive. If one autocorrelation were negative, eventually a shock would be sufficiently large to force the conditional variance negative resulting in an ill-defined model.

Kurtosis

The second interesting property of the ARCH models is that the kurtosis of shocks is strictly greater than the kurtosis of a normal. This may seem strange; all of the shocks ϵ_t are normal by assumption. However, an ARCH model is a *variance-mixture* of normals which must produce a kurtosis greater than three. An intuitive proof is simple,

$$\kappa = \frac{\mathsf{E}\left[\epsilon_t^4\right]}{\mathsf{E}\left[\epsilon_t^2\right]^2} = \frac{\mathsf{E}\left[\mathsf{E}_{t-1}\left[\epsilon_t^4\right]\right]}{\mathsf{E}\left[\mathsf{E}_{t-1}\left[\sigma_t^2 e_t^2\right]\right]^2} = \frac{\mathsf{E}\left[\mathsf{E}_{t-1}\left[e_t^4\right]\sigma_t^4\right]}{\mathsf{E}\left[\mathsf{E}_{t-1}\left[e_t^2\right]\sigma_t^2\right]^2} = \frac{\mathsf{E}\left[3\sigma_t^4\right]}{\mathsf{E}\left[\sigma_t^2\right]^2} = 3\frac{\mathsf{E}\left[\sigma_t^4\right]}{\mathsf{E}\left[\sigma_t^2\right]^2} > 3$$

The key steps in this derivation are that $\epsilon_t^4 = e_t^4\sigma_t^4$ and that $\mathsf{E}_{t-1}\left[e_t^4\right] = 3$ since it is a standard normal. The final conclusion that $\mathsf{E}\left[\sigma_t^4\right]/\mathsf{E}\left[\sigma_t^2\right]^2 > 1$ follows from the Cauchy-Schwartz inequality.[3] Alternatively, this can be deduced decomposing $\mathsf{E}\left[\sigma_t^4\right]$ into $\mathsf{V}\left[\sigma_t^2\right] + \mathsf{E}\left[\sigma_t^2\right]$. The formal derivation of the kurtosis is tedious, but the kurtosis, κ, of an ARCH(1) can be shown to be

$$\kappa = \frac{3\left(1-\alpha^2\right)}{1-3\alpha^2} > 3$$

8.2.2 The GARCH model

The ARCH model has been deemed a sufficient contribution to economics to warrant a Nobel prize for its inventor. Unfortunately, like most models in econometrics, it has problems. ARCH models typically require 5-8 lags of the squared shock to adequately model conditional variance. Generalized ARCH (GARCH) processes improve on ARCH models by including a *smoothing* term to produce a substantially more parsimonious specification. First, consider the complete

[3] If you are not convinced, try inverting the fraction and see if you can convince yourself that the inverted fraction must be less than one. Once you decompose the $\mathsf{E}\left[\epsilon_t^2\right]^2$ into $\mathsf{E}\left[\epsilon_t^2\right]\mathsf{E}\left[\epsilon_t^2\right]$, the inverted fraction should look like a correlation: $\mathsf{E}\left[\epsilon_t^2\right]^2 \leq \mathsf{E}\left[\epsilon_t^2\right]\mathsf{E}\left[\epsilon_t^2\right]$ since $\mathsf{E}\left[XY\right]^2 \leq \mathsf{E}\left[X^2\right]\mathsf{E}\left[Y^2\right]$.

GARCH(P,Q) process,

$$\begin{aligned}
r_t &= \mu_t + \epsilon_t \\
\mu_t &= \phi_0 + \phi_1 r_{t-1} + ... + \phi_S r_{t-S} \\
\sigma_t^2 &= \omega + \sum_{p=1}^{P} \alpha_p \epsilon_{t-p}^2 + \sum_{q=1}^{Q} \beta_q \sigma_{t-q}^2 \\
\epsilon_t &= \sigma_t e_t \\
e_t &\sim \mathsf{WN}(0,1)
\end{aligned}$$

which builds on an ARCH(P) model by including Q lags of the conditional variance, $\sigma_{t-1}^2, ..., \sigma_{t-Q}^2$. Rather than focusing on the general specification with all of its complications, consider a simpler GARCH(1,1) model where the conditional mean is assumed to be zero,

$$\begin{aligned}
r_t &= \epsilon_t \\
\sigma_t^2 &= \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\
\epsilon_t &= \sigma_t e_t \\
e_t &\sim \mathsf{WN}(0,1)
\end{aligned}$$

In this specification, tomorrow's variance will be an average of today's shock, ϵ_{t-1}^2 and today's variance σ_{t-1}^2 plus a constant. The effect of including today's variance is to produce a model which is actually and ARCH(∞) in disguise. Begin by backward substituting,

$$\begin{aligned}
\sigma_t^2 &= \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\
&= \omega + \alpha \epsilon_{t-1}^2 + \beta \left(\omega + \alpha \epsilon_{t-2}^2 + \beta \sigma_{t-2}^2 \right) \\
&= ... = \omega \sum_{i=0}^{\infty} \beta^i + \sum_{i=0}^{\infty} \beta^i \alpha \epsilon_{t-i} \\
&= \frac{\omega}{1-\beta} + \alpha \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}
\end{aligned}$$

and the ARCH(∞) representation can be derived. From this form, it can be determined that the conditional variance in period t is a constant, $\omega / (1-\beta)$, and a weighted average of past squared innovations with weights α, $\alpha\beta$, $\alpha\beta^2$...

As was the case in the ARCH(P) model, the coefficients of a GARCH model must also be restricted to ensure positive conditional variances. In a GARCH(1,1), these restrictions are $\omega > 0, \alpha \geq 0$ and $\beta \geq 0$. In a GARCH(P,1) model the restriction change to $\alpha_p \geq 0$, $p = 1, 2, ..., P$. However, in a complete GARCH(P,Q) model the parameter restriction are difficult to derive. For

example, in a GARCH(2,2), one of the two $\beta's$ (β_2) can be negative while ensuring that all conditional variances are positive.

As was the case in the ARCH(1) model, the GARCH(1,1) model can be transformed into a standard time series model for ϵ_t^2, by adding $\epsilon_t^2 - \sigma_t^2$ to both sides

$$\begin{aligned}
\sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \\
\sigma_t^2 + \epsilon_t^2 - \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \epsilon_t^2 - \sigma_t^2 \\
&= \omega + (\alpha+\beta)\epsilon_{t-1}^2 - \beta\left(\epsilon_{t-1}^2 - \sigma_{t-1}^2\right) + \epsilon_t^2 - \sigma_t^2 \\
&= \omega + (\alpha+\beta)\epsilon_{t-1}^2 + \nu_t - \beta\nu_{t-1}
\end{aligned}$$

However, unlike the ARCH(1) process which can be transformed into an AR(1), the GARCH(1,1) is transformed into an ARMA(1,1) where $\nu_t = \epsilon_t^2 - \sigma_t^2$ is the volatility surprise. In the general GARCH(P,Q), the ARMA representation takes the form of an ARMA(max(P,Q),Q).

$$\epsilon_t^2 = \omega + \sum_{i=1}^{\max(P,Q)} (\alpha_i + \beta_i)\epsilon_{t-i}^2 - \sum_{q=1}^{Q} \beta_q \nu_{t-q} + \nu_t$$

Using the same derivation in the ARCH(1) model, the unconditional variance can be shown to be

$$\begin{aligned}
\mathsf{E}\left[\sigma_t^2\right] &= \omega + \alpha\mathsf{E}\left[\epsilon_{t-1}^2\right] + \beta\mathsf{E}\left[\sigma_{t-1}^2\right] \\
\overline{\sigma}^2 &= \omega + \alpha\overline{\sigma}^2 + \beta\overline{\sigma}^2 \\
\overline{\sigma}^2 &= \frac{\omega}{1-\alpha-\beta}
\end{aligned}$$

Inspection of the ARMA model, leads to an alternative derivation of $\overline{\sigma}^2$ since the AR(1) coefficient is $\alpha+\beta$ and the intercept is ω, the unconditional mean is intercept divided by one minus the AR(1) coefficient. In mathematical notation this is simply $\omega/(1-\alpha-\beta)$. In a general GARCH(P,Q) the unconditional variance is

$$\overline{\sigma}^2 = \frac{\omega}{1 - \sum_{p=1}^{P}\alpha_p - \sum_{q=1}^{Q}\beta_q}$$

As was the case in the ARCH(1) model, the requirements for stationarity are that $1-\alpha-\beta \geq 0$ and $\alpha \geq 0$, $\beta \geq 0$ and $\omega > 0$.

The ARMA(1,1) form can be used directly to solve for the autocovariances. Recall the definition of a mean zero ARMA(1,1),

$$y_t = \phi y_{t-1} + \eta_t + \theta\eta_{t-1}$$

and the 1st autocovariance can be computed

$$\begin{aligned}
\mathsf{E}\left[y_t y_{t-1}\right] &= \mathsf{E}\left[\left(\phi y_{t-1} + \eta_t + \theta\eta_{t-1}\right) y_{t-1}\right] \\
&= \phi\mathsf{E}\left[y_{t-1}^2\right] + \theta\mathsf{E}\left[\eta_{t-1}y_{t-1}\right] \\
&= \phi\mathsf{V}\left[y_{t-1}\right] + \theta\mathsf{E}\left[\eta_{t-1}^2\right] \\
\gamma_1 &= \phi\gamma_0 + \theta\sigma_\eta^2
\end{aligned}$$

and the jth autocovariance is $\gamma_j = \phi^{j-1}\gamma_1$. Returning to the GARCH(1,1) model, $\phi = \alpha+\beta, \theta = -\beta$, y_{t-1} is ϵ_{t-1}^2, and η_{t-1} is $\sigma_{t-1}^2 - \epsilon_{t-1}^2$. Thus $\mathsf{V}\left[\epsilon_{t-1}^2\right]$ and $\mathsf{V}\left[\sigma_{t-1}^2 - \epsilon_{t-1}^2\right]$ must be solved for. However, this is tedious. The key to understanding the autocovariance (and autocorrelation) of a GARCH is to use the ARMA mapping. First note that $\mathsf{E}\left[\sigma_{t-1}^2 - \epsilon_{t-1}^2\right] = 0$ so $\mathsf{V}\left[\sigma_{t-1}^2 - \epsilon_{t-1}^2\right]$ is simply $\mathsf{E}\left[\left(\sigma_{t-1}^2 - \epsilon_{t-1}^2\right)^2\right]$. This can be expanded to $\mathsf{E}\left[\sigma_{t-1}^4\right] - 2\mathsf{E}\left[\sigma_{t-1}^2\epsilon_{t-1}^2\right] + \mathsf{E}\left[\epsilon_{t-1}^4\right]$ which can be shown to be $(1-2+3)\,\mathsf{E}\left[\sigma_{t-1}^4\right]$. The only remaining step is to complete the tedious derivation of the expectation of these fourth powers,

$$\begin{aligned}
\mathsf{E}\left[\epsilon_{t-1}^4\right] &= \mathsf{E}\left[\mathsf{E}_t\left[\epsilon_{t-1}^4\right]\right] = \mathsf{E}\left[3\sigma_t^4\right] \\
\mathsf{E}\left[\sigma_t^4\right] &= \mathsf{E}\left[\left(\omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2\right)^2\right] \\
&= \mathsf{E}\left[\omega^2 + 2\omega\alpha\epsilon_{t-1}^2 + 2\omega\beta\sigma_{t-1}^2 + 2\alpha\beta\epsilon_{t-1}^2\sigma_{t-1}^2 + \alpha^2\epsilon_{t-1}^4 + \beta^2\sigma_{t-1}^4\right] \\
&= \omega^2 + 2\omega\alpha\mathsf{E}\left[\epsilon_{t-1}^2\right] + 2\omega\beta\mathsf{E}\left[\sigma_{t-1}^2\right] + 2\alpha\beta\mathsf{E}\left[\epsilon_{t-1}^2\sigma_{t-1}^2\right] + \alpha^2\mathsf{E}\left[\epsilon_{t-1}^4\right] + \beta^2\mathsf{E}\left[\sigma_{t-1}^4\right]
\end{aligned}$$

Noting that $\mathsf{E}\left[\epsilon_{t-1}^2\right] = \mathsf{E}\left[\sigma_{t-1}^2\right] = \overline{\sigma}^2$, $\mathsf{E}\left[\epsilon_{t-1}^2\sigma_{t-1}^2\right] = \mathsf{E}\left[e_{t-1}^2\sigma_{t-1}^4\right] = \mathsf{E}\left[\sigma_{t-1}^4\right]$ and that $\mathsf{E}\left[\epsilon_{t-1}^4\right] = \mathsf{E}\left[\mathsf{E}_{t-2}\left[\epsilon_{t-1}^4\right]\right] = \mathsf{E}\left[\mathsf{E}_{t-2}\left[e_{t-1}^4\sigma_{t-1}^4\right]\right] = 3\mathsf{E}\left[\sigma_{t-1}^4\right]$,

$$\mathsf{E}\left[\sigma_t^4\right] = \omega^2 + 2\omega\alpha\overline{\sigma}^2 + 2\omega\beta\overline{\sigma}^2 + 2\alpha\beta\mathsf{E}\left[\sigma_{t-1}^4\right] + 3\alpha^2\mathsf{E}\left[\sigma_{t-1}^4\right] + \beta^2\mathsf{E}\left[\sigma_{t-1}^4\right]$$

$\mathsf{E}\left[\sigma_t^4\right]$ can be solved for (replacing $\mathsf{E}\left[\sigma_t^4\right]$ with $\overline{\sigma}_4$),

$$\overline{\sigma}_4 = \frac{\omega^2 + 2\omega\left(\alpha+\beta\right)\overline{\sigma}^2}{1 - 2\alpha\beta - 3\alpha^2 - \beta^2}$$

finally substituting in for $\overline{\sigma}^2 = \omega/\left(1-\alpha-\beta\right)$ and

$$\begin{aligned}
\overline{\sigma}_4 &= \frac{\omega^2\left(1+\alpha+\beta\right)}{\left(1-\alpha-\beta\right)\left(1-2\alpha\beta-3\alpha^2-\beta^2\right)} \\
\mathsf{E}\left[\epsilon_t^4\right] &= \frac{3\omega^2\left(1+\alpha+\beta\right)}{\left(1-\alpha-\beta\right)\left(1-2\alpha\beta-3\alpha^2-\beta^2\right)}
\end{aligned}$$

Kurtosis

The kurtosis can be shown to be

$$\kappa = \frac{3\left(1+\alpha+\beta\right)\left(1-\alpha-\beta\right)}{\left(1-2\alpha\beta-3\alpha^2-\beta^2\right)} > 3$$

Once again, the kurtosis is greater than that of a normal despite all of the innovation having a normal distribution.

8.2.3 The EGARCH model

The Exponential GARCH model represents a major shift from the ARCH and GARCH models. Rather than model the variance directly, EGARCH models the natural logarithm of the variance.

One advantage of this choice is obvious: there are no restrictions on any parameters needed to ensure that variance is positive. The evolution of a general EGARCH(P,O,Q) model is given by

$$\begin{aligned} r_t &= \mu_t + \epsilon_t \\ \mu_t &= \phi_0 + \phi_1 r_{t-1} + ... + \phi_S r_{t-S} \\ \ln\left(\sigma_t^2\right) &= \omega + \sum_{p=1}^{P} \alpha_p \left(|e_{t-p}| - \sqrt{\frac{2}{\pi}} \right) + \sum_{o=1}^{O} \gamma_o e_{t-o} + \sum_{q=1}^{Q} \beta_q \ln\left(\sigma_{t-q}^2\right) \\ \epsilon_t &= \sigma_t e_t \\ e_t &\sim \mathsf{WN}\left(0,1\right) \end{aligned}$$

Again, rather than working with the complicated form, consider a simpler version, an EGARCH(1,1,1) with constant mean,

$$\begin{aligned} r_t &= \epsilon_t \\ \ln\left(\sigma_t^2\right) &= \omega + \alpha \left(|e_{t-1}| - \sqrt{\frac{2}{\pi}} \right) + \gamma e_{t-1} + \beta \ln\left(\sigma_{t-1}^2\right) \\ \epsilon_t &= \sigma_t e_t \\ e_t &\sim \mathsf{WN}\left(0,1\right) \end{aligned}$$

which shows that log variance is a constant plus three terms. The first term, $|e_{t-1}| - \sqrt{\frac{2}{\pi}}$, appears strange. However this is just the absolute value of a normal random variable e_{t-1} minus its expectation, $\sqrt{2/\pi}$. Thus, this term has mean zero. The second term is e_{t-1} which is also a mean zero variable while the last term is lagged log variance. The two shocks (the e_{t-1} terms behave differently. The first causes a symmetric rise in the log variance while the second is asymmetric. Whenever the sign of e_t and γ agree, the volatility rises.

When the signs disagree, the volatility is reduced. Thus, the EGARCH(1,1,1) model is an AR model for log variance with two mean-zero shocks. More importantly, the EGARCH model often provides a superior fit to a standard GARCH model.

The S&P 500 and IBM

The use of GARCH models will be demonstrated using daily returns on both the S&P 500 and IBM from January 1, 1993 until December 31, 2003. All data were taken from WRDS. 100 times the returns will be used throughout. The returns are in figure 8.1, the squared returns are in figure 8.2 and the absolute value of returns is in figure 8.3. Summary statistics are in table 8.1 and estimated from the three models presented thus far are in table 8.2. The summary statistics are typical of daily returns. The mean is positive and the annualized volatility is between 15 and 40% per year. Neither return series is extremely skewed but both are leptokurtic (fat-tailed).

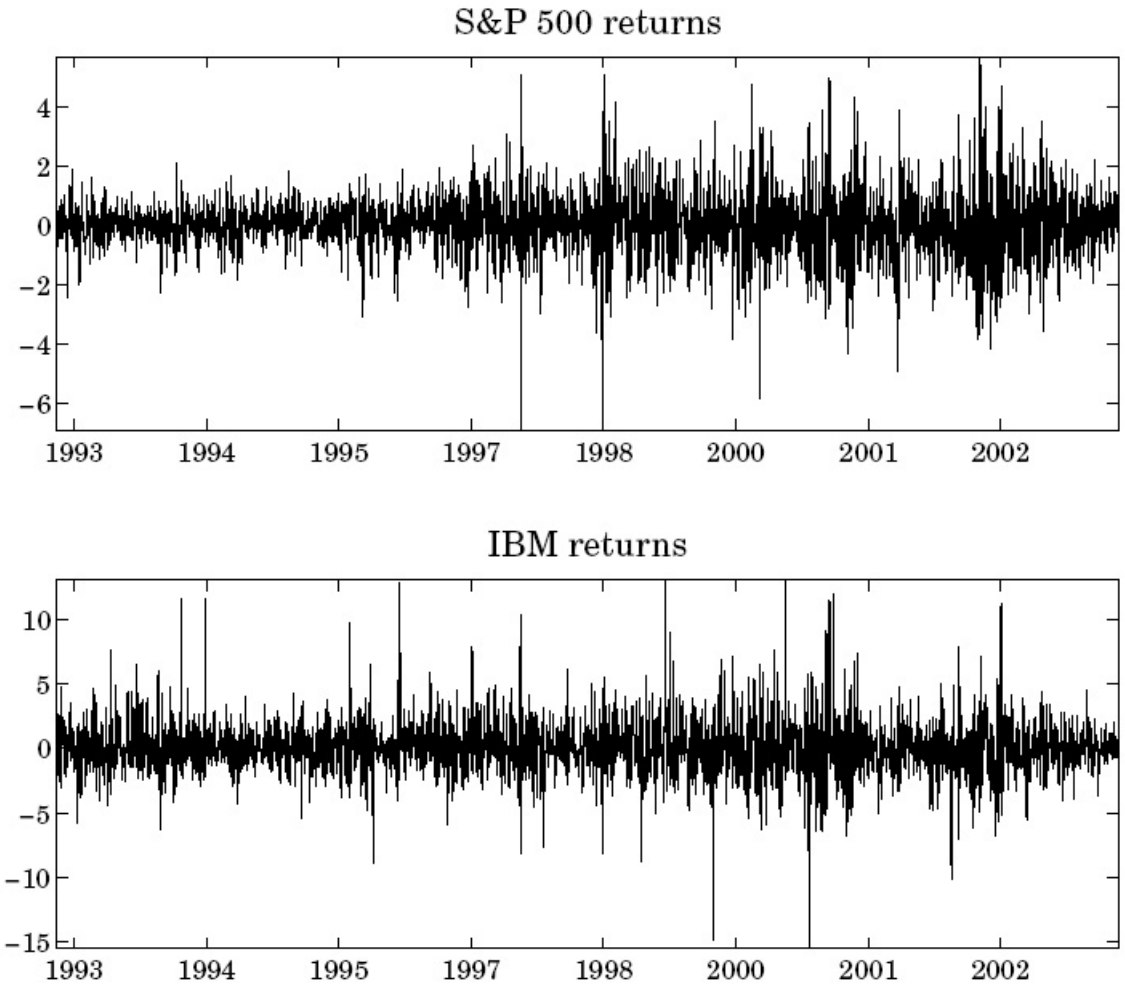

Figure 8.1: Plots of the returns on the S&P 500 and IBM from 1993 until 2003. The bulges in the return plots are graphical evidence of time-varying volatility.

	S&P 500	IBM
Ann. Mean	10.04	25.36
Ann. Volatility	17.48	35.18
Skewness	-0.02	0.35
Kurtosis	6.45	8.30

Table 8.1: Summary statistics for the S&P 500 and IBM.

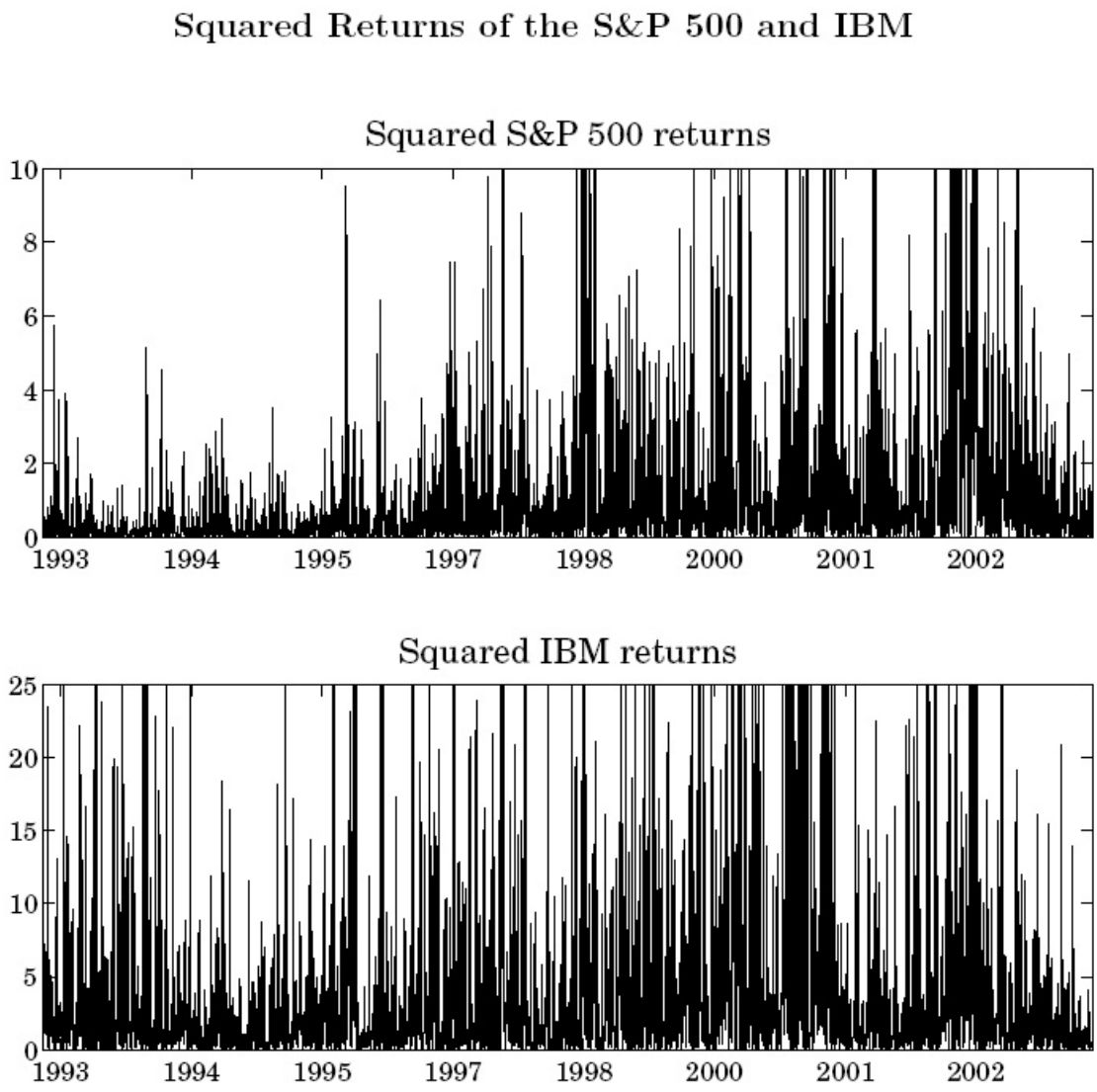

Figure 8.2: Plots of the squared returns of the S&P 500 Index and IBM. Time-variation in the squared returns is evidence of ARCH.

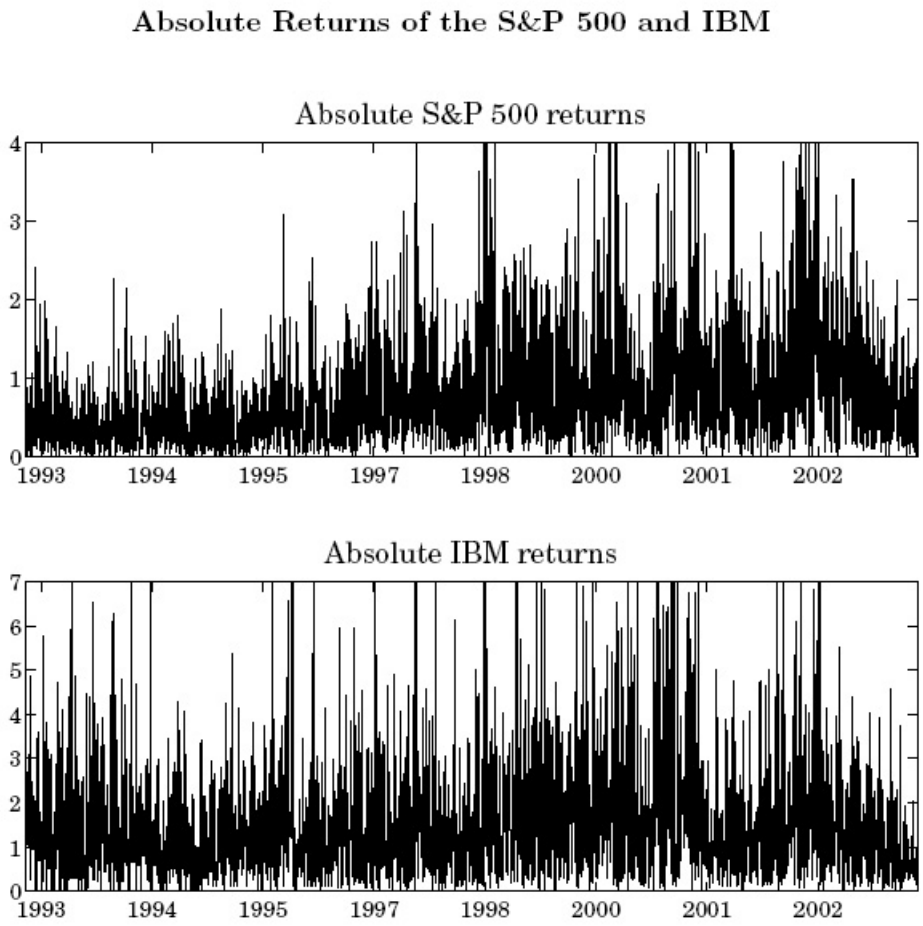

Figure 8.3: Plots of the absolute value of returns for the S&P 500 Index and IBM. Plots of the absolute value are often more useful in detecting ARCH as they are less noisy than squared returns yet still reveal dynamics in variance.

ARCH(5)

	ω	α_1	α_2	α_3	α_4	α_5	LL
Coeff.	0.309	0.093	0.213	0.161	0.201	0.145	-3890
P-val	(.000)	(.002)	(.000)	(.000)	(.000)	(.000)	

GARCH(1,1)

	ω	α	β	LL
Coeff.	0.005	0.067	0.930	-3805
P-val	(.053)	(.000)	(.000)	

EGARCH(1,1,1)

	ω	α	γ	β	LL
Coeff.	0.005	0.133	-0.109	0.978	-3746
P-val	(.052)	(.000)	(.000)	(.000)	

IBM

ARCH(5)

	ω	α_1	α_2	α_3	α_4	α_5	LL
Coeff.	2.963	0.126	0.095	0.020	0.097	0.097	-6066
P-val	(.000)	(.011)	(.042)	(.576)	(.063)	(.147)	

GARCH(1,1)

	ω	α	β	LL
Coeff.	0.130	0.066	0.909	-6024
P-val	(.125)	(.016)	(.000)	

EGARCH(1,1,1)

	ω	α	γ	β	LL
Coeff.	0.037	0.115	-0.077	0.980	-5955
P-val	(.001)	(.000)	(.000)	(.000)	

Table 8.2: Simple estimates from an ARCH(5), a GARCH(1,1) and an EGARCH(1,1,1) for the S&P 500 and IBM. The log-likelihood indicates the EGARCH model is preferred in both cases. p-values are in parentheses.

Table 8.2 contains estimates from an ARCH(5), a GARCH(1,1) and an EGARCH(1,1,1) model. All estimates were computed using maximum likelihood assuming the innovations were conditionally normal. Examining the table, there is strong evidence of time varying variance in these models indicated by numerous p-values near 0. The highest log-likelihood (a measure of fit) is produced by the EGARCH model in both series. This is likely due to the EGARCH's inclusion of asymmetries, a feature excluded from both the ARCH and GARCH models.

8.2.4 Alternative Specifications

Many specification have been introduced to capture the dynamics of volatility. This section outlines three of the most important.

GRJ-GARCH

The GJR-GARCH model was named after the authors who first described it, Glosten, Jagannathan and Runkle. It extends the GARCH to include asymmetric terms. These asymmetric terms capture an important phenomena in the conditional variance of equities: the propensity for the volatility to rise more subsequent to large negative shocks than to large positive shocks. This is an example of a GJR-GARCH(1,1,1) model.

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \gamma\epsilon_{t-1}^2 I_{\{\epsilon_{t-1}<0\}} + \beta\sigma_{t-1}^2, \quad \alpha + \gamma \geq 0$$

where $I_{\{\epsilon_{t-1}<0\}}$ is an indicator variable which takes the value 1 $I_{\{\epsilon_{t-1}<0\}}$. And more general GJR-GARCH(P,O,Q) models can be described in a natural way (Note: O always refers to the asymmetric terms in these notes).

AVARCH/TARCH/ZARCH

The Threshold ARCH (TARCH) model (also known as AVGARCH and ZARCH) makes one fundamental change to the GJR-GARCH model. Rather than model the variance directly using squared innovations, a TARCH model parameterizes the *conditional standard deviation* as a function of lagged absolute value of the shocks. It also captures asymmetries using an asymmetric term similar to that of the GJR-GARCH model. Below is an example of a TARCH(1,1,1) model.

$$\sigma_t = \omega + \alpha\left|\epsilon_{t-1}\right| + \gamma\left|\epsilon_{t-1}\right| I_{\{\epsilon_{t-1}<0\}} + \beta\sigma_{t-1}, \quad \alpha + \gamma \geq 0$$

APARCH

The third model extends the TARCH model to directly parameterize the form of the evolution in the variance. While a GJR-GARCH model uses 2 and a TARCH model uses 1, the Asymmetric Power ARCH (APARCH) parameterizes this values as an unknown (λ). This form provides

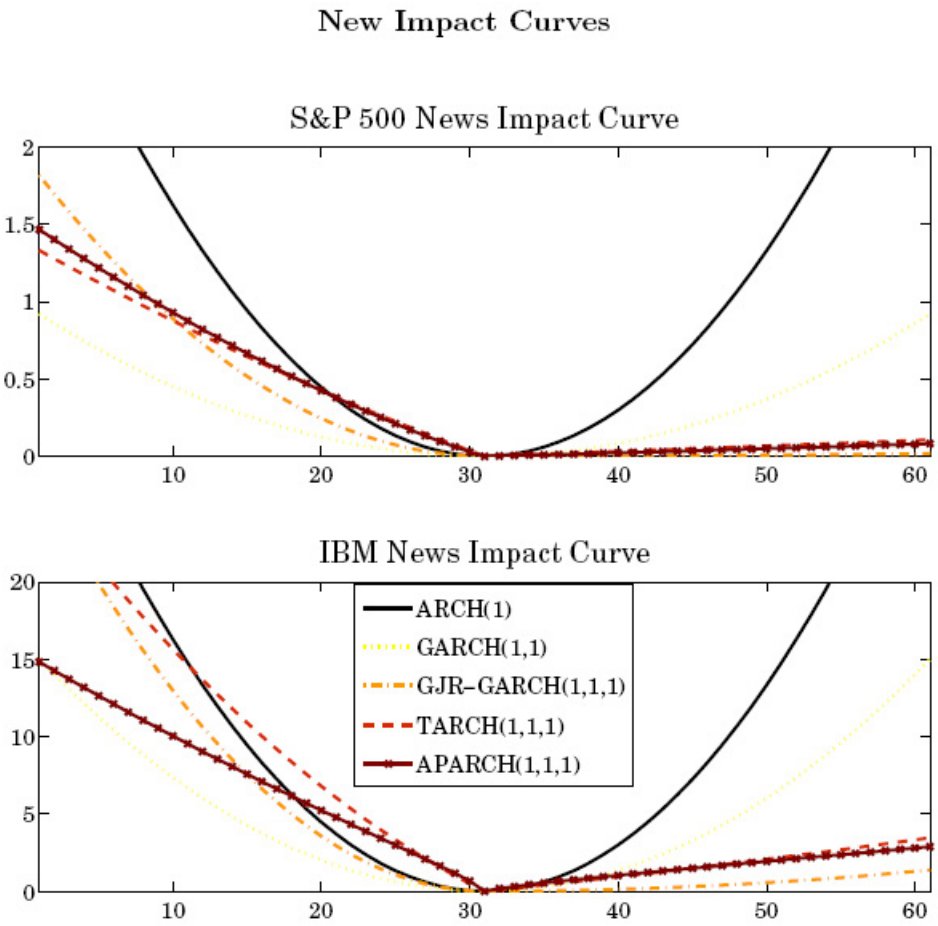

Figure 8.4: News impact curves for both the S&P 500 and IBM returns. While the ARCH and GARCH curves are symmetric, the other show substantial asymmetries to positive and negative news. Additionally, the APARCH model chose a parameterization with $\widehat{\lambda} \approx 1$ and its NIC appears similar to that of a TARCH(1,1,1).

greater flexibility in modeling the memory of volatility while remaining parsimonious. Below is an example of an APARCH(1,1,1) model.[4]

$$\sigma_t^\lambda = \omega + \alpha \left|\epsilon_{t-1}\right|^\lambda + \gamma \left|\epsilon_{t-1}\right|^\lambda I_{\{\epsilon_{t-1}<0\}} + \beta\sigma_{t-1}^\lambda, \quad \alpha + \gamma \geq 0$$

The APARCH is also interesting since it nests ARCH(P), GARCH(P,Q), GJR-GARCH(P,O,Q) and TARCH(P,O,Q) models as special cases.

8.2.5 The News Impact Curve

With a wide range of volatility models, it can be difficult to determine the precise effect of a shock to the conditional variance. Much like how the impulse response solved similar issues in VARs,

[4]This form differs slightly from what Eviews estimates, but the intuition is identical.

the news impact curve solves this problem in ARCH models. The news impact curve is nearly self descriptive. It measures the effect of a shock on the next period's conditional variance. To normalize the curve, the variance in the current period is set to the unconditional variance. The following two equation define the news impact curve,

$$\begin{aligned} n\left(e_t\right) &= \sigma_{t+1}^2\left(e_t|\sigma_t^2=\overline{\sigma}^2\right) \\ \mathsf{NIC}\left(e_t\right) &= n\left(e_t\right)-n\left(0\right) \end{aligned}$$

News impact curve for ARCH and GARCH models are simply the terms which involve ϵ.

GARCH(1,1)

$$\begin{aligned} n\left(e_t\right) &= \omega+\alpha\overline{\sigma}^2e_t^2+\beta\overline{\sigma}^2 \\ \mathsf{NIC}\left(e_t\right) &= \alpha\overline{\sigma}^2e_t^2 \end{aligned}$$

However, for models which are not linear in ϵ_t^2, the news impact curve can be fairly complicated.

TARCH(1,1,1)

$$\begin{aligned} \sigma_t &= \omega+\alpha\left|\epsilon_{t-1}\right|+\gamma\left|\epsilon_{t-1}\right|I_{\{\epsilon_{t-1}<0\}}+\beta\sigma_{t-1} \\ n\left(e_t\right) &= \omega^2+2\omega\left(\alpha+\gamma I_{\{\epsilon_t<0\}}\right)\left|\epsilon_t\right|+2\beta\left(\alpha+\gamma I_{\{\epsilon_t<0\}}\right)\left|\epsilon_t\right|\overline{\sigma}^2+\beta^2\overline{\sigma}^2+2\omega\beta\overline{\sigma}+\left(\alpha+\gamma I_{\{\epsilon_t<0\}}\right)^2\epsilon_t^2 \end{aligned}$$

The S&P 500 and IBM

Figure 8.4 contains plot of the news impact curves for both the S&P 500 and IBM. When the models include asymmetries, the news impact curves are very asymmetric and show a much larger response to negative shocks than to positive shocks.

8.2.6 Estimation

ARCH family models are substantially more complicated than what we have previously encountered in this course. Consider a simple GARCH(1,1) specification,

$$\begin{aligned} r_t &= \mu_t+\epsilon_t \\ \sigma_t^2 &= \omega+\alpha\epsilon_{t-1}^2+\beta\sigma_{t-1}^2 \\ \epsilon_t &= \sigma_t e_t \\ e_t &\sim \mathsf{WN}\left(0,1\right) \end{aligned}$$

Since the errors are assumed to be conditionally normal, the obvious strategy to estimate the unknown parameters ($\theta=(\omega,\alpha,\beta)$) is to use maximum likelihood. The normal likelihood is given by

$$f\left(\mathbf{r};\theta\right)=\prod_{t=1}^{T}\left(2\pi\sigma_t^2\right)^{-1/2}\exp\left(-\frac{\left(r_t-\mu_t\right)^2}{2\sigma_t^2}\right)$$

and the log-likelihood function is

$$l(\mathbf{r};\theta) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log\left(\sigma_t^2\right) - \sum_{t=1}^{T}\frac{(r_t-\mu_t)^2}{2\sigma_t^2}$$

If the mean is set to 0, the log-likelihood simplifies to

$$l(\mathbf{r};\theta) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log\left(\sigma_t^2\right) - \sum_{t=1}^{T}\frac{r_t^2}{2\sigma_t^2}$$

and is maximized by solving the first order conditions.

$$\frac{\partial l}{\partial \sigma_t^2}(\mathbf{r};\theta) = \sum_{t=1}^{T} -\frac{1}{2\sigma_t^2} + \frac{r_t^2}{2\sigma_t^4} = 0$$

which can be rewritten in an intuitive form,

$$\frac{\partial l}{\partial \sigma_t^2}(\mathbf{r};\theta) = \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\sigma_t^2}\left(\frac{r_t^2}{\sigma_t^2}-1\right)$$

which demonstrates that the parameters are chosen to make $\left(\frac{r_t^2}{\sigma_t^4}-1\right)$ as close to zero as possible. This term can be viewed as a *generalized error* in a GMM framework. These first order conditions are not complete since θ_i, not σ_t^2, are the parameters of the model and

$$\frac{\partial l}{\partial \theta_i}(\mathbf{r};\theta) = \frac{\partial l}{\partial \sigma_t^2}(\mathbf{r};\theta)\frac{\partial \sigma_t^2}{\partial \theta_i}$$

The derivatives take a form most of you won't have previously encountered; they are recursive.

$$\begin{aligned}
\frac{\partial \sigma_t^2}{\partial \omega} &= 1 + \beta\frac{\partial \sigma_{t-1}^2}{\partial \omega} \\
\frac{\partial \sigma_t^2}{\partial \alpha} &= \epsilon_{t-1}^2 + \beta\frac{\partial \sigma_{t-1}^2}{\partial \alpha} \\
\frac{\partial \sigma_t^2}{\partial \beta} &= \sigma_{t-1}^2 + \beta\frac{\partial \sigma_{t-1}^2}{\partial \beta}
\end{aligned}$$

although the recursion in the first FOC can be removed noting that

$$\frac{\partial \sigma_t^2}{\partial \omega} = 1 + \beta\frac{\partial \sigma_{t-1}^2}{\partial \omega} \approx \frac{1}{1-\beta}$$

The use of the normal likelihood has one incredibly strong justification; using it to estimate the parameters of the conditional variance produces parameter estimates which are **strongly consistent**. Strong consistency in a concept that states that the estimates converge to the true parameters even if the model assumes the wrong conditional distribution. For example, in a standard GARCH(1,1), if the errors were actually conditionally Student's T rather than normal, the parameter estimates

would still converge to their true value if estimated with the normal likelihood. The intuition behind this result comes form the generalized error:

$$\left(\frac{r_t^2}{\sigma_t^2} - 1\right)$$

Whenever $\sigma_t^2 = \mathsf{E}_{t-1}\left[r_t^2\right]$, then

$$\mathsf{E}\left[\left(\frac{r_t^2}{\sigma_t^2} - 1\right)\right] = \mathsf{E}\left[\left(\frac{\mathsf{E}_{t-1}\left[r_t^2\right]}{\sigma_t^2} - 1\right)\right] = \mathsf{E}\left[\left(\frac{\sigma_t^2}{\sigma_t^2} - 1\right)\right] = 0$$

Thus, as long as the GARCH model contains the correct specification, the parameters will be chosen to make the conditional expectation 0; these parameters correspond to those of the original DGP even if the conditional distribution is not normal. This is a very special property of the normal distribution and is not found in any other commonly used distribution. For example, if the conditional distribution was actually

$$e_t \overset{iid}{\sim} \frac{\chi^2 - 1}{\sqrt{2}}$$

and a Student's T were specified, the parameters would not be strongly consistent (because the likelihood would not be based on the normal distribution and the generalized errors would not appear).

8.2.7 Inference

Once the parameters have been estimated, parameter inference is next order of business. In maximum likelihood estimation problems (MLE) the estimated parameters can be shown to be asymptotically normally distributed,

$$\sqrt{T}\left(\theta - \theta_0\right) \rightarrow \mathsf{N}\left(\mathbf{0}, \mathbf{A}^{-1}\right)$$

where

$$\mathbf{A} = -\mathsf{E}\left[\frac{\partial^2 l}{\partial\theta\partial\theta'}\left(\mathbf{r};\theta\right)\right]$$

is the negative of the expected Hessian. Intuitively, the Hessian measures how much curvature there is in the log-likelihood at the optimum just like the second-derivative measures the rate-of-change in the rate-of-change of the function in a standard calculus problem. To estimate $\mathbf{A}$, the sample analogue is used,

$$\widehat{\mathbf{A}} = T^{-1}\sum_{t=1}^{T}\frac{\partial^2 l}{\partial\theta\partial\theta'}\left(\mathbf{r};\theta\right)$$

You may recall from your notes on linear regression that for MLE problems, the Information Matrix Equality (IME) holds and

$$\mathbf{A} = \mathbf{B}$$

where

$$\mathbf{B} = \mathsf{E}\left[\frac{\partial l}{\partial \theta}(\mathbf{r};\theta)\frac{\partial l}{\partial \theta'}(\mathbf{r};\theta)\right]$$

is the covariance of the scores. Again, this covariance measures how much information there is in the data to estimate the parameters and larger variances of the scores indicates that small parameter changes have a large impact and that the parameters are precisely estimated. Once again, the estimator of $\mathbf{B}$ is simply the sample analogue,

$$\widehat{\mathbf{B}} = T^{-1}\sum_{t=1}^{T}\frac{\partial l}{\partial \theta}(\mathbf{r};\theta)\frac{\partial l}{\partial \theta'}(\mathbf{r};\theta)$$

The conditions for the IME to hold require that the models be MLE which requires the specified conditional distribution of returns to be correct. Thus, when normal is assumed the conditional distribution of returns must be normal. When one specification is used but the data actually have a different conditional distribution, these problems are known as Quasi maximum likelihood estimation problems (QMLE) and, in general, the IME does not hold. When this is the case, the parameters are still asymptotically normal but with a different covariance,

$$\sqrt{T}(\theta - \theta_0) \rightarrow \mathsf{N}\left(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right)$$

When the IME holds, $\mathbf{A} = \mathbf{B}$ and this form simplifies to the earlier form. However, in most applications of ARCH models, the conditional distribution of shocks is not normal and exhibits excess kurtosis (kurtosis beyond that of a normal) and some skewness. An alternative method to derive the form of the covariance, $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ is to treat the estimation problem as a GMM problem on the scores. In this setup, Bollerslev and Wooldridge were the first to show that the IME doesn't generally hold for GARCH models when the distribution is misspecified and this so-called "sandwich" form $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ of the covariance estimator is often referred to as the **Bollerslev-Wooldridge** standard error (E-Views uses this expression).

The S&P 500 and IBM

To demonstrate the effect of the form of the covariance estimator, a TARCH(1,1,1) model was estimated for both the S&P 500 returns and the returns on IBM. Table 8.3 contains the estimated parameters and their p-values using both the IME VCV and the BW VCV. There is little change in the S&P model but the symmetric term in the IBM model changes form highly significant to insignificant at 5%.

The independence of the mean and variance

One important but not obvious issue when using MLE assuming conditionally normal errors or QMLE when conditional normality is wrongly assumed is that the parameters in the mean and

S&P 500

	ω	α	γ	β
Coeff.	0.016	0.012	0.115	0.929
Std. VCV	(.00)	(0.11)	(.00)	(.00)
BW VCV	(.00)	(0.10)	(.00)	(.00)

IBM

	ω	α	γ	β
Coeff.	0.036	0.021	0.082	0.936
Std. VCV	(.00)	(0.00)	(.00)	(.00)
BW VCV	(.01)	(0.08)	(.00)	(.00)

Table 8.3: Estimates from a TARCH(1,1,1) for the S&P 500 and IBM. The BW (Bollerslev-Wooldridge) VCV (variance/covariance matrix) makes a difference in the significance of the symmetric term in the IBM model.

the parameters in the variance are asymptotically independent. As a result, you can estimate the mean and variance parameters separately and their covariances will still be correct. The intuition comes form analyzing the cross-partial derivative of the log-likelihood with respect to the mean and variance parameters,

$$l\left(\mathbf{r};\theta\right)=-\frac{T}{2}\log\left(2\pi\right)-\frac{1}{2}\sum_{t=1}^{T}\log\left(\sigma_{t}^{2}\right)-\sum_{t=1}^{T}\frac{\left(r_{t}-\mu_{t}\right)^{2}}{2\sigma_{t}^{2}}$$

The first order condition is,

$$\frac{\partial l}{\partial\mu_{t}}\left(\mathbf{r};\theta\right)\frac{\partial\mu_{t}}{\partial\phi}=\sum_{t=1}^{T}\frac{\left(r_{t}-\mu_{t}\right)}{\sigma_{t}^{2}}\frac{\partial\mu_{t}}{\partial\phi}$$

and the second order condition is

$$\frac{\partial^{2}l}{\partial\mu_{t}\partial\sigma_{t}^{2}}\left(\mathbf{r};\theta\right)\frac{\partial\mu_{t}}{\partial\phi}\frac{\partial\sigma_{t}^{2}}{\partial\psi}=-2\sum_{t=1}^{T}\frac{\left(r_{t}-\mu_{t}\right)}{\sigma_{t}^{4}}\frac{\partial\mu_{t}}{\partial\phi}\frac{\partial\sigma_{t}^{2}}{\partial\psi}$$

where ϕ is a parameter of the conditional mean and ψ is a parameter of the conditional variance. For example, in a simple ARCH model with a constant mean, μ, $\phi=\mu$ and ψ can be either ω or α. Taking expectations of the cross-partial,

$$\mathsf{E}\left[\frac{\partial^{2}l}{\partial\mu_{t}\partial\sigma_{t}^{2}}\left(\mathbf{r};\theta\right)\frac{\partial\mu_{t}}{\partial\phi}\frac{\partial\sigma_{t}^{2}}{\partial\psi}\right]=0$$

it can be shown using the law of iterated expectations that the expectation is 0. The intuition behind this result is also simple: if the mean model is correct for the conditional expectation of r_t,

the term $r_t - \mu_t$ has conditional expectation 0 and the variance parameters play no role. This is a similar argument to the validity of OLS when the errors are heteroskedastic.

8.2.8 GARCH-in-Mean

The GARCH-in-mean model (GIM) makes a significant change to the role of time-varying volatility by explicitly relating the level of volatility to the expected return. A simple GIM model can be described as

$$r_t = \mu + \delta\sigma_t^2 + \epsilon_t$$

$$\begin{aligned} r_t &= \mu + \delta\sigma_t^2 + \epsilon_t \\ \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \\ \epsilon_t &= \sigma_t e_t \\ e_t &\sim \mathsf{WN}\left(0,1\right) \end{aligned}$$

While this model uses a GARCH(1,1) specification for the conditional variance, any ARCH-family model can be used. The obvious difference between the GIM and a standard GARCH(1,1) is that the variance causes the returns. However, you should notice that the shock causing the changes in variance is not the mean return but still ϵ_{t-1}^2 so the ARCH portion is fundamentally unaffected. Other forms of the GIM model have been employed where the conditional standard deviation or the log variance are used in mean equation,

$$r_t = \mu + \delta\sigma_t + \epsilon_t$$

or

$$r_t = \mu + \delta \ln\left(\sigma_t^2\right) + \epsilon_t$$

Note: Despite the previous discussion of estimating the mean and variance parameters separately, in a GIM model this is not the case since there is feedback between the variance and the size of shock to variance through he mean equation. One interesting property of the GIM model is that is will be stationary as long as the variance process is stationary. The intuition for this result follows from noting that the mean term is exactly that; it does not feed back into the variance process.

Asset pricing likes to believe that there is a risk-return trade off. GIM models provide a natural method to examine whether this is the case. Using the S&P 500 data, three GIM models were estimated (one for each for of variance in the mean equation) and the results are presented in table 8.4. Based on these estimates, there does appear to be a trade off between mean and variance and higher variances produce higher expected means.

Mean Specification	μ	δ	ω	α	β
σ_t	0.064	0.015	0.006	0.072	0.924
	(.207)	(.718)	(.000)	(0.000)	(.000)
σ_t^2	0.040	0.037	0.006	0.073	0.923)
	(.105)	(.102)	(.000)	(0.000)	(.000)
$\ln(\sigma_t)$	0.018	0.076	0.006	0.072	0.924
	(.395)	(.000)	(.000)	(0.000)	(.000)

Table 8.4: GARCH-in-mean estimated for the S&P 500 series. δ, the parameter which measures the GIM effect, is the most interesting parameter and it is significant in both the log variance specification and the variance specification. The GARCH model estimated was a standard GARCH(1,1). p-values are in parentheses.

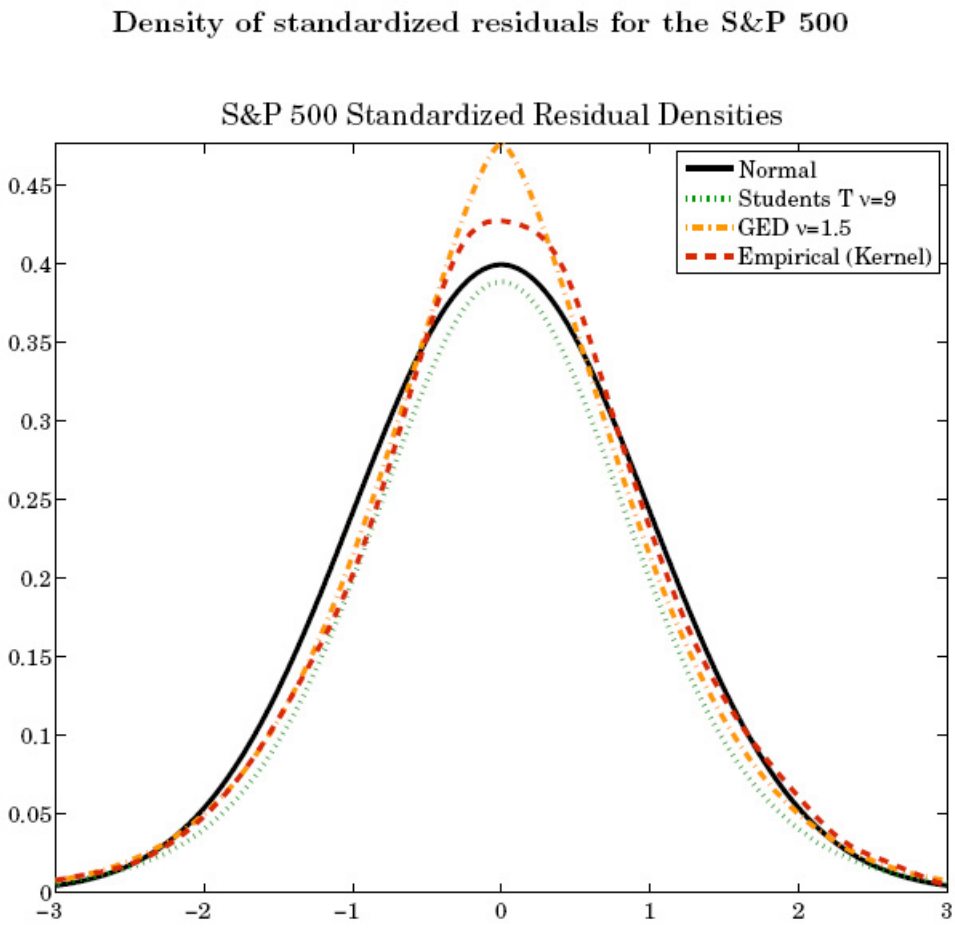

Figure 8.5: This picture contains the estimated density for the S&P 500 and the density implied by the three distributions, normal, Student's T and GED. In the latter two, the degree of freedom parameter, ν was jointly estimated with the variance parameters.

8.2.9 Alternative Distributional Assumptions

Despite the strengths of the conditional normal (estimation is easy and it produces parameters which are strongly consistent for the true parameters), empirical research in to conditional variances have made use of alternative distributions. These two most popular are the **standardized Student's T** whose density is

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\pi\left(\nu-2\right)}}\frac{1}{\sigma_t}\frac{1}{\left(1+\frac{x^2}{\sigma_t(\nu-2)}\right)^{\frac{\nu+1}{2}}}$$

which is simply the density of a t_ν which has been standardized to have unit variance; and the **Generalized Error Distribution**

$$\frac{\nu\exp\left(-\frac{1}{2}\left|\frac{x}{\sigma_t}\right|^\nu\right)}{\lambda 2^{\frac{\nu+1}{\nu}}\Gamma\left(\frac{1}{\nu}\right)}$$

$$\lambda=\left(\frac{2^{-\frac{2}{\nu}}\Gamma\left(\frac{1}{\nu}\right)}{\Gamma\left(\frac{3}{\nu}\right)}\right)^{\frac{1}{2}}$$

which nests the normal when $\nu=2$ (different ν from that of the T). Both of these distribution may be better approximations to the true distribution since they can produce conditional distributions where the kurtosis is greater than that of the conditional normal, an important empirical fact. Next chapter will return to these distributions in the context of value-at-risk and density forecasting.

The S&P 500

To quickly explore the role of alternative distribution, a TARCH(1,1,1) was fit to the S&P series using the conditional normal, the Student's T and the GED. Figure 8.5 contains the empirical density (constructed with a kernel) and the density implied by the three distributions. Note that the degree of freedom parameters, ν, was jointly estimated with the conditional variance parameters. Figure 8.6 shows plots of the estimated conditional variance from the three models. The most important aspect of this figure is that the fit variances are essentially identical. This is a common phenomena when trying alternative distribution; few make an economically meaningful difference in the fit conditional variances.

8.3 Model Building

Since ARCH and GARCH models are very similar to AR and ARMA models, the Box-Jenkins methodology is a natural way to approach the problem. The first step is to analyze the ACF and PACF of the squared returns. Figures 8.7 and 8.8 contain the ACF and PACF for the squared

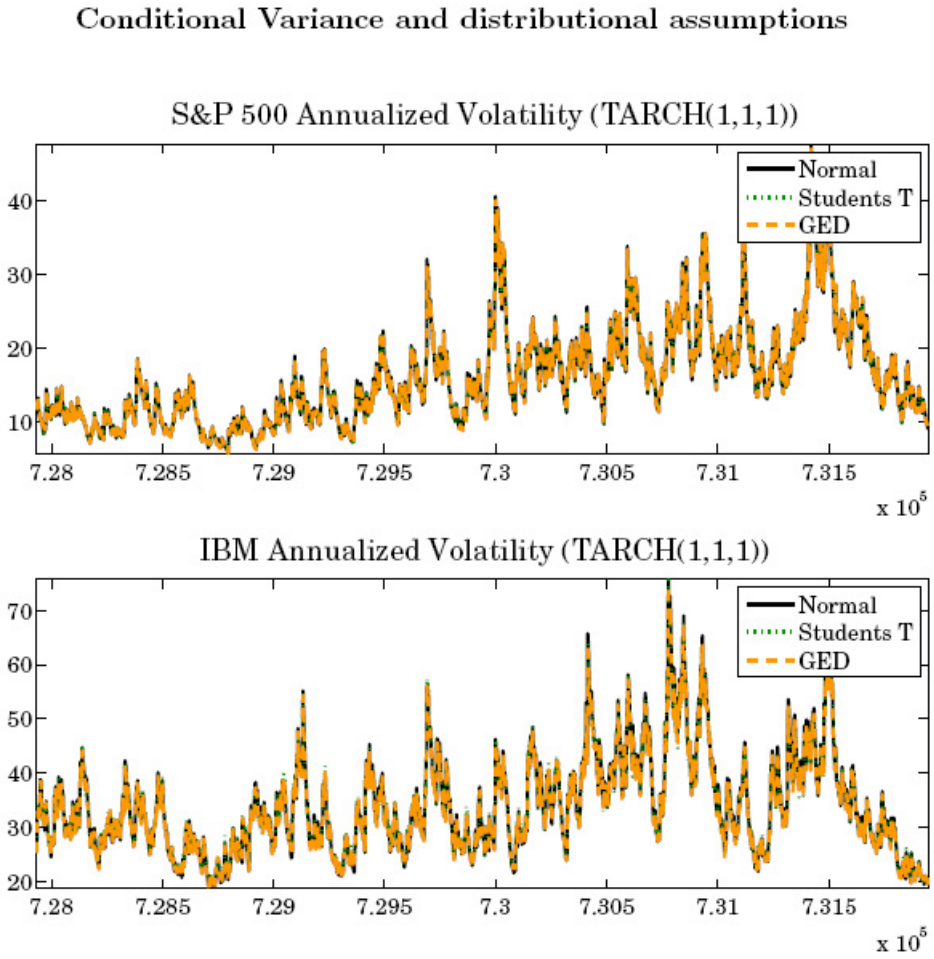

Figure 8.6: The most important aspect of using alternative distributional assumptions is that they make essentially no difference to the fit of variances or the estimated parameters.

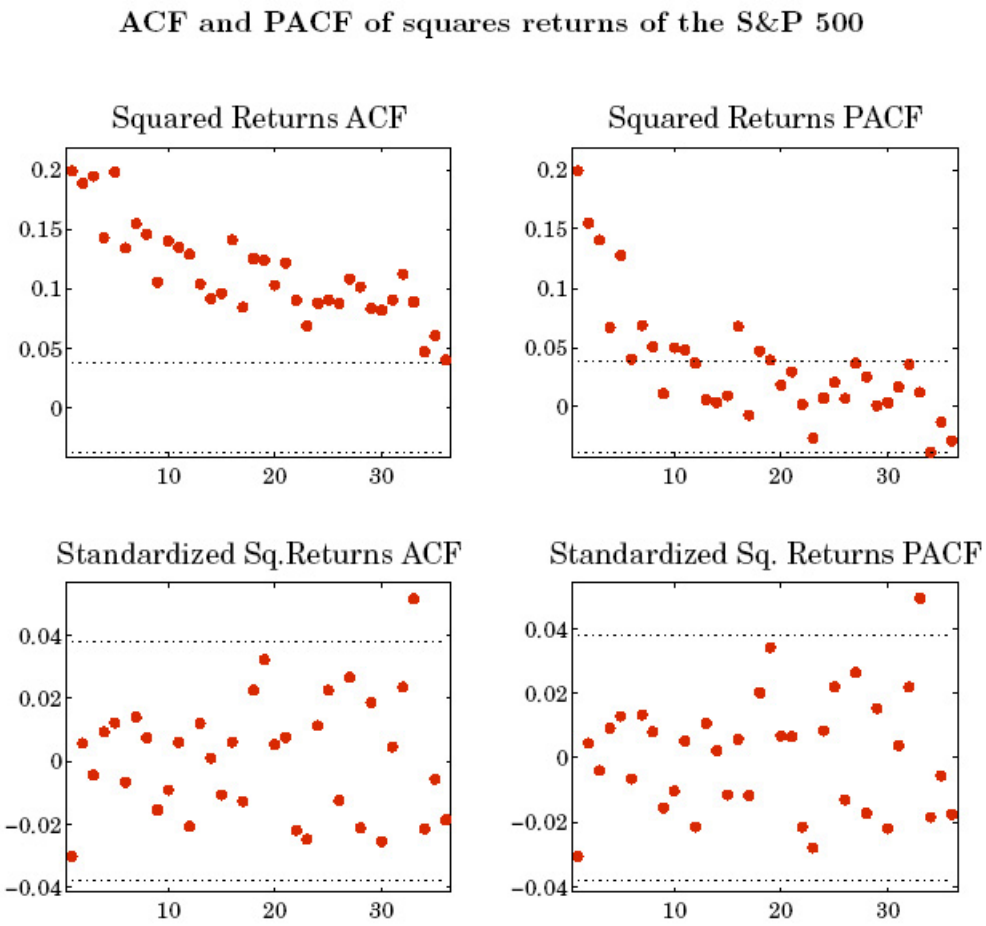

Figure 8.7: ACF and PACF of the squared returns for the S&P 500. The bottom two figures contain the ACF and PACF of $\widehat{e}_t^2 = \widehat{\epsilon}_t^2/\widehat{\sigma}_t^2$. The top figures show persistence in both the ACF and PACF which indicate an ARMA model is needed (hence a GARCH model) while the ACF and PACF of the standardized residuals is white noise.

returns of the S&P 500 and IBM respectively. The models used in selecting the final model are contained in tables 8.5 and 8.6 respectively. Both instances began with a simple GARCH(1,1). The next steps were to check is more lags were needed but fitting GARCH(2,1) and GARCH(1,2) to each. Neither of these improved the fit meaningfully and a GARCH(1,1) was assumed to be sufficient to capture the important dynamics.

The next step was to examine whether there are any asymmetries sing a GJR-GARCH(1,1,1). The asymmetry term was significant so other forms of the GJR model were explored and all found to provide little improvement in the fit. Once a GJR-GARCH(1,1,1) model was decided upon, a TARCH(1,1,1) was used to test whether evolution in variances or standard deviations was better. Both series preferred the TARCH form to the GARCH form (compare the log-likelihoods), and the TARCH(1,1,1) was selected. In comparing alternative specification, an EGARCH was fit and found to also provide a good description of the data. In both cases the EGARCH was expanded. In the S&P 500 model and EGARCH(1,2,1), a model with 2 asymmetry terms was selected since both were statistically significant. In the IBM return data, the EGARCH model did not improve on the TARCH fit and the TARCH(1,1,1) was selected.

If alternative distributions were going to be considered, fitting these models would be the final step of the model building. However, this will be reserved for next chapter.

Testing for (G)ARCH

After 30 pages this chapter hasn't formally described how ARCH (or GARCH) is detected in the first place. The standard method is to use a test known as the ARCH-LM test which is simple a regression of residuals squared on lagged residuals squared. This should seem familiar as the ARCH-LM test directly exploits the AR representation of an ARCH process to specify a test. The model estimated is

$$\epsilon_t^2 = \phi_0 + \phi_1 \epsilon_{t-1}^2 + \phi_2 \epsilon_{t-2}^2 + ... + \phi_P \epsilon_{t-P}^2 + \eta_t$$

and the test is simple $T\mathsf{R}^2$ from this regression which has an asymptotic χ_P^2 distribution. The null is that $\phi_1 = ... = \phi_P = 0$ which would indicate that there is no persistence in the conditional variance.

8.4 Forecasting Volatility

Forecasting conditional variances with GARCH models ranges from trivial for simple ARCH and GARCH specifications to difficult for non-linear specifications. Consider the simple ARCH(1)

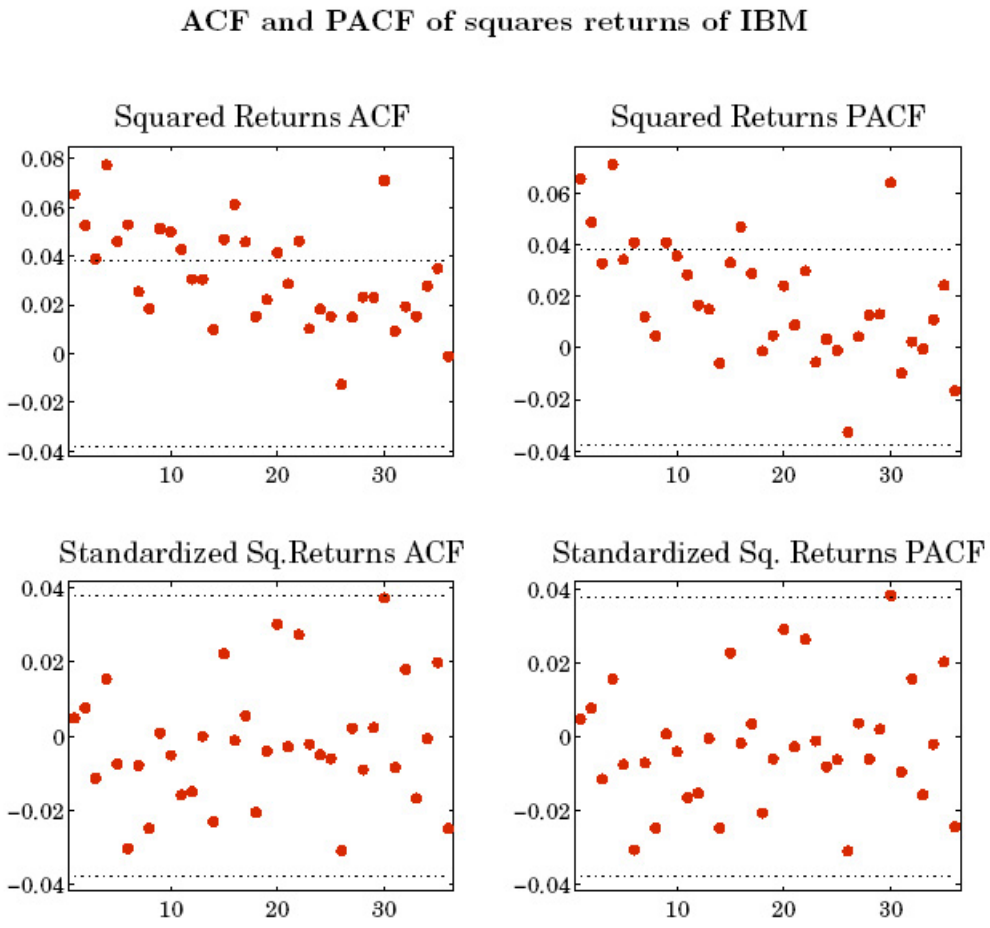

Figure 8.8: ACF and PACF of the squared returns for IBM. The bottom two figures contain the ACF and PACF of $\widehat{e}_t^2 = \widehat{\epsilon}_t^2/\widehat{\sigma}_t^2$. The top figures show persistence in both the ACF and PACF which indicate an ARMA model is needed (hence a GARCH model) while the ACF and PACF of the standardized residuals is white noise. Compared to the S&P 500 ACF and PACF, IBM appears to have weaker dynamics.

	ω	α_1	α_2	γ_1	γ_2	β_1	β_2	LL
GARCH(1,1)	0.006	0.070	-	-	-	0.928	-	-3798
	(.05)	(.00)	-	-	-	(.00)	-	
GARCH(2,1)	0.006	0.046	0.031	-	-	0.920	-	-3797
	(.09)	(.02)	(.30)	-	-	(.00)	-	
GARCH(1,2)	0.005	0.071	-	-	-	0.929	0.000	-3798
	(.06)	(.00)	-	-	-	(.00)	(.99)	
GJR-GARCH(1,1,1)	0.011	0.001	-	0.136	-	0.922	-	-3754
	(.01)	(.88)	-	(.00)	-	(.00)	-	
GJR-GARCH(2,2,1)	0.011	0.001	-	0.136	0.000	0.922	-	-3754
	(.02)	(.89)	-	(.00)	(.99)	(.00)	-	
TARCH(1,1,1)	0.016	0.012	-	0.116	-	0.929	-	-3741
	(.00)	(.10)	-	(.00)	-	(.00)	-	
EGARCH(1,1,1)	0.000	0.131	-	-0.104	-	0.982	-	-3744
	(.87)	(.00)	-	(.00)	-	(.00)	-	
EGARCH(2,1,1)**	0.001	0.122	-	-0.236	0.145	0.985	-	-3732
	(.73)	(.00)	-	(.00)	(.00)	(.00)	-	
EGARCH(1,2,1)	0.000	0.012	0.131	-0.111	-	0.979	-	-3739
	(.91)	(.92)	(.253	(.00)	-	(.00)	-	
EGARCH(1,1,2)	0.000	0.142	-	-0.115	-	0.875	0.106	-3744
	(.88)	(.00)	-	(.00)	-	(.00)	(.35)	

Table 8.5: Models used to build a model for the conditional variance of the S&P 500 Index. ** indicates the selected model.

	ω	α_1	α_2	γ_1	γ_2	β_1	β_2	LL
GARCH(1,1)	0.130	0.069	-	-	-	0.908	-	-6017
	(.10)	(.00)	-	-	-	(.00)	-	
GARCH(2,1)	0.130	0.069	0.000	-	-	0.908	-	-6017
	(.14)	(.07)	(.99)	-	-	(.00)	-	
GARCH(1,2)	0.144	0.078	-	-	-	0.760	0.136	-6017
	(.12)	(.00)	-	-	-	(.00)	(0.42)	
GJR-GARCH(1,1,1)	0.111	0.006	-	0.116	-	0.920	-	-5969
	(.00)	(.47)	-	(.00)	-	(.00)	-	
GJR-GARCH(2,2,1)	0.111	0.006	-	0.116	0.000	0.920	-	-5969
	(.00)	(.45)	-	(.02)	1.000	(.00)	-	
TARCH(1,1,1)**	0.036	0.021	-	0.082	-	0.937	-	-5949
	(.00)	(.07)	-	(.00)	-	(.00)	-	
EGARCH(1,1,1)	0.031	0.116	-	-0.073	-	0.983	-	-5952
	(.00)	(.00)	-	(.00)	-	(.00)	-	
EGARCH(2,1,1)	0.034	0.119	-	-0.023	-0.056	0.981	-	-5950
	(.00)	(.00)	-	(.56)	0.178	(.00)	-	
EGARCH(1,2,1)	0.030	0.153	-0.040	-0.073	-	0.983	-	-5951
	(.00)	(.00)	(.52)	(.00)	-	(.00)	-	
EGARCH(1,1,2)	0.027	0.100	-	-0.064	-	1.134	-0.149	-5952
	(.03)	(.05)	-	(.02)	-	(.01)	(.75)	

Table 8.6: Models used to build a model for the conditional variance of IBM. ** indicates the selected model.

process,

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 \\ \epsilon_t &= \sigma_t e_t \\ e_t &\sim \mathsf{WN}(0,1)\end{aligned}$$

Iterating forward, $\sigma_{t+1}^2 = \omega+\alpha\epsilon_t^2$, and taking conditional expectations, $\mathsf{E}_t\left[\sigma_{t+1}^2\right] = \mathsf{E}_t\left[\omega + \alpha\epsilon_t^2\right] = \omega+\alpha\epsilon_t^2$ since all of these quantities are known at time t. This is a property common to **all** GARCH models: forecasts of σ_{t+1}^2 are always known at time t. Continuing forward to $t+2$

$$\mathsf{E}_t\left[\sigma_{t+2}^2\right] = \mathsf{E}_t\left[\omega + \alpha\epsilon_{t+1}^2\right] = \omega + \alpha\left(\omega + \alpha\epsilon_t^2\right) = \omega + \alpha\omega + \alpha^2\epsilon_t^2$$

a pattern emerges,

$$\mathsf{E}_t\left[\sigma_{t+h}^2\right] = \omega\sum_{i=0}^{h-1}\alpha^i + \alpha^h\epsilon_t^2$$

This should look familiar; it is the multi-step forecasting formula for an AR(1). This shouldn't be surprising since an ARCH(1) is an AR(1).

Forecasts from GARCH(1,1) models can be derived in a similar fashion,

$$\begin{aligned}\mathsf{E}_t\left[\sigma_{t+1}^2\right] &= \mathsf{E}_t\left[\omega + \alpha\epsilon_t^2 + \beta\sigma_t^2\right] \\ &= \omega + \alpha\epsilon_t^2 + \beta\sigma_t^2 \\ \mathsf{E}_t\left[\sigma_{t+2}^2\right] &= \mathsf{E}_t\left[\omega + \alpha\epsilon_{t+1}^2 + \beta\sigma_{t+1}^2\right] \\ &= \omega + \alpha\mathsf{E}_t\left[\epsilon_{t+1}^2\right] + \beta\mathsf{E}_t\left[\sigma_{t+1}^2\right] \\ &= \omega + \alpha\mathsf{E}_t\left[e_{t+1}^2\sigma_{t+1}^2\right] + \beta\mathsf{E}_t\left[\sigma_{t+1}^2\right] \\ &= \omega + (\alpha+\beta)\,\mathsf{E}_t\left[\sigma_{t+1}^2\right] \\ &= \omega + (\alpha+\beta)\,\omega + (\alpha+\beta)\left(\alpha\epsilon_t^2 + \beta\sigma_t^2\right)\end{aligned}$$

Finally note that $\mathsf{E}_t\left[\sigma_{t+3}^2\right] = \omega + (\alpha+\beta)\,\mathsf{E}_t\left[\sigma_{t+2}^2\right]$, so

$$\begin{aligned}\mathsf{E}_t\left[\sigma_{t+3}^2\right] = \omega &+ (\alpha+\beta)\,\omega + (\alpha+\beta)^2\,\omega \\ &+ (\alpha+\beta)^2\left(\alpha\epsilon_t^2 + \beta\sigma_t^2\right)\end{aligned}$$

and the pattern emerges,

$$\mathsf{E}_t\left[\sigma_{t+h}^2\right] = \omega\sum_{i=0}^{h-1}(\alpha+\beta)^i + (\alpha+\beta)^{h-1}\left(\alpha\epsilon_t^2 + \beta\sigma_t^2\right)$$

which reduces to the ARCH forecast formula when $\beta = 0$. Forecasts from GJR-GARCH models are equally simple but forecasts form other models, particularly those which are not linear combinations of ϵ_t^2, are nontrivial and generally do not have a simple recursive formulas. For example, consider forecasting the variance from a simple TARCH(1,0,0) model,

$$\sigma_t = \omega + \alpha\left|\epsilon_{t-1}\right|$$

As is always the case, the 1-step ahead forecast is known at time t,

$$\begin{aligned}\mathsf{E}_t\left[\sigma_{t+1}^2\right] &= \mathsf{E}_t\left[\left(\omega+\alpha\left|\epsilon_t\right|\right)^2\right] \\ &= \omega^2+2\alpha\omega\left|\epsilon_t\right|+\alpha^2\epsilon_t^2\end{aligned}$$

but the two step ahead forecast is much, much more complicated.

$$\begin{aligned}\mathsf{E}_t\left[\sigma_{t+2}^2\right] &= \mathsf{E}_t\left[\left(\omega+\alpha\left|\epsilon_{t+1}\right|\right)^2\right] \\ &= \mathsf{E}_t\left[\omega^2+2\alpha\omega\left|\epsilon_{t+1}\right|+\alpha^2\epsilon_{t+1}^2\right] \\ &= \omega^2+2\alpha\omega\mathsf{E}_t\left[\left|e_{t+1}\right|\sigma_{t+1}\right]+\alpha^2\mathsf{E}_t\left[e_{t+1}^2\sigma_{t+1}^2\right] \\ &= \omega^2+2\alpha\omega\mathsf{E}_t\left[\left|e_{t+1}\right|\right]\mathsf{E}_t\left[\sigma_{t+1}\right]+\alpha^2\mathsf{E}_t\left[e_{t+1}^2\right]\mathsf{E}_t\left[\sigma_{t+1}^2\right] \\ &= \omega^2+2\alpha\omega\mathsf{E}_t\left[\left|e_{t+1}\right|\right]\left(\omega+\alpha\omega\left|\epsilon_t\right|\right)+\alpha^2\left(\omega^2+2\alpha\omega\left|\epsilon_t\right|+\alpha^2\epsilon_t^2\right)\end{aligned}$$

The issues with multi-step ahead forecasting arise because the forecast depends on more than $\mathsf{E}_t\left[e_{t+h}^2\right]=1$. In the above example, the forecast depends on both $\mathsf{E}_t\left[e_{t+1}^2\right]=1$ and $\mathsf{E}_t\left[\left|e_{t+1}\right|\right]=\sqrt{2/\pi}$ when returns are normal. The latter identity is derived by assuming shocks are conditionally normal and the final form is a mess!

$$\mathsf{E}_t\left[\sigma_{t+2}^2\right]=\omega^2+2\alpha\omega\sqrt{\frac{2}{\pi}}\left(\omega+\alpha\omega\left|\epsilon_t\right|\right)++\alpha^2\left(\omega^2+2\alpha\omega\left|\epsilon_t\right|+\alpha^2\epsilon_t^2\right)$$

The intuition behind the difficulties in forecasting with 'nonlinear'' GARCH models is simple and follows directly from Jensen's inequality. In the TARCH case,

$$\mathsf{E}_t\left[\sigma_{t+h}\right]^2\neq\mathsf{E}_t\left[\sigma_{t+h}^2\right]$$

and in the general case

$$\mathsf{E}_t\left[\sigma_{t+h}^{\lambda}\right]^{2/\lambda}\neq\mathsf{E}_t\left[\sigma_{t+h}^2\right]$$

8.4.1 Evaluating Volatility Forecasts

Evaluating forecasts from conditional variance models is essentially identical to evaluating forecasts from standard models with one caveat. In standard time series models, once time $t+h$ has arrived, the value of the variable being forecast is known. However, in volatility models, the value is unknown and must be replaced with a proxy. The standard choice if to use the squared return, r_t^2. This is reasonable if the mean squared is small relative to the variance, a reasonable approximation for daily and possibly weekly returns.

If using longer horizon measurements of returns, such as monthly, squared residuals (ϵ_t^2) from a model can be used instead. An alternative, and likely better choice if to use $RV_{it}\left(m\right)$, realized volatility, to proxy for the unobserved volatility. Once a choice of proxy has been chosen, Generalized Mincer-Zarnowitz regressions can be used to assess forecast optimality,

$$r_{t+h}^2-\widehat{\sigma}_{t+h|t}^2=\gamma_0+\gamma_1\widehat{\sigma}_{t+h|t}^2+\gamma_2 z_{1t}+...+\gamma_{K+1}z_{Kt}+\eta_t$$

where z_{jt} are any instruments known at time t. Common choices for z_{jt} include r_t^2, $|r_t|$, r_t or indicator variables for the sign of the return. One may notice that the GMZ regression above has a heteroskedastic variance and that a better GMZ-GLS can be constructed,

$$\frac{r_{t+h}^2 - \widehat{\sigma}_{t+h|t}^2}{\widehat{\sigma}_{t+h|t}^2} = \gamma_0 \frac{1}{\widehat{\sigma}_{t+h|t}^2} + \gamma_1 + \gamma_2 \frac{z_{1t}}{\widehat{\sigma}_{t+h|t}^2} + ... + \gamma_{K+1} \frac{z_{Kt}}{\widehat{\sigma}_{t+h|t}^2} + \nu_t$$

by dividing both sized by the time t forecast, $\widehat{\sigma}_{t+h|t}^2$ where $\nu_t = \eta_t / \widehat{\sigma}_{t+h|t}^2$. If one were to use the realized variance, the GMZ and GMZ-GLS regressions become

$$\begin{aligned} RV_{t+h} - \widehat{\sigma}_{t+h|t}^2 &= \gamma_0 + \gamma_1 \widehat{\sigma}_{t+h|t}^2 + \gamma_2 z_{1t} + ... + \gamma_{K+1} z_{Kt} + \eta_t \\ \frac{RV_{t+h} - \widehat{\sigma}_{t+h|t}^2}{RV_{t+h}} &= \gamma_0 \frac{1}{RV_{t+h}} + \gamma_1 \frac{\widehat{\sigma}_{t+h|t}^2}{RV_{t+h}} + \gamma_2 \frac{z_{1t}}{RV_{t+h}} + ... + \gamma_{K+1} \frac{z_{Kt}}{RV_{t+h}} + \nu_t \end{aligned}$$

where RV_{t+h} is used in place of the forecast since RV_{t+h} is a consistent estimate of the variance on day $t+h$.

Diebold-Mariano tests can also be used to assess the relative performance of two models. To perform a DM test, a loss function must be specified. Two choices for the loss function are MSE,

$$\left(r_{t+h}^2 - \widehat{\sigma}_{t+h|t}^2 \right)^2$$

and QML-loss (which is simply the kernel of the normal log-likelihood)

$$\left(\frac{r_{t+h}^2}{\widehat{\sigma}_{t+h|t}^2} + \ln\left(\widehat{\sigma}_{t+h|t}^2 \right) \right)^2$$

The DM statistic is a t-test of the null that $\overline{\delta} = 0$ where for models A and B

$$\delta_t = \left(r_{t+h}^2 - \widehat{\sigma}_{A,t+h|t}^2 \right)^2 - \left(r_{t+h}^2 - \widehat{\sigma}_{B,t+h|t}^2 \right)^2$$

in the case of the MSE loss and

$$\delta_t = \left(\frac{r_{t+h}^2}{\widehat{\sigma}_{A,t+h|t}^2} + \ln\left(\widehat{\sigma}_{A,t+h|t}^2 \right) \right)^2 - \left(\frac{r_{t+h}^2}{\widehat{\sigma}_{B,t+h|t}^2} + \ln\left(\widehat{\sigma}_{B,t+h|t}^2 \right) \right)^2$$

in the case of the QML-loss. Statistically significant positive values of $\overline{\delta} = R^{-1} \sum_{r=1}^{R} \delta_r$ indicate that B is a better model than A while negative values indicate the opposite (R is used to denote the number of out-of-sample periods used to construct the DM statistic). The QML-loss should be generally preferred as it is essentially a heteroskedasticity corrected version of the MSE.

Chapter 9

Value-at-Risk, Expected Shortfall and Density Forecasting

Note: The primary reference for these notes is Gourieroux & Jasiak (2001), although it is fairly technical. An alternative and less technical textbook treatment can be found in Christoffersen (2003) while a comprehensive and technical treatment can be found in McNeil, Frey & Embrechts (2005).

The American Heritage dictionary, Fourth Edition, defines risk as the possibility of suffering harm or loss; danger. In finance, harm or loss has a specific meaning: decreases in the value of a portfolio. This chapter provides an overview of three methods used to assess the risk of a portfolio: Value-at-Risk (VAR), Expected Shortfall, and modeling the entire density.

9.1 Defining Risk

In order to be precise about what is meant by risk, it is useful to decompose risk into the types of risk encountered in actual portfolios.

Market Risk contains all uncertainty about the future price of an asset. For example, changes in the share price of IBM due to earnings news represent market risk.

Liquidity Risk complements market risk by measuring the extra loss involves if a position must be rapidly changed. For example, if a fund wished to sell 10,000,000 shares of IBM on a single day (typical daily volume 6,000,000), this sale would have an effect on the price. This effect is separate from the market risk since it presumably represents a transitory distortion due to increased sales pressure.

Credit Risk, also known as default risk, covers cases where a 2nd party is unable to pay per previously agreed terms. Holders of corporate bonds are exposed to credit risk since a company may not be able to meet some or all of its regular coupon payments due to varying business

conditions.

Model Risk represents an econometric form of risk which captures uncertainty over the correct form of the model.

Estimation Risk captures the aspect of risk present even if a model is completely correct due the unknown parameters of this model. In practical application, parameter estimation error can result in a substantial misstatement of risk. Model and estimation risk are always present and are generally substitutes. More parsimonious models are increasingly likely to be misspecified but have less parameter estimation uncertainty.

This chapter deals with exclusively with market risk. This is not to say that credit and liquidity risk are not important, only that they typically require specialty tools beyond the scope of this course. Credit risk modeling has evolved rapidly over the past decade from using poorly conceived specifications to employing sophisticated models capable of capturing the highly nonlinear nature of defaults. Additionally, whenever building a model, estimation and model risk should always be considered. While bad models tend to produce bad forecasts, the opposite is not necessarily (i.e. good models producing good forecasts). Good models may fail due to poorly estimated parameters. This is a particularly acute problem when examining VaR since only 1 to 5% of the data contain information about the VaR.

9.2 Value-at-Risk (VaR)

The first measure examined is Value-at-Risk. Simply put, the VaR of a portfolio is the amount you are risking with some fixed probability. VaR provides a more sensible measure of the risk of the portfolio than variance since it considers losses. However, VaR is not without its own issues. These will be discussed in more detail in the section on coherent risk measures.

9.2.1 Definition

The VaR of a portfolio is usually positive. This is because it is measuring the value (£, $, , ¥, etc.) which an investor should lose with some small probability, usually between 1 and 10%.

Definition 15 (Value-at-Risk) *The α Value-at-Risk (*VaR*) of a portfolio is defined as the largest number such that the probability that the loss in portfolio value over some period of time is greater than the* VaR *is α,*

$$\Pr\left(R < -VaR\right) = \alpha$$

where $R = W_1 - W_0$ is the total return on the portfolio and W is the value of the assets in the portfolio and 1 and 0 measure an arbitrary length of time (e.g. one day or two weeks).

For example, if an investor had a portfolio value of 1,000,000 and had a daily portfolio return

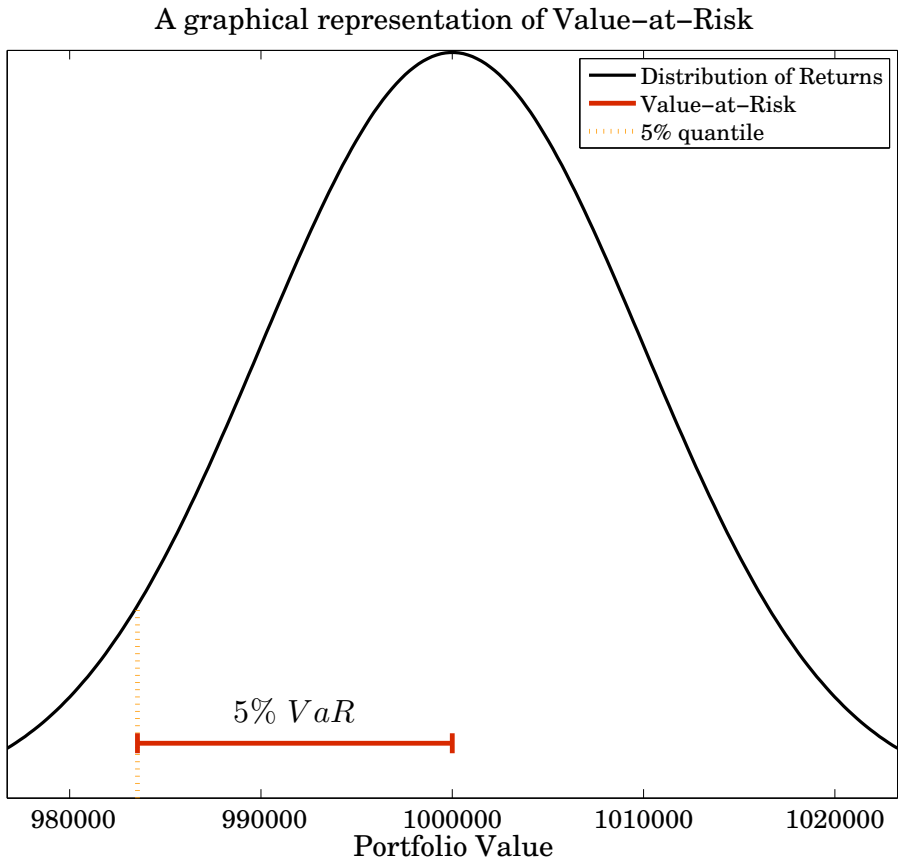

Figure 9.1: A graphical representation of VaR. The VaR is represented by the magnitude of the horizontal bar and measures the distance between the value of the portfolio today and its α quantile. In this example, α = 5%.

which was $\mathsf{N}(0.004, 0.1^2)$, the daily α Value-at-Risk of this portfolio would be

$$1,000,000 \times \left[-0.004 - 0.1\Phi^{-1}(\alpha)\right]$$

where $\Phi^{-1}(\cdot)$ is the inverse CDF of a standard normal. This expression may seem backward; it is not. The negative sign on the mean indicates that increases in the mean decrease the VaR and the negative sign on the standard deviation term indicates that increases in the volatility raise the VaR since for $\alpha < 0.5$, $\Phi^{-1}(\alpha) < 0$.

The relationship between VaR and quantiles

Understanding that VaR and quantiles are fundamentally related provides 90%+ of what you need to understand about VaR. If W represents the value of your portfolio, the α VaR is the $-1\times q_\alpha(W)$ where $q_\alpha(W)$ is the α quantile of the portfolio value. This relationship highlights two important aspects of VaR: (a) it is measured in the same units as the portfolio value is measured in (e.g. £, \$, , ¥, etc.) and (b) it should generally bo positive.[1]

9.2.2 Conditional Value-at-Risk

Most application of VaR are used to control for risk over short horizons and require a conditional version that employs information up to time t to produce a VaR for some time period $t+h$. Specifically, the conditional VaR is defined

$$\Pr\left(R_{t+h} < -VaR_{t+h|t}|\mathcal{F}_t\right) = \alpha$$

where $R_{t+h} = W_{t+h} - W_t$ is the total (, \$, etc.) return on a portfolio at time $t+h$. Most conditional models for VaR forecast the density directly, although some only attempt to estimate the required quantile of the distribution. Five standard methods will be presented here in order of the restrictiveness of the assumptions needed to justify them, from strongest to weakest.

RiskMetrics©

The RiskMetrics group has produced a surprisingly simple yet robust method for producing conditional VaR. The basic structure of the RiskMetrics model is similar the variance evolution in a GARCH(1,1) model where the coefficients sum to 1 and no constant is included,

$$\sigma_{t+1}^2 = (1-\lambda)\, r_t^2 + \lambda\sigma_t^2$$

where r_t is the (percentage) return on the portfolio in period t. In the RiskMetrics specification σ_{t+1}^2 follows an exponentially weighted moving average with weights $\lambda^0(1-\lambda)$, $\lambda^1(1-\lambda)$, $\lambda^2(1-\lambda)$... on r_t^2, r_{t-1}^2, ... By $t-100$, the weight is essentially 0. This model includes no explicit mean model for returns and is only applicable to assets with returns that are close to zero

[1] If your portfolio has a positive VaR, you either have no risk or are an incredible fund manager.

or when the time horizon is short (e.g. one day to one month). The VaR is derived from the α quantile of a normal distribution,

$$VaR_{t+1} = -W_t \sigma_{t+1} \Phi^{-1}(\alpha)$$

where W_t is the portfolio value at time t and $\Phi^{-1}(\cdot)$ is the inverse normal CDF. The attractiveness of the RiskMetrics model is that there are no parameters to estimate; λ is fixed at 0.06 for daily data (0.03 for weekly and 0.01 for monthly).[2] Additionally, this model can be trivially extended to large portfolios using a vector-matrix switch by replacing returns with a vector of returns and σ_{t+1} with a covariance matrix, $\mathbf{\Sigma}_{t+1}$. The disadvantages of the procedure are that the parameters are not estimated (which was also an advantage), it cannot be modified to incorporate a leverage effect, and the VaR follows a random walk since $\lambda + (1 - \lambda) = 1$.

GARCH Models

Fully parametric GARCH-family models provide a natural method to compute VaR. For simplicity, only a constant mean GARCH(1,1) will be described, although the mean and variance could be described using rich, sophisticated time-series models.

$$\begin{aligned} r_{t+1} &= \mu + \epsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma \epsilon_t^2 + \beta \sigma_t^2 \\ \epsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\overset{i.i.d}{\sim} f(0,1) \end{aligned}$$

where $f(0,1)$ is used to indicate that the distribution of innovations need not be normal but must have mean 0 and variance 1. For example, f could be a standardized Student's t with ν degrees of freedom or Hansen's skewed-t with degree of freedom parameter ν and asymmetry parameter λ. The parameters of the model are estimated using maximum likelihood and the time t conditional VaR is

$$VaR_{t+1} = W_t(-\widehat{\mu} - \widehat{\sigma}_{t+1} F_\alpha)$$

where F_α is the α quantile of the distribution of e_{t+1}. The most substantial limitations of this procedure are (a) required knowledge of a density family which includes f and (b) limitation to only location-scale families (i.e. where quantiles are fully characterized by the expectation and variance). The second limitation requires that all of the dynamics of the model be summarized by the time-varying mean and variance.

[2] $\lambda = 0.06$ is provided by RiskMetrics based on large studies of estimating VaR.

Semiparametric GARCH Models (*)

Semiparametric estimation mixes parametric models like ARMA or GARCH processes with nonparametric estimators of the distribution.[3] Again, consider a simple constant mean GARCH(1,1) model

$$\begin{aligned} r_{t+1} &= \mu + \epsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma\epsilon_t^2 + \beta\sigma_t^2 \\ \epsilon_{t+1} &= \sigma_{t+1}e_{t+1} \\ e_{t+1} &\overset{i.i.d}{\sim} g(0,1) \end{aligned}$$

where $g(0,1)$ is an unknown distribution with mean zero and variance 1. When $g(\cdot)$ is unknown, standard maximum likelihood estimation is not available. However, recall from the univariate volatility notes that assuming conditional normality if the errors, even if false, produces estimates which are strongly consistent. Thus, ω, γ and β will converge to their true values for essentially any $g(\cdot)$. The model can be estimated using QMLE with normal errors and the Value-at-Risk for the α quantile can be computed

$$VaR_{t+1} = W_t\left(-\widehat{\mu} - \widehat{\sigma}_{t+1}\widehat{G}_\alpha\right)$$

where $\widehat{G}_\alpha$ is the empirical α-quantile of $e_{t+1} = \epsilon_{t+1}/\sigma_{t+1}$. To estimate this quantile, define $\widehat{e}_{t+1} = \widehat{\epsilon}_{t+1}/\widehat{\sigma}_{t+1}$ and order the errors such that

$$\widehat{e}_1 < \widehat{e}_2 < ... < \widehat{e}_{N-1} < \widehat{e}_N$$

where N replaces T to indicate the residuals are no longer time ordered. $\widehat{G}_\alpha = \widehat{e}_{\lfloor \alpha N \rfloor}$ or $\widehat{G}_\alpha = \widehat{e}_{\lceil \alpha N \rceil}$ where $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor (largest integer smaller than) and ceiling (smallest integer larger than) of x. In other words, the estimate of $\widehat{G}_\alpha$ is the α quantile of the empirical distribution of $\widehat{e}_{t+1}$ which corresponds to the αT ordered $\widehat{e}_n$.

Semiparametric GARCH models provide one clear advantage over their parametric GARCH cousins; the quantile, and hence the VaR, will be consistent under weaker conditions. The primary disadvantage of the semiparametric approach is that $\widehat{G}_\alpha$ may be poorly estimated and slow to converge. It also shares the limitation of parametric GARCH models that it is limited to location scale families.

[3] This is only one example of a semiparametric estimator. Any semiparametric estimator has element of both a parametric estimator and a nonparametric estimator.

Cornish-Fisher Approximation (*)

The Cornish-Fischer approximation splits the difference between a fully parametric model and a semi parametric model. The setup is identical to that of the semiparametric model

$$\begin{aligned} r_{t+1} &= \mu + \epsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma\epsilon_t^2 + \beta\sigma_t^2 \\ \epsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\overset{i.i.d}{\sim} g(0,1) \end{aligned}$$

where $g(\cdot)$ is again an unknown distribution. The unknown parameters are estimated by quasi-maximum likelihood assuming conditional normality to produce standardized residuals, $\widehat{e}_{t+1} = \widehat{\epsilon}_{t+1}/\widehat{\sigma}_{t+1}$. The Cornish-Fisher approximation is a Taylor-series like expansion of the α VaR around the α VaR of a normal and is given by

$$\begin{aligned} VaR_{t+1} &= W_t\left(-\widehat{\mu} - \widehat{\sigma}_{t+1} CF^{-1}(\alpha)\right) \\ CF^{-1}(\alpha) &= \Phi^{-1}(\alpha) + \frac{\varsigma}{6}\left(\left[\Phi^{-1}(\alpha)\right]^2 - 1\right) \\ &\quad + \frac{\kappa - 3}{24}\left(\left[\Phi^{-1}(\alpha)\right]^3 - 3\Phi^{-1}(\alpha)\right) - \frac{\varsigma^2}{36}\left(2\left[\Phi^{-1}(\alpha)\right]^3 - 5\Phi^{-1}(\alpha)\right) \end{aligned}$$

where ς and κ are the skewness and kurtosis of $\widehat{e}_{t+1}$, respectively. From the expression for $CF^{-1}(\alpha)$, negative skewness and excess kurtosis ($\kappa > 3$, the kurtosis of a normal) decrease the estimated quantile and increases the VaR. The Cornish-Fischer approximation shares the strength of the semiparametric distribution in that it can be accurate without a parametric assumption. However, it is not necessarily consistent, a drawback. Additionally, estimates of higher order moments of standardized residuals may be problematic or the moments may not even exist.

Conditional Autoregressive Value-at-Risk (CaViaR)

Engle & Manganelli (2004) extended the techniques of standard GARCH models to estimate the conditional Value-at-Risk using quantile regression.[4] Like the variance in a GARCH model, the α quantile of the return distribution, $F_{\alpha,t+1}$, is a sum of the last periods quantile, a constant, and a 'shock'. The shock can be almost anything although a 'HIT', defined as a Value-at-Risk exceedance, is often used

$$HIT_{t+1} = I_{[r_{t+1} < F_{t+1}]} - \alpha$$

where r_{t+1} the (percentage) return and $F_{\alpha,t+1}$ is the time t α-quantile of this distribution and $I_{[r_{t+1} < F_{\alpha,t+1}]}$ is the indicator value that takes value 1 if $r_{t+1} < F_{\alpha,t+1}$ and 0 otherwise.

[4] Conditional Autoregressive Value-at-Risk is actually a misnomer. The model is actually Autoregressive Conditional Value-at-Risk, although this is less amendable to acronyms.

The evolution in a standard CaViaR model is defined

$$F_{\alpha,t+1} = \omega + \gamma HIT_t + \beta F_{\alpha,t}$$

Other forms which have been considered are the symmetric absolute value,

$$F_{\alpha,t+1} = \omega + \gamma \left| r_t \right| + \beta F_{\alpha,t}$$

the asymmetric absolute value,

$$F_{\alpha,t+1} = \omega + \gamma_1 \left| r_t \right| + \gamma_1 \left| r_t \right| I_{[r_t<0]} + \beta F_{\alpha,t}$$

and the indirect GARCH,

$$F_{\alpha,t+1} = \left(\omega + \gamma r_t^2 + \beta F_{\alpha,t}^2\right)^{\frac{1}{2}}$$

The the parameters can be estimated by minimizing the 'tick' loss function

$$\begin{aligned}
&\arg\min_{\theta} T^{-1} \sum_{t=1}^{T} \alpha \left(r_t - F_{\alpha,t}\right) \left(1 - I_{[r_t<F_{\alpha,t}]}\right) + \left(1-\alpha\right) \left(F_{\alpha,t} - r_t\right) I_{[r_t<F_{\alpha,t}]} \\
&= \arg\min_{\theta} T^{-1} \sum_{t=1}^{T} \alpha \left(r_t - F_{\alpha,t}\right) + \left(F_{\alpha,t} - r_t\right) I_{[r_t<F_{\alpha,t}]}
\end{aligned}$$

Estimation of the parameters in this problem is tricky since the objective function has many flat plots and is non-differentiable. Derivative free methods, such as simplex methods or generic algorithms, can be used to overcome these issues. The VaR in a CaViaR framework is

$$VaR = -W_t F_{\alpha,t+1}$$

Because a CaViaR model does not specify a distribution of returns, it can be valid under much weaker assumptions and does not require information about the entire density. Additionally, its parametric form provides reasonable convergence of the unknown parameters. The main drawbacks to the CaViaR modeling strategy are (a) it can produce out-of order quantiles (i.e. 5% VaR is less then 10% VaR) and (b) estimation of the unknown parameters is difficult.

Conditional Value-at-Risk for the S&P 500

The concepts of VaR will be illustrated using the S&P 500 returns form January 1, 1993 until December 31, 2003. This is the same data used in the univariate volatility notes. A number of models have been estimated which produce very similar VaR. Specifically the GARCH models, whether using a normal likelihood, a Students t, a semiparametric or Cornish-Fischer approximation all produce very similar fits and generally only differ in the quantile estimated. Table 9.1 reports parameter estimated from these a few models. Note that the TARCH parameters are extremely

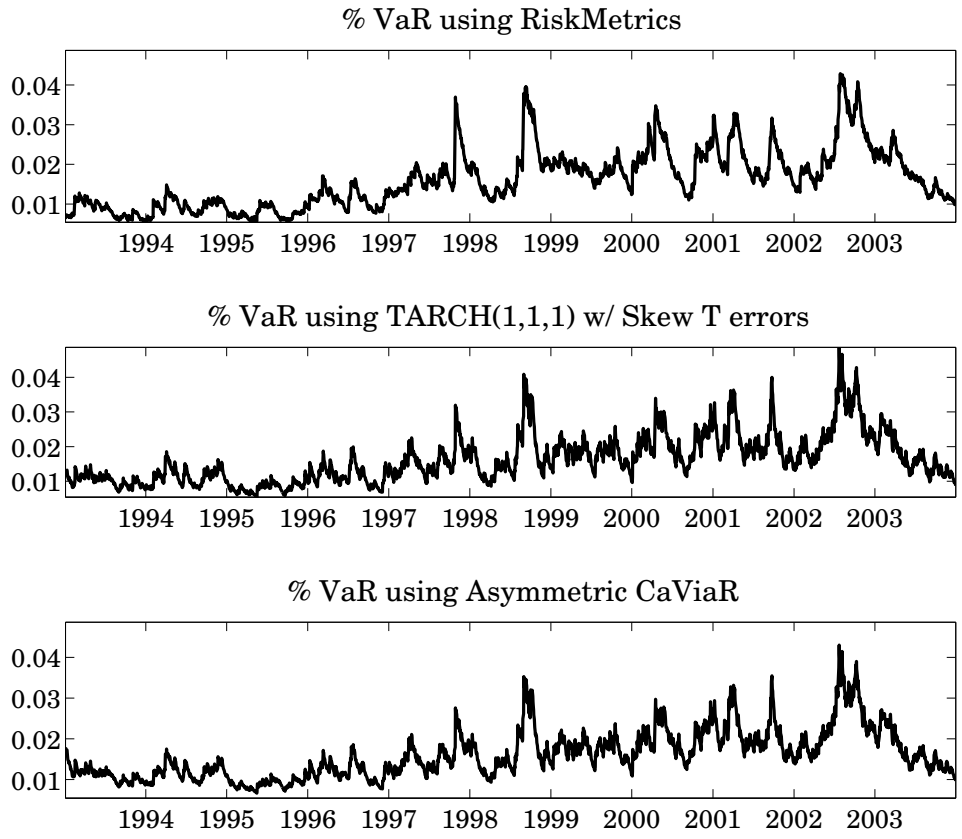

Figure 9.2: The figure contains the estimated % VaR for the S&P 500 using data from 1993 until 2003. While these three models are clearly distinct, the estimated VaR are remarkably similar.

similar in all three models, $\widehat{\mu} \approx 9$ indicating that the standardizes residuals are fat tailed and that $\widehat{\lambda} \approx 0$ indicating little skewness in the skewed t. The CaViaR estimates indicate no change in the conditional quantile for a symmetric shock, a large decrease when the return is negative and that the quantile is fairly persistent.

The table also contains estimated quantiles using the various techniques outlines in this chapter. Since the variance models were extremely similar, the only difference in the VaRs comes from the various estimates of the quantiles.

9.2.3 Unconditional Value at Risk

While the conditional VaR is often the object of interest, there may be situations which call for the unconditional (or marginal) VaR. Unconditional VaR expands the set of choices from the conditional to include ones which make do not use of an information set and derive the VaR from the unmodified distribution of returns.

Model Parameters

TARCH(1,1,1)

$$\sigma_{t+1} = \omega + \gamma_1|r_t| + \gamma_2|r_t|I_{[r_t<0]} + \beta\sigma_t$$

	ω	γ_1	γ_2	β	ν	λ
Normal	0.000	0.012	0.115	0.929		
Stud t	0.000	0.007	0.113	0.935	9.002	
Skew-t	0.000	0.008	0.113	0.934	9.292	-0.0455

CaViaR

$$F_{t+1} = \omega + \gamma_1|r_t| + \gamma_2|r_t|I_{[r_t<0]} + \beta F_t$$

	ω	γ_1	γ_2	β
Asym CaViaR	-0.000	0.004	-0.154	0.948

Estimated Quantiles

	Normal	Stud. t	Skew t	Semiparam.	CF
1%	-2.326	-2.488	-2.549	-2.639	-2.226
5%	-1.644	-1.616	-1.646	-1.639	-1.755
10%	-1.281	-1.219	-1.234	-1.254	-1.410

Table 9.1: Estimated model parameters and quantiles. Since the variance models were extremely similar, the only difference in the VaRs comes from the various estimates of the quantiles.

Parametric Estimation

The simplest form of VaR specifies a parametric model for the unconditional returns and derives the VaR from the α quantile of this distribution. For example, if $r_t \sim \mathsf{N}\left(\mu, \sigma^2\right)$, the α-VaR is

$$VaR = W\left(-\widehat{\mu} - \widehat{\sigma}\Phi^{-1}\left(\alpha\right)\right)$$

and the parameters can be directly estimated using Maximum likelihood with the usual estimators,

$$\widehat{\mu} = T^{-1}\sum_{t=1}^{T} r_t; \quad \widehat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}\left(r_t - \widehat{\mu}\right)^2$$

In a general parametric VaR models, some distribution for returns indexed a finite number of unknown parameters θ is assumed, $r_t \sim F\left(\theta\right)$ and parameters are estimated by maximum likelihood. The VaR is simply $-W_t F_\alpha$, where F_α is the α quantile of the estimated distribution. The advantages and disadvantages to parametric unconditional VaR are identical to parametric conditional VaR. The models are parsimonious and parameters estimates are generally precise yet finding a specification which necessarily includes the true distribution is difficult (or impossible).

Nonparametric Estimation (Historical Simulation)

At the other end of the spectrum is a pure nonparametric estimate of the unconditional VaR. As was the case in the semiparametric conditional VaR, the first step is to order the returns such that

$$r_1 < r_2 < ... < r_{N-1} < r_N$$

where $N = T$ is used to denote an ordering not based on time. The VaR is simple $r_{\lfloor \alpha N \rfloor}$ or alternatively $r_{\lceil \alpha N \rceil}$ or an average of the two where $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor (largest integer smaller than) and ceiling (smallest integer larger than) of x. In other words, the estimate of the VaR is simply the α quantile of the empirical distribution of $\{r_t\}$.

$$VaR = -W\widehat{G}_\alpha$$

where $\widehat{G}_\alpha$ is the estimated quantile. Historical simulation estimates are very rough and can generally be improved upon using a smoothed estimate, particularly if the data sample is small.

The advantage of nonparametric estimates of VaR is that they are generally consistent under very weak conditions and that they are trivial to compute. The disadvantage is that the VaR estimates can be highly variable and are very slow to converge to the actual VaR.

Parametric Monte Carlo

Parametric Monte Carlo is meaningfully different from either straight parametric or nonparametric estimation of the density. Rather than fit a model to the returns directly, parametric Monte Carlo

	Normal	Stud. t	Skew t	Nonparam.
1% VaR	-0.025	-0.030	-0.031	-0.028
5% VaR	-0.017	-0.016	-0.016	-0.017

Table 9.2: Unconditional estimated quantiles using a Normal, Student's t, skew t and a nonparametric estimator. The 5% quantiles are very similar while the 1% show some differences.

fits a parsimonious conditional model which is then used to simulate the unconditional distribution. For example, suppose that returns followed an AR(1) with GARCH(1,1) errors and normal innovations,

$$\begin{aligned} r_{t+1} &= \phi_0 + \phi_1 r_t + \epsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma \epsilon_t^2 + \beta \sigma_t^2 \\ \epsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\overset{i.i.d}{\sim} \mathsf{N}(0,1) \end{aligned}$$

Parametric Monte Carlo can be implemented by first estimating the parameters of the model, $\widehat{\theta} = \left[\widehat{\phi}_0, \widehat{\phi}_1, \widehat{\omega}, \widehat{\gamma}, \widehat{\beta}\right]'$ and the simulating the process for a long period to time (generally much longer than the actual number of data points available). The VaR from this models is simply the α quantile of the simulated data $\widetilde{r}_t$.

$$VaR = -W\widehat{\widetilde{G}}_\alpha$$

where $\widehat{\widetilde{G}}_\alpha$ is the empirical α quantile of the simulated $\widetilde{r}_t$. Generally the amount of simulated data is sufficient that no smoothing is needed and the empirical quantile can be reliably used. The advantage of this procedure is that it efficiently makes use of conditional information which is ignored in either parametric or nonparametric estimators of VaR and that complicated unconditional distributions can be generated from relatively simple models. The disadvantage of this procedure is that an incorrect conditional specification leads to an inconsistent estimate of the unconditional VaR.

Unconditional Value-at-Risk for the S&P 500

Using the S&P 500 data, 3 parametric models, a normal, a Student's t and a skewed t, and a nonparametric estimator were used to estimate the unconditional VaR. The estimates are largely similar although the normal differs meaningfully from the other three at the 1% VaR.

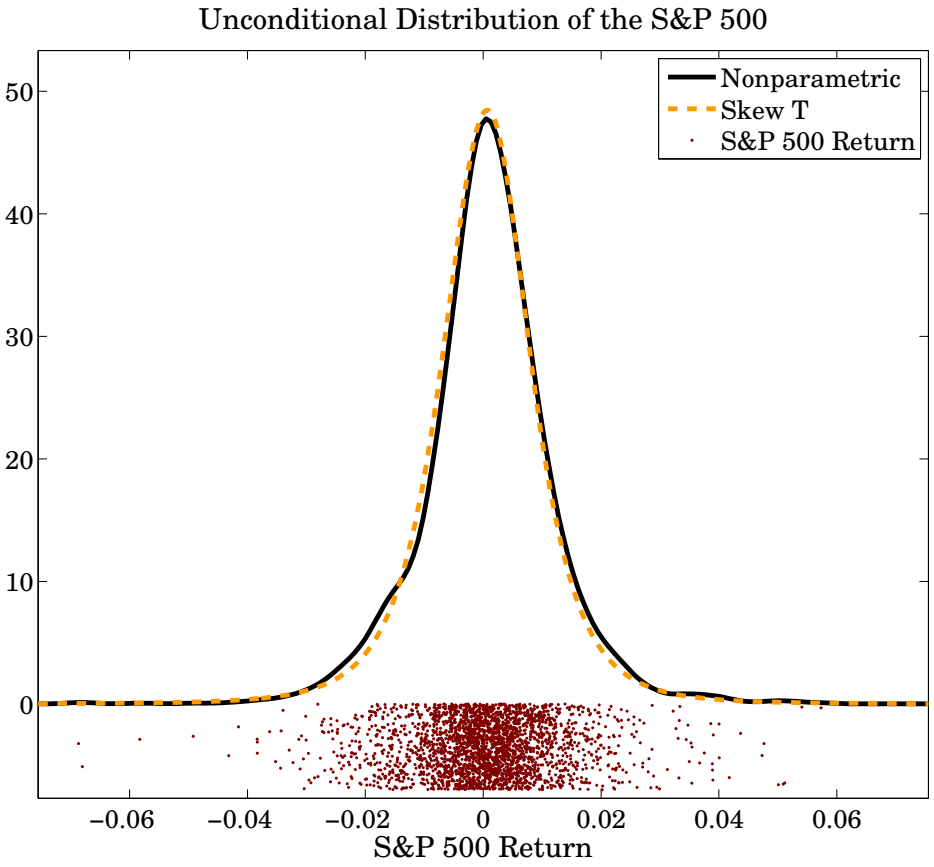

Figure 9.3: Plot of the S%P 500 returns as well as a parametric density using Hansen's skew-t and a nonparametric density estimator constructed using a kernel.

9.2.4 Evaluating VaR models

Evaluating the performance of VaR models is not fundamentally different from either ARMA or GARCH models. The key insight of VaR evaluation comes from the loss-function for VaR errors,

$$T^{-1}\sum_{t=1}^{T}\alpha\left(r_t-F_{\alpha,t}\right)\left(1-I_{[r_t<F_t]}\right)+\left(1-\alpha\right)\left(F_{\alpha,t}-r_t\right)I_{[r_t<F_t]}$$

where r_t is the return at time t and $F_{\alpha,t}$ is α quantile of the return distribution at date t. From this loss function, the generalized error can be directly computed by differentiating with respect to VaR, and is simply

$$ge_t = I_{[r_t<F_t]}-\alpha$$

which is nothing but HIT_t. When there is a VaR exceedance, $HIT_t = 1-\alpha$ and when there is none, $HIT_t = -\alpha$. If the model is correct the mean of HIT_t should be 0. Moreover, when the VaR is conditional on time t information, $\mathsf{E}_t\left[HIT_{t+1}\right]=0$ which follows from the properties of optimal forecasts (see the univariate time series notes).

A test of no conditional expectation can be performed by a general Mincer-Zarnowitz (GMZ) regression of $HIT_{t+1|t}$ on any time t available variable. For example, the estimated quantile $F_{t+1|t}$ for $t+1$ could be included as well as lagged HITs to form a regression,

$$HIT_{t+1|t}=\gamma_0+\gamma_1F_{t+1|t}+\gamma_2HIT_t+\gamma_3HIT_{t-1}+...+\gamma_KHIT_{t-K+1}+\eta_t$$

If the model is correctly specified, all of the coefficients should be zero and the null

$$\mathsf{H}_0:\gamma=\left(\gamma_1,...,\gamma_K\right)'=0$$

can be tested against an alternative

$$\mathsf{H}_1:\gamma_j\neq 0$$

for some j.

While the generalized errors can be tested in the GMZ framework, there are some improvements possible. Specifically, note that HIT_t is actually a Bernoulli random variable which takes the value $1-\alpha$ with probability α and takes the value $-\alpha$ with probability $1-\alpha$. There are ways to use this to construct Likelihood ratio test of correct specification.

9.3 Density Forecasting

VaR (a quantile) provides only limited insight into the total risk of an asset. More importantly, it may not adequately describe the types of risk an arbitrary forecast user may care about. The same cannot be said for density forecasting which contains everything there is to know about the riskiness of the asset. Density forecasts also nest both VaR and Expected Shortfall (ES) as

special cases. In light of this relationship, why do we bother with VaR and ES in the first place? Density forecasting is hard and can be very imprecise. Density forecasting suffers form a number of problems:

- Since the density contains all of the information about the random variable being studied, a flexible form is generally needed. The cost of this flexibility is increased parameter estimation error which can be magnified when evaluating the expectation of nonlinear functions of the asset (e.g. options) over the density.

- Multi-step ahead density forecasts are difficult to impossible since densities do not aggregate. This contrasts form standard results in ARMA and GARCH models. (note that we do not derive multi-step VaR in these notes that they constitute a straightforward extension).

- Unless the user has preferences over the entire distribution, density forecasting inefficiently uses to information.

9.3.1 Density Forecasts from GARCH models

Density forecasting from GARCH models is simple. The setup is identical to that of using GARCH models to compute VaR and again, for simplicity, only a model with a constant mean and GARCH(1,1) variances will be used. As was the case in the VaR application of GARCH, the mean and variance can be described using much richer, more sophisticated time-series models.

$$\begin{aligned} r_{t+1} &= \mu + \epsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma\epsilon_t^2 + \beta\sigma_t^2 \\ \epsilon_{t+1} &= \sigma_{t+1}e_{t+1} \\ e_{t+1} &\overset{i.i.d}{\sim} g(0,1) \end{aligned}$$

where $g(0,1)$ is used to indicate that the distribution of innovations need not be normal but must have mean 0 and variance 1. Standard choices for $g(\cdot)$ include the standardized Student's t, the generalized error distribution, and Hansen's skew T. The 1-step ahead density forecast if simply

$$f_{t+1|t} \sim g\left(\mu, \sigma_{t+1|t}^2\right)$$

where $f(\cdot)$ is the distribution of returns. This follow directly from the original model since $r_{t+1} = \mu + \sigma_{t+1}e_{t+1}$ and $e_{t+1} \overset{i.i.d}{\sim} g(0,1)$.

9.3.2 Semiparametric Density forecasting

Semiparametric Density forecasting is also essentially the same as it's VaR counterpart. The model begins by assuming that innovations are generated according to some unknown distribution $g(\cdot)$,

$$\begin{aligned} r_{t+1} &= \mu + \epsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma\epsilon_t^2 + \beta\sigma_t^2 \\ \epsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\overset{i.i.d}{\sim} g(0,1) \end{aligned}$$

As was the case in the VaR application, estimates of σ_t^2 are computed assuming that the innovations are conditionally normal. The justification for this choice follows from the strong consistency of the variance model estimates even when the innovations are not normal. Using the estimated variances, standardized innovations are estimated as $\widehat{e}_t = \widehat{\epsilon}_t/\widehat{\sigma}_t$. The final step is to compute the density. The simplest method to accomplish this is to compute the empirical CDF as

$$G(e) = T^{-1}\sum_{t=1}^{T} I_{[\widehat{e}_t<e]}$$

which simply sums up the number of standardized residuals than e. This method is trivial but has some limitations. First, the density does not exist since is function is not differentiable. This makes taking expectation of arbitrary functions difficult. This can be overcome by using a small number of bins and computing the histogram. Second, the CDF is very jagged and is generally inefficient.

An alternative, and more efficient method, is to compute a kernel density of residuals. The kernel density is simply a local average of how many $\widehat{e}_t$ there are in a small neighborhood of x. The more in this neighborhood the higher the probability in the region. The kernel density is

$$g(e) = \frac{1}{Th}\sum_{t=1}^{T} K\left(\frac{\widehat{e}_t - e}{h}\right)$$

where $K(\cdot)$ can be one of many kernels, although most commonly used are the normal

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-x^2/2\right)$$

and the Epanechnikov

$$K(x) = \begin{cases} \frac{3}{4}\left(1-x^2\right) & -1 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

The bandwidth (h) is usually set to Silverman's bandwidth , $h = 1.06\sigma_x T^{-1/5}$ where σ_x is the standard deviation of x, the input to the kernel. However, larger or smaller bandwidths can be used to produce smoother or rougher densities, respectively. If the CDF is needed, $G(e)$ can be computed using numerical techniques such as a trapezoidal approximation to the Riemann integral or quadrature on $f(e)$.

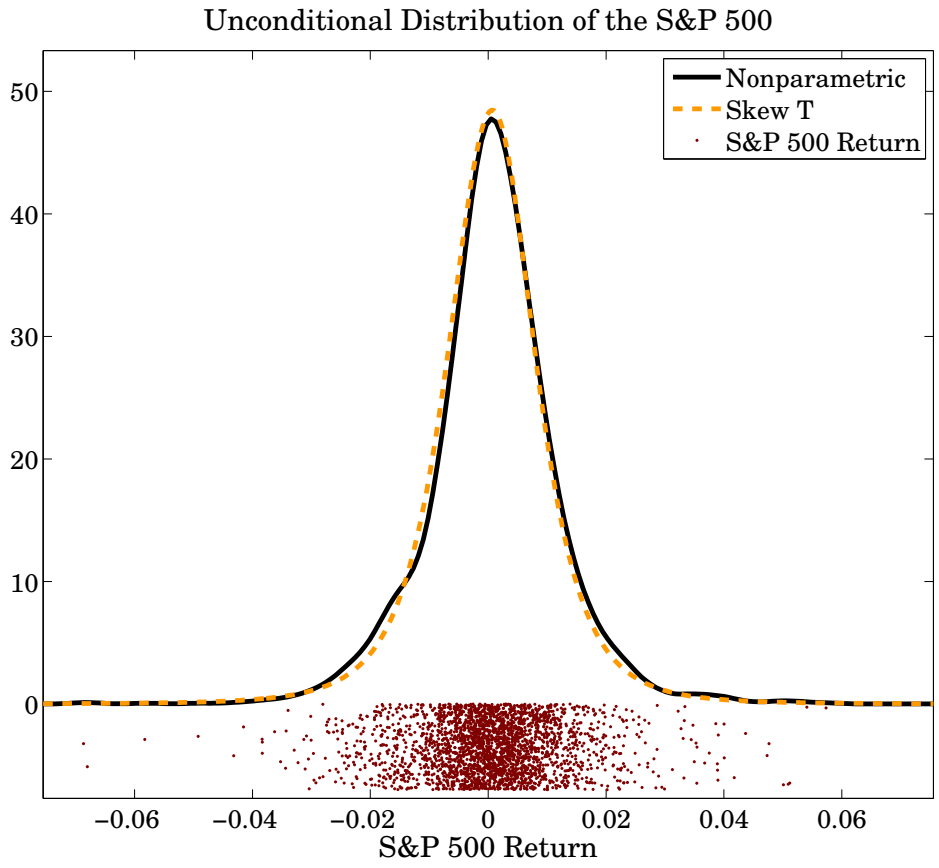

Figure 9.4: Unconditional estimated quantiles using a Normal, Student's t, skew t and a nonparametric estimator. The 5% quantiles are very similar while the 1% show some differences.

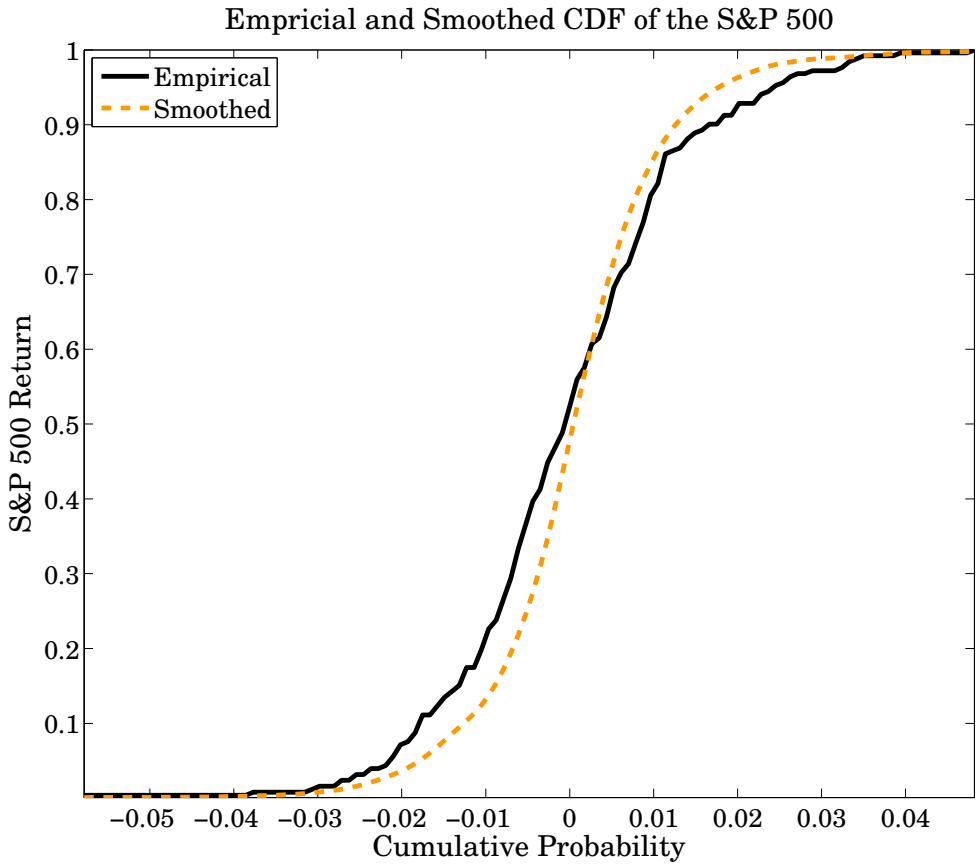

Figure 9.5: The rough empirical and smoothed empirical CDF for the returns to the S&P 500 in 2001. The smoothed CDF is generally more accurate and puts probability in regions where the rough CDF has none.

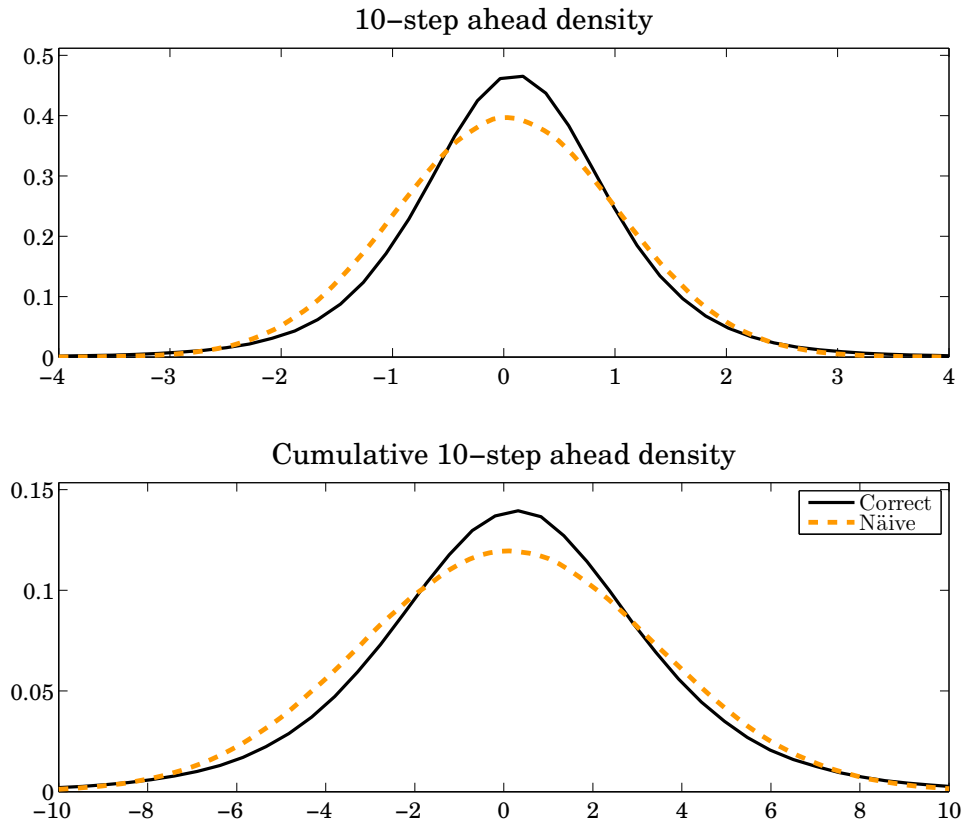

Figure 9.6: Naïve and correct 10-step ahead density forecasts from a simulated GARCH(1,1) model. The correct density forecasts have substantially fatter tails then the naïve forecast and are more centrally peaked. The naïve forecast assumes a normal density for the conditional forecast of the returns $r_{t+h|t}$. In reality, the conditional forecast is a weighted sum of successive one-step ahead forecasts; the weights are random and function of the relative volatilities. Hence $r_{t+h|t}$ is not truly normally distributed.

9.3.3 Multi-step density forecasting and the fan plot

Multi-step ahead density forecasting if not trivial and we do not consider it here.

Fan plots are a simple method to convey information about future changes in uncertainty. Their use has been made popular by the Bank of England and they are a good way to "wow" a less-sophisticated-than-you audience. Figure 9.7 contains a simple example of a fan plot which contains the density of a standard random walk which begins at 0 and has $i.i.d.$ standard normal increments. Darker regions indicate higher probability while progressively lighter regions indicate positive but low probability events.

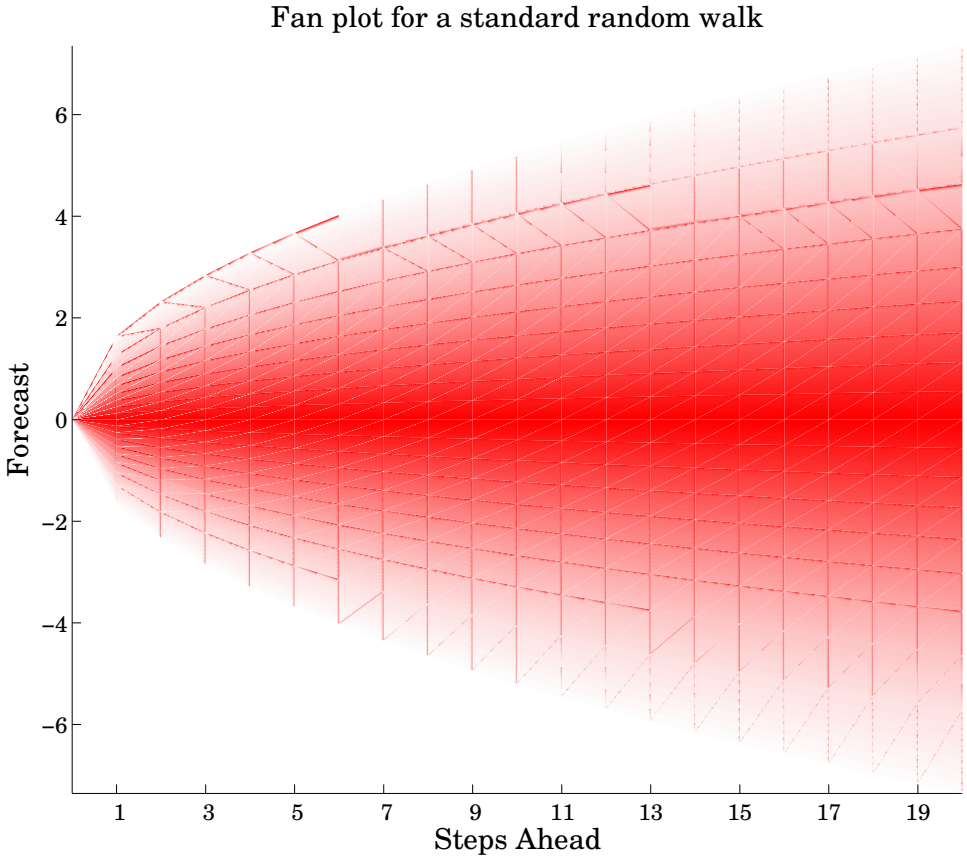

Figure 9.7: Future density of a standard random walk beginning at 0 with $i.i.d.$ standard normal increments. Darker regions indicate higher probability while progressively lighter regions indicate positive but low probability events.

9.4 Expected Shortfall

9.4.1 Definition

Expected shortfall combines aspects of the VaR methodology with more information in the tail. Expected Shortfall (ES) is the expected value of the portfolio loss given a Value-at-Risk exceedance has occurred.[5] Mathematically, the *unconditional* ES is

$$ES = \mathsf{E}\left[W_1 - W_0 | W_1 - W_0 < -VaR\right]$$

while the *conditional*, and generally more useful ES is

$$ES_{t+1} = \mathsf{E}_t\left[W_{t+1} - W_t | W_{t+1} - W_t < -VaR_{t+1}\right]$$

Because computation of ES requires both a quantile and an expectation, these are generally computed from complete distributional models, either parametric or semi-parametric.

9.4.2 Evaluating ES models

Evaluation of Expected Shortfall models uses standard techniques. ES is just a conditional mean,

$$\mathsf{E}_t\left[ES_{t+1}\right] = \mathsf{E}_t\left[W_{t+1} - W_t | W_{t+1} - W_t < -VaR_{t+1}\right]$$

and a generalized Mincer-Zarnowitz regression can be used to test whether this mean is zero. Let $I_{[r_t<VaR_t]}$ indicate that the portfolio return was less than the VaR, the GMZ regression for testing ES is

$$\left(ES_{t+1} - R_{t+1}\right) I_{[R_{t+1}<VaR_{t+1}]} = \mathbf{x}_t\gamma$$

where $\mathbf{x}_t$, as always, is any set of time t measurable instruments and $R_{t+1} = W_{t+1} - W_t$. The only obvious choices for $\mathbf{x}_t$ are a constant and $ES_{t+1|t}$, the forecast expected shortfall, although other regressors which capture characteristics of the tail, such as recent volatility $\left(\sum_{i=0}^{\tau} r_{t-i}^2\right)$ or recent VaR ($\sum_{i=0}^{\tau}\mathrm{VaR}_{t-i}$) would be reasonable choices. The fundamental difficulty with this regression is the lack of data; ES can only be measured when there is a VaR exceedance and, at 5%, this would only result in 50 observations out of a sample size of 1000. This generally makes evaluating ES models difficult and can lead to a failure to reject in many cases. If the the ES model is correct, $\mathsf{H}_0 : \gamma = \mathbf{0}$ should not be rejected. If it is, then the ES is predictable and can be improved.

[5] ES is a special case of a broad class of expectations known as exceedance measures which describe interesting relationship focusing on the behavior in teh tails. In the case of ES, it is an exceedance mean. Other exceedance measures which have been studies include exceedance variance, $\mathsf{V}[x|x < q]$, exceedance correlation, $\mathsf{Corr}(x, y|x < q_{\alpha,x}, y < q_{\alpha,y})$, where $q_{\alpha,\cdot}$ is the quantile of the distribution of x or y. All of these statistics capture the behavior of standard measures in the tail of the distribution.

9.5 Bibliography

Christoffersen, P. (2003), Elements of Financial Risk Management, Academic Press, London.

Engle, R. F. & Manganelli, S. (2004), 'Caviar: Conditional autoregressive value at risk by regression quantiles', Journal of

Business & Economic Statistics 22(4), 367–381.

Gourieroux, C. & Jasiak, J. (2001), Value at Risk. Handbook of Financial Econometrics.

McNeil, A. J., Frey, R. & Embrechts, P. (2005), Quantitative Risk Management : Concepts, Techniques, and Tools, Princeton University Press, Woodstock, Oxfordshire.

Chapter 10

Exercises

EXERCISE 1

This exercise is for you to practice your skills with summations.

Let two random variables X and Y; show that

$$\mathsf{Var}\left[X\right] = \mathsf{E}\left[X^2\right] - \mathsf{E}\left[X\right]^2$$

and

$$\mathsf{Cov}\left[X, Y\right] = \mathsf{E}\left[XY\right] - \mathsf{E}\left[X\right]\mathsf{E}\left[Y\right].$$

Show also that these expressions hold for sample (or empirical) means, variances and covariances (i.e. computed from the sample, as opposed to the true population values).

EXERCISE 2

Let the data generating process:

$$y_t = \alpha y_{t-1} + u_t, \tag{10.1}$$

where $|\alpha| < 1$ and $\{u_t\}$ is a standard Gaussian White Noise:

$$u_t \sim \mathsf{NID}\left(0, 1\right).$$

Under each of the following hypotheses, determine whether the process y_t is weakly stationary:

(i) the process $\{y_t\}$ started at $t_0 = -\infty,$ so that at each point in time, the process has infinite history.

(ii) the process starts at $t_0 = 0,$ with fixed initial value $y_0 = \overline{y}_0$ and $\{y_t\}$ obeys (1) for all $t > 0$.

(iii) the process starts at $t = 0$ with $y_0 \sim \mathsf{N}\left(0, \frac{1}{1-\alpha^2}\right)$ and obeys (1) for all $t > 0$.

EXERCISE 3

Let $\{u_t\}$ a Gaussian white noise, i.e. $u_t \sim \mathsf{NID}\left(0, \sigma^2\right)$. The stochastic process $\{y_t\}$ derived from $\{u_t\}$ is defined as:

$$\begin{array}{lll} (i) & y_t = & \alpha y_{t-1} + u_t, \quad \text{for } t > 0, \text{ where } \alpha = \frac{1}{2} \text{ and } y_0 = 1. \\ (ii) & y_t = & u_t - \beta u_{t-1}, \quad \text{with } \beta = 1 \text{ for } t = 0, \pm 1, \pm 2, ... \\ (iii) & y_t = & \begin{cases} 1 + u_t & \text{for } t = 1, 3, 5, ... \\ -1 + u_t & \text{for } t = 2, 4, 6, ... \end{cases} \end{array}$$

For each of the processes defined above:

(a) Find the mean μ_t and autocovariance function $\gamma_t(h)$ of $\{y_t\}$;
(b) Determine whether the process is (weakly) stationary.

EXERCISE 4

Consider the following ARMA$(2, 1)$ process:

$$y_t = \alpha_2 y_{t-2} + \varepsilon_t + \beta_1 \varepsilon_{t-1},$$

where it is assumed that (i) $\beta_1 \neq \pm\sqrt{\alpha_2}$ and that (ii) $\{\varepsilon_t\}$ is a standard Gaussian white noise:

$$\varepsilon_t \sim \mathsf{NID}(0, 1).$$

(a) Why is it assumed that (i) $\beta_1 \neq \pm\sqrt{\alpha_2}$?

(b) Under what conditions is the process $\{y_t\}$ stationary?

(c) Assuming that $\{y_t\}$ is stationary, what is the autocovariance function $\gamma(h)$ for $h \geq 0$.

EXERCISE 5

You will find below eight figures with four graphs each. These represent from left to right, top to bottom, (a) the time series, (b) its autocorrelation function (ACF), (c) its partial autocorrelation function, and (d) a scatter plot of y_t against y_{t-1}. These graphs were drawn from series generated

using the same white noise series $\{\varepsilon_t\}$ as:

$$
\begin{aligned}
&(1): y_t = y_{t-1} + \varepsilon_t - 0.5\varepsilon_{t-1} \\
&(2): y_t = 0.95y_{t-1} + \varepsilon_t \\
&(3): y_t = 0.6y_{t-1} + \varepsilon_t \\
&(4): y_t = 0.6y_{t-4} + \varepsilon_t - 0.3\varepsilon_{t-1} \\
&(5): y_t = 0.2y_{t-1} + 0.8y_{t-4} + \varepsilon_t \\
&(6): y_t = 1.3y_{t-1} - 0.4y_{t-2} + \varepsilon_t - 0.5\varepsilon_{t-1} \\
&(7): y_t = \varepsilon_t + 0.8\varepsilon_{t-1} \\
&(8): y_t = y_{t-1} + \varepsilon_t
\end{aligned}
$$

(a) Using the information at your disposal, including any computation that you may think useful, find which of series 1 to 8 corresponds to graphs A to H.

(b) Which of the series are stationary?

(c) Which of the series present empirical properties that differ from their theoretical properties?

EXERCISE 6

From an observed series $\{y_t\}$, two stationary processes are considered as potential candidates for the data generating process (DGP):

$$y_t = \nu + \alpha y_{t-2} + u_t, \quad u_t \sim \mathsf{NID}\left(0, \sigma_u^2\right), \tag{10.2}$$

$$y_t = \mu + \varepsilon_t + \beta\varepsilon_{t-2}, \quad \varepsilon_t \sim \mathsf{NID}\left(0, \sigma_\varepsilon^2\right). \tag{10.3}$$

Let the empirical moments be, for three different cases:

$$
\begin{aligned}
(i) \quad & \overline{y} = 0, \quad \widehat{\mathsf{Var}}\left[y_t\right] = 2, \quad \widehat{\mathsf{Corr}}\left[y_t, y_{t-1}\right] = 0,9, \quad \widehat{\mathsf{Corr}}\left[y_t, y_{t-2}\right] = 0,4\,; \\
(ii) \quad & \overline{y} = 1, \quad \widehat{\mathsf{Var}}\left[y_t\right] = 2, \quad \widehat{\mathsf{Corr}}\left[y_t, y_{t-1}\right] \approx 0, \quad \widehat{\mathsf{Corr}}\left[y_t, y_{t-2}\right] = 0,4\,; \\
(iii) \quad & \overline{y} = 1, \quad \widehat{\mathsf{Var}}\left[y_t\right] = 2, \quad \widehat{\mathsf{Corr}}\left[y_t, y_{t-1}\right] \approx 0, \quad \widehat{\mathsf{Corr}}\left[y_t, y_{t-2}\right] = 0,8\,;
\end{aligned}
$$

Find, if possible, the values the model parameters (1), $\left\{\nu, \alpha, \sigma_u^2\right\}$ and (2), $\left\{\mu, \beta, \sigma_\varepsilon^2\right\}$ from the empirical moments and discuss the additional information that could help you identify the DGP.

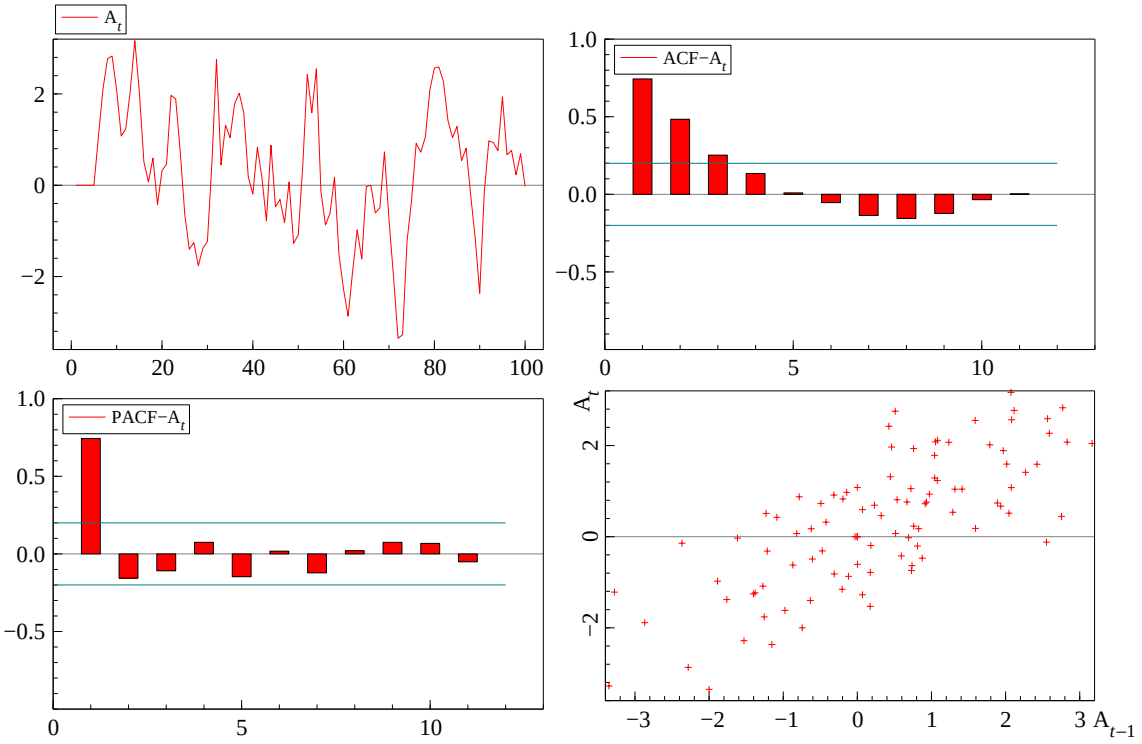

Figure 10.1: Série A

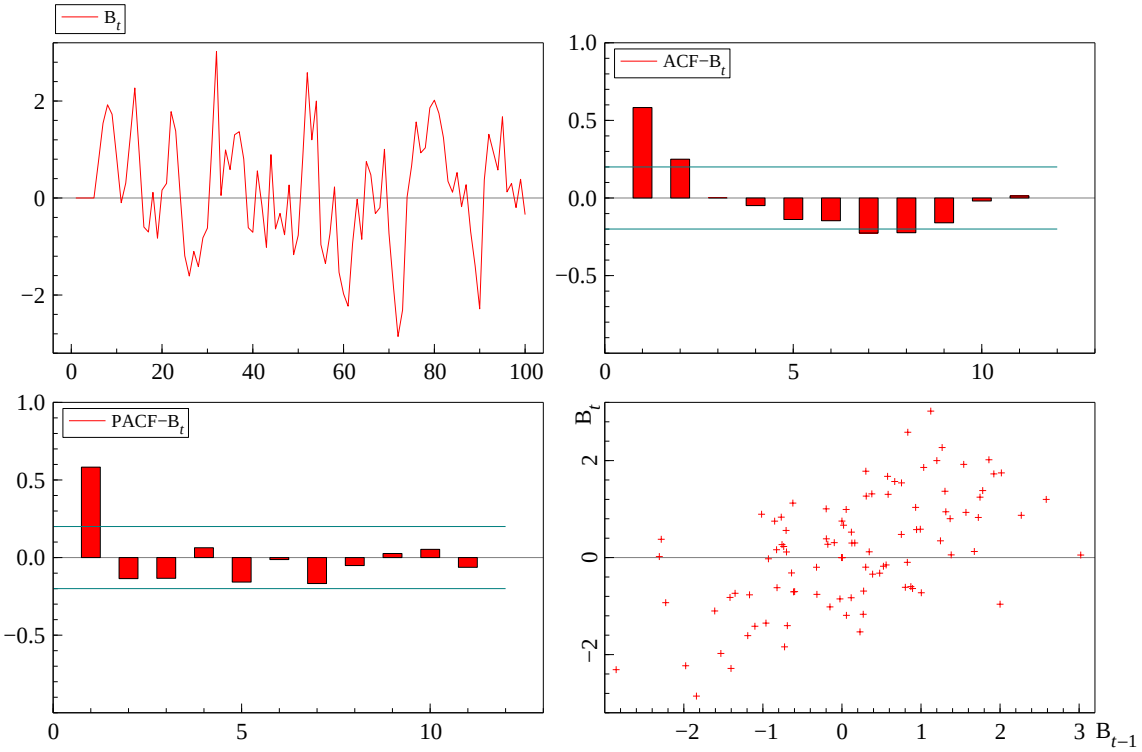

Figure 10.2: Série B

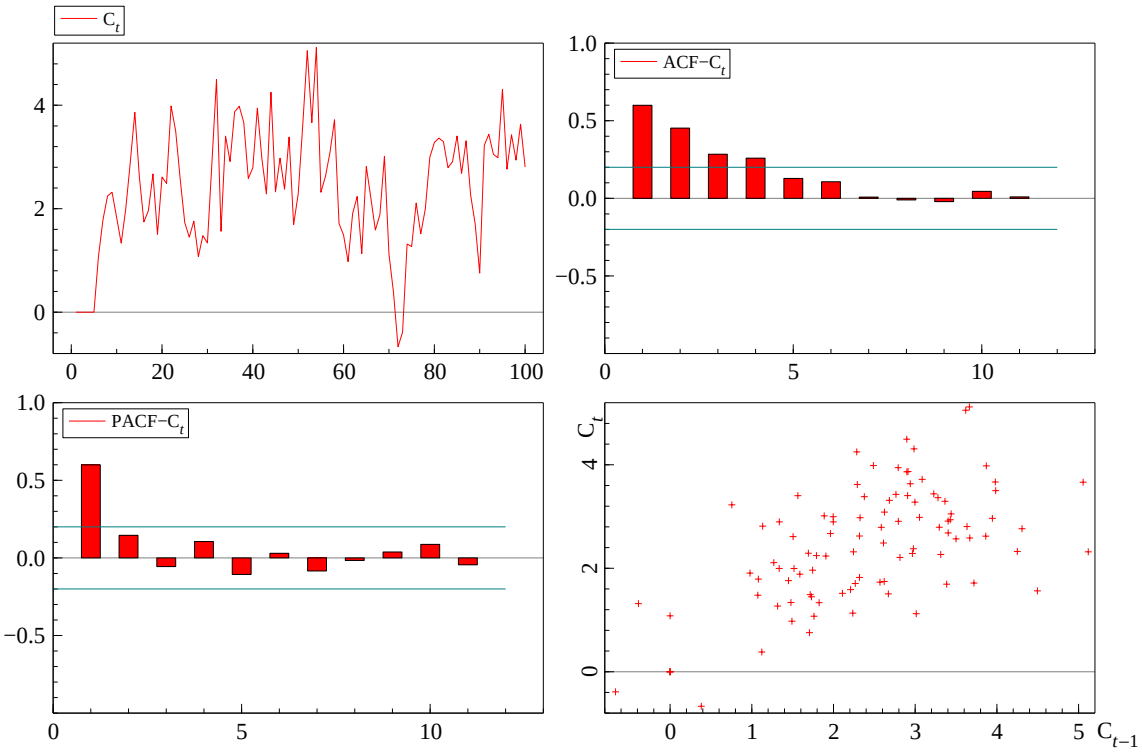

Figure 10.3: Série C

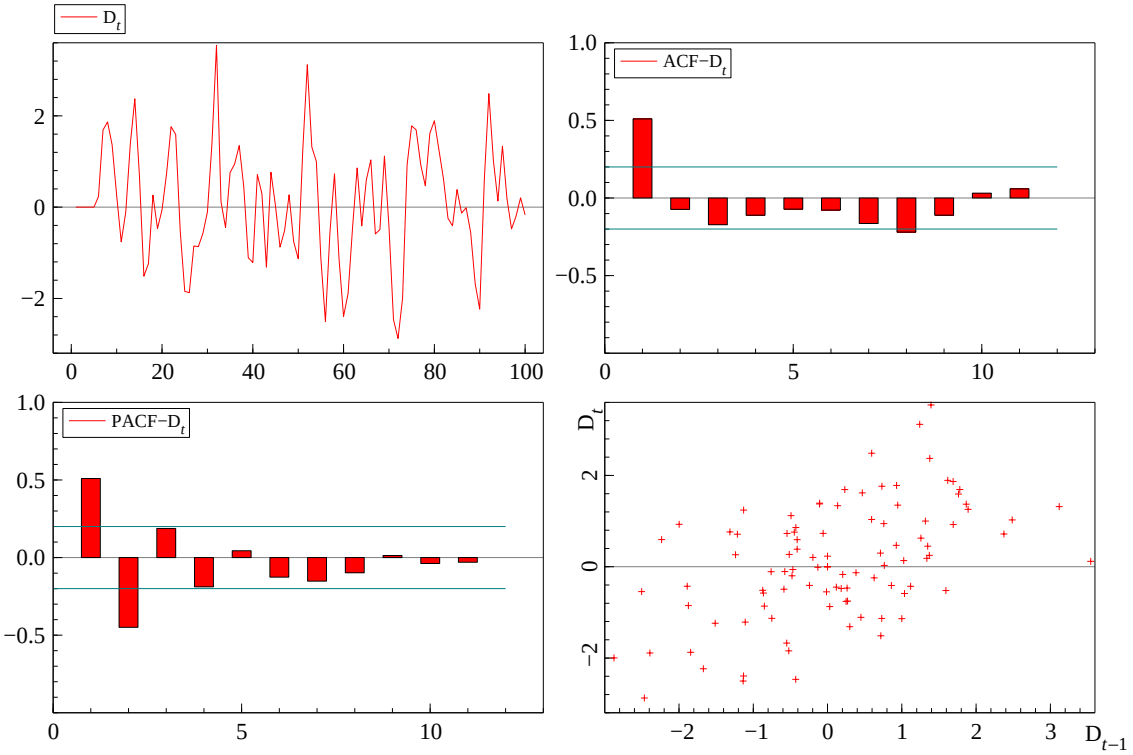

Figure 10.4: Série D

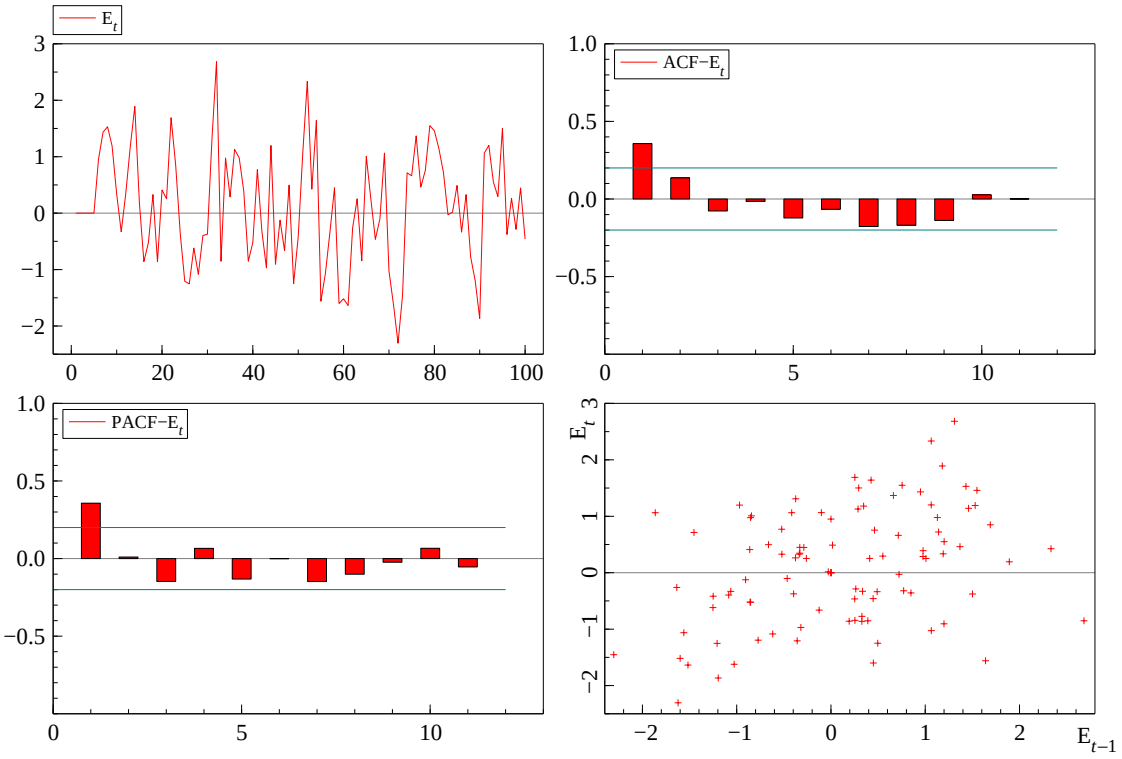

Figure 10.5: Série E

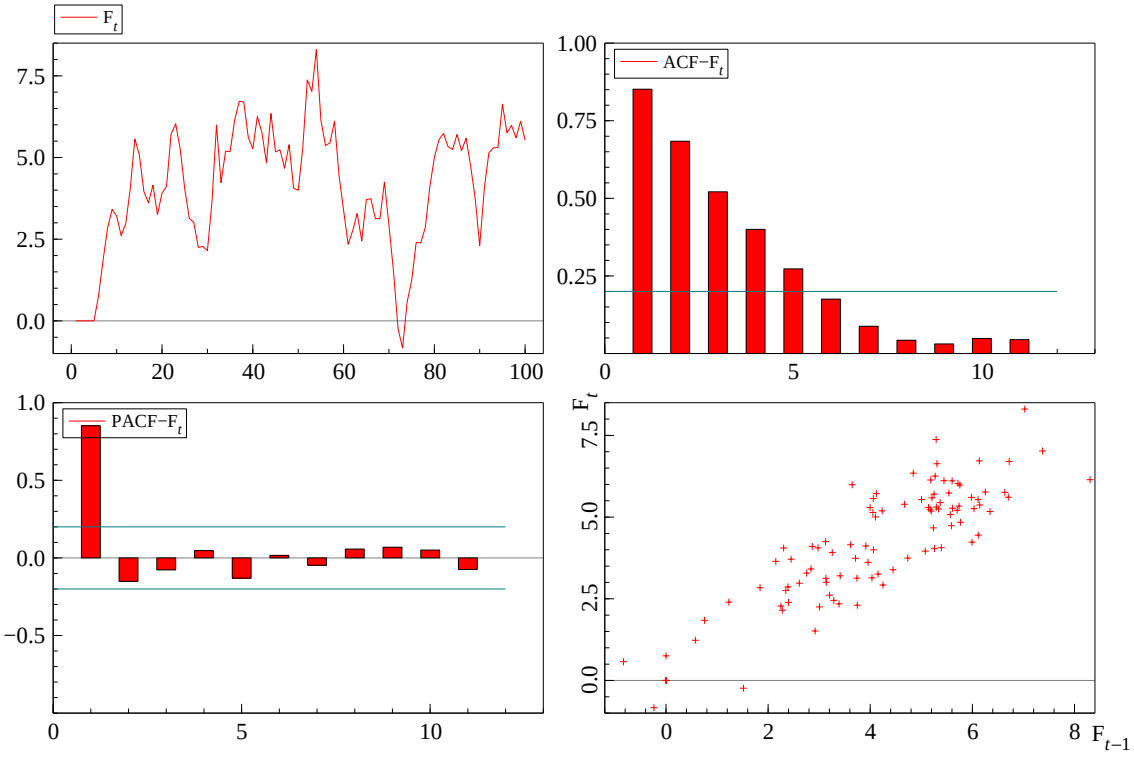

Figure 10.6: Série F

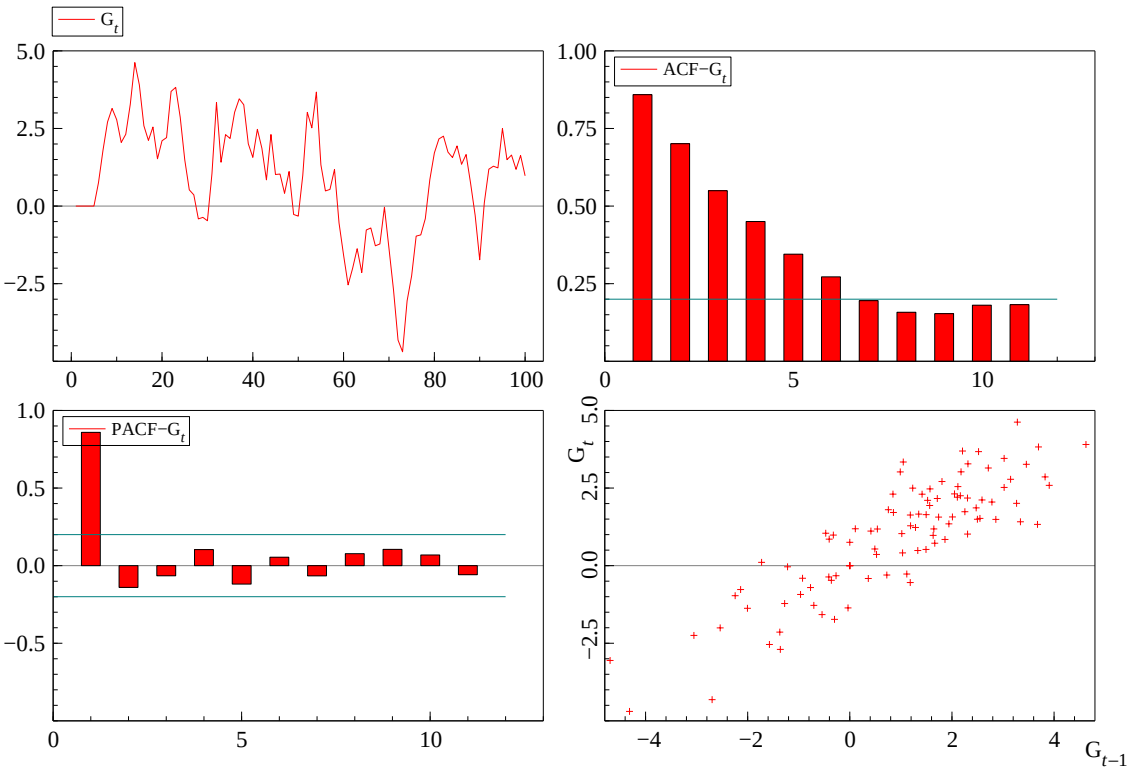

Figure 10.7: Série G

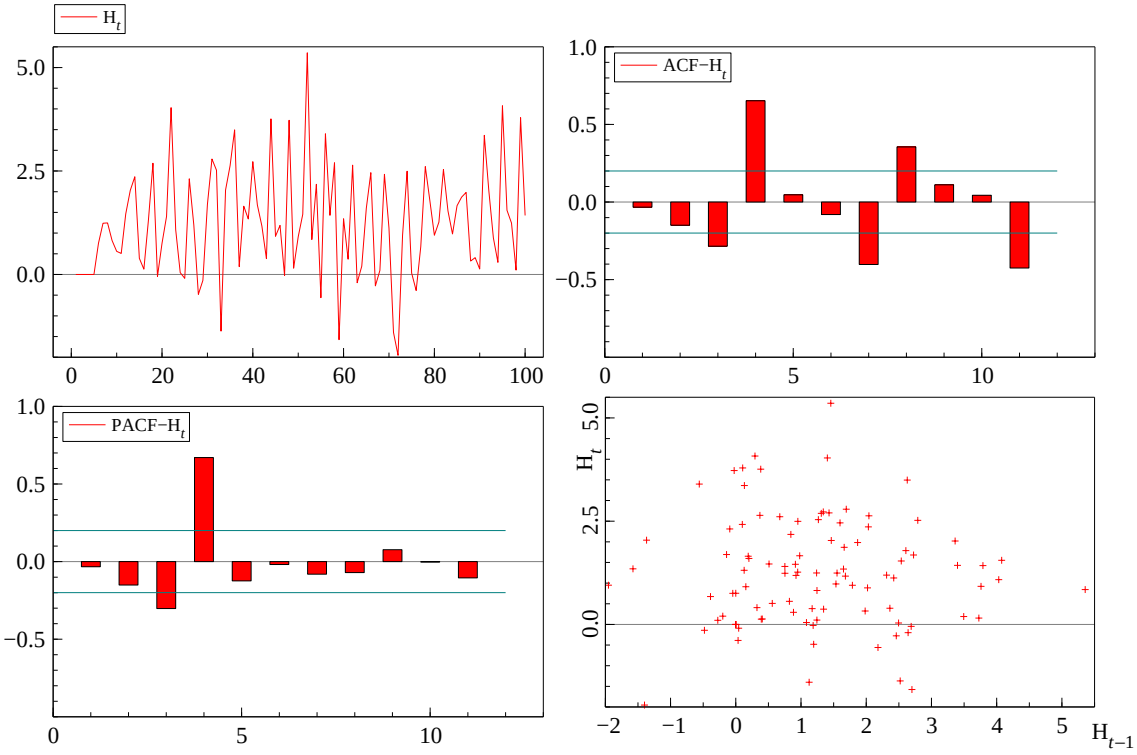

Figure 10.8: Série H

EXERCISE 7

We define the exponential smoother (or Exponentially Weighted Moving Average (EWMA) model) for a series y_t as

$$\begin{aligned}\widetilde{y}_{T+1} &= \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} \ldots . \\ &= \alpha \sum_{i=0}^{\infty} (1-\alpha)^i y_{T-i},\end{aligned}$$

for $\alpha \in (0,1)$.

1. Show that the exponential smoother is mean preserving.
2. What is the effect of varying α?
3. At a date T, the exponential smoother is used for forecasting y_{T+h} for $h > 0$. Show that the forecast is the same at all horizons.

EXERCISE 8

Let $\{u_t\}$ a Gaussian white noise, i.e. $u_t \sim \mathsf{NID}\left(0, \sigma^2\right)$. The stochastic process $\{y_t\}$ derived from $\{u_t\}$ is defined as:

$$\begin{array}{lll} (i) & y_t = & \alpha y_{t-1} + u_t, \qquad \text{for } t > 0, \text{ where } \alpha = 1 \text{ and } y_0 = 2. \\ (ii) & y_t = & \tau + \alpha y_{t-1} + u_t, \quad \text{for } t > 0, \text{ where } \alpha = 1, \tau \neq 0, \text{ and } y_0 = 0\,; \end{array}$$

For each of the processes defined above:

(a) Find the mean μ_t and autocovariance function $\gamma_t(h)$ of $\{y_t\}$;
(b) Determine whether the process is (weakly) stationary.

EXERCISE 9

Download, for a country of your choice, the quarterly or monthly series concerning:
– macroeconomic activity (GDP, index of industrial production...) ;
– exchange rates
– prices (e.g. GDP deflator, wage index, consumer price index) ;
Then analyze them graphically as follows :

1. Compute the following transforms for the original series (say X_t) and their logarithms ($x_t = \ln X_t$): obtain the first differences $\Delta X_t = X_t - X_{t-1}$ and Δx_t, the second differences $\Delta(\Delta X_t)$ and $\Delta^2 x_t$

2. Plot the series graphically (using E-Views) and their ACF

3. Discuss the properties of the time series, in particular whether they are stationary. You can use any information that E-Views provides.

4. Fit a univariate time-series model to **one** (and one only) of the series and explain your choice of data transformation and model specification. Discuss the model fit, diagnostic tests, etc. Do you think this is a reasonable model?

   Note: A number of websites are available to download such data. Examples include: UK Office for National Statistics; INSEE; Organisation for Economic Cooperation and Development; Federal Reserve Bank of St. Louis; http://www.economagic.com/; etc.

EXERCISE 10

We dispose of a series $\{y_t\}$ for $t = -\infty$ to T, but only consider the observations for $t = 1, ..., T$. This series seems to present a linear deterministic trend but we hesitate between two models for its representation:

$$\mathsf{TS} : y_t = \alpha + \beta t + u_t, \tag{10.4}$$

$$\mathsf{DS} : y_t = y_{t-1} + u_t \tag{10.5}$$

where $u_t \sim \mathsf{IN}\left(0, \sigma_u^2\right)$.

1. *Briefly* comment on the properties of the two TS and DS models. Why do we hesitate between them?

2. What would your approach be, should you wish to find which models fit the series best?

   In order to make a choice between the models, we focus on their forecasting properties. We assume that, at time T, we wish to generate a forecast h periods into the future, with $h > 0$.

3. If y_t follows model TS, what is the value of y_{T+h} as a function of y_T? Same question if y_t follows model DS.

   We assume that for each model, we generate a forecast for y_{T+h}. We denote them by $\widehat{y}_{T+h}^{\mathsf{TS}}$ and $\widehat{y}_{T+h}^{\mathsf{DS}}$ respectively. In order to compute them, we assume that the forecast is the expectation of y_{T+h} conditional on y_T; more precisely:

   $$\widehat{y}_{T+h}^{\mathsf{TS}} = \mathsf{E}\left[y_{T+h}|y_T\right] \text{ assuming that } y_t \text{ follows model TS},$$
   $$\widehat{y}_{T+h}^{\mathsf{DS}} = \mathsf{E}\left[y_{T+h}|y_T\right] \text{ assuming that } y_t \text{ follows model DS},$$

   Assume to simplify that $\mathrm{E}[u_T|y_T] = 0$.

4. Compute $\widehat{y}_{T+h}^{\mathsf{TS}}$ and $\widehat{y}_{T+h}^{\mathsf{DS}}$.

5. We assume in this question that y_t follows model TS

   (a) Using model TS to compute the forecast, we define the forecast error as:

   $$e_{\mathsf{TS|TS}} = y_{T+h} - \widehat{y}_{T+h}^{\mathsf{TS}}$$

   Derive $e_{\mathsf{TS|TS}}$, its expectation and variance.

   (b) Alternatively, we erroneously use model DS to compute the forecast and define the forecast error, when y_{T+h} follows model TS:

   $$e_{\mathsf{DS|TS}} = y_{T+h} - \widehat{y}_{T+h}^{\mathsf{DS}}$$

   Derive $e_{\mathsf{DS|TS}}$, its expectation and variance.

6. Now, we assume that y_t follows model DS and define as before:

   $$e_{\mathsf{TS|DS}} = y_{T+h} - \widehat{y}_{T+h}^{\mathsf{TS}}$$

   and

   $$e_{\mathsf{DS|DS}} = y_{T+h} - \widehat{y}_{T+h}^{\mathsf{DS}}$$

   Derive $e_{\mathsf{TS|DS}}$ and $e_{\mathsf{DS|DS}}$, their expectations and variances.

7. Assuming that it is possible that we are mistaken in our model choice. To minimize the error risk, which model would you pick? What do you think of this criterion for model choice.

EXERCISE 11: VAR

Consider the VAR(1) given by

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} .1 \\ 1 \end{bmatrix} + \begin{bmatrix} .8 & 0 \\ .2 & .4 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

1. Is the model stationary, cointegrated, or with two independent unit roots?

2. Trace out the 6 steps of the impulse response function for a unit shock to ϵ_{1t} and ϵ_{2t}

3. Does x_t Granger cause y_t, and vice versa?

EXERCISE 12: VAR

Consider the following VAR model

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 0 & 1/2 \\ 1 & 1/2 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix} \tag{10.6}$$

also denoted

$$\begin{aligned} \mathbf{z}_t &= \mu + \mathbf{A}\mathbf{z}_{t-1} + \epsilon_t \\ \mathbf{\Phi}(L)\,\mathbf{z}_t &= \mu + \epsilon_t \end{aligned}$$

with ϵ_t a white noise.

1. Check whether $\mathbf{z}_t$ is integrated of order 1
2. Compute $\mathbf{\Phi}(1)$: what is its rank? what do you conclude from this?
3. Write the Error Correction form for this model
4. How many stochastic trends does $\mathbf{z}_t$ exhibit?

EXERCISE 13

What does it mean for two variables to be cointegrated? why is it problematic? why is it interesting?

EXERCISE 14: Volatility

Consider an ARCH(1) process given by

$$\begin{aligned} r_t &= \mu_t + \epsilon_t \\ \mu_t &= \phi_0 + \phi_1 r_{t-1} \\ \epsilon_t &\sim \mathsf{N}\left(0, \sigma_t^2\right) \\ \sigma_t^2 &= \omega + \alpha_1 \epsilon_{t-1}^2 \end{aligned}$$

1. What is $\mathsf{E}[r_t]$]?
2. What is $\mathsf{E}_{t-1}[r_t]$]?
3. What is $\mathsf{E}\left[\epsilon_t^2\right]$]?
4. What is $\mathsf{E}_{t-1}\left[\epsilon_t^2\right]$]?

5. What is $\mathsf{Cov}[\epsilon_t, \epsilon_{t-1}]$]?
6. How does the conditional variance evolve in a GARCH(1,1)? Describe the evolution both mathematically and in words.
7. What is the h-step ahead forecast formula for an ARCH(1) model?
8. What is the h-step ahead forecast formula for a GARCH(1,1) model?
9. What is the main advantage of a GARCH process over an ARCH process?
10. Name and describe one or two important features the classic GARCH model is missing?

EXERCISE 15: The impact of permanent and transitory shocks to output

The purpose of this exercise is to study a decomposition of a time series y_t, Gross Domestic Product in particular, into a permanent and a transitory components. This is often called a *Trend/Cycle (T/C) decomposition* or simply a permanent/transitory decomposition

$$y_t = T_t + C_t$$

where T_t represents the trend and C_t the cyclic, and possibly noisy, component.

The first part leads you through the Beveridge and Nelson (1980) decomposition, the second through aspects of the Unobserved Component model.

The Beveridge-Nelson (BN) decomposition.

Beveridge and Nelson (1980) proposed a definition of the permanent component of an $\mathsf{I}(1)$ time series y_t with drift δ as the limiting forecast as the horizon grows to infinity, adjusted for the mean rate of growth over the forecast horizon: if

$$y_t = \mu + \delta t + z_t$$

where $\mathsf{E}[z_t] = 0$ then the BN permanent component is

$$BN_t = \lim_{h \to \infty} \left(\widehat{y}_{t+h|t} - \delta h \right)$$

where $\widehat{y}_{t+h|t}$ denotes the forecast of y_{t+h} using the information available at time t.

We follow Stock and Watson (1987) into assuming that y_t, the log of the U.S. real GDP, follows an ARIMA(0,1,1) model, i.e. $x_t = \Delta y_t$ follows an MA(1). They estimated, using postwar quarterly data over 1947:2–1985:4 that, for $t = 1, ...,$

$$x_t = 0.008 + \epsilon_t + 0.3\epsilon_{t-1}$$

where

$$\epsilon_t \sim \mathsf{iid}(0, \sigma^2); \;\; \widehat{\sigma} = 0.0106.$$

1. Stationarity and integration. Answer the following questions with detailed explanations.

   (a) What does it mean whether a series is stationary or not?

   (b) What do we mean when we say that a series is integrated? Give examples of models that generate integrated series.

   (c) Why do we aim in economic modeling to find out whether a variable is integrated? Does it really matter?

   (d) How would you test whether an observed variable $y_1, ..., y_t$ is stationary, whether is is integrated?

2. Study of x_t.

   (a) What are the expectation, variance and autocovariance functions of x_t?

   (b) Is x_t stationary?

   (c) What do we mean when we refer to invertibility of the MA lag polynomial? Does such a property hold here for x_t?

   (d) What are the 'optimal' forecasts of x_{t+h}, $h \geq 0$, given information available at time t? What do we mean by optimal?

   (e) What is the long run response of x_{t+h} to a shock ϵ_t?

3. Study of y_t.

   (a) Find the ARMA representation for y_t. Is is stationary?

   (b) Express y_t as a function of y_0 and the sequence of shocks $\epsilon_1, ...\epsilon_t$ (assume $\epsilon_0 = 0$). Does y_t exhibit a linear (deterministic) trend?

   (c) Derive the optimal forecasts $\widehat{y}_{t+h|t}$, for $h \geq 1$.

   (d) What is the BN permanent component derived from the long run response of y_t?

   (e) The cyclical component is defined as $c_t = y_t - BN_t$. Find an expression for it here. What are its dynamic properties?

The Unobserved Component (UC) decomposition

The basic idea behind the UC model is to give structural equations for the components on the trend-cycle decomposition. For example, Watson (1986) considers UC-ARIMA models of the form

$$\begin{aligned} y_t &= \mu_t + C_t \\ \mu_t &= \alpha + \mu_{t-1} + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2) \\ \phi(L)C_t &= \theta(L)\eta_t, \quad \eta_t \sim iid(0, \sigma_\eta^2) \end{aligned} \tag{10.7}$$

where $\phi(L)$ is of order p and $\theta(L)$ of order q. Notice that in Watson's example, the trend component, μ_t, is a random walk with drift and the cyclical component, C_t, is an ARMA(p, q) process. As it stands, however, the parameters of the UC model are not identified without further restrictions. Restrictions commonly used in practice to identify all of the parameters are:
(1) the roots of $\phi(z) = 0$ are outside the unit circle;
(2) $\theta(L) = 1$, and
(3) $\mathsf{Cov}(\epsilon_t, \eta_t) = 0$.

These restrictions identify C_t as a transitory autoregressive `cyclical´component, and μ_t as the permanent trend component. The restriction $\mathsf{Cov}(\epsilon_t, \eta_t) = 0$ states that shocks to C_t and μ_t are uncorrelated. As shown in Morley, Nelson and Zivot (2003), for certain models the assumption that $\mathsf{Cov}(\epsilon_t, \eta_t) = 0$ turns out to be an over-identifying restriction.

1. We consider the model by Blanchard, L'Huillier and Lorenzoni (2009). They assume that (log) productivity a_t can be decomposed into the sum of a permanent x_t and a transitory z_t component.

   $$a_t = x_t + z_t$$

   where x_t is integrated of order 1: $\Delta x_t = \rho \Delta x_{t-1} + \epsilon_t$ and z_t is stationary $z_t = \rho z_{t-1} + \eta_t$ where ρ is identical in both equations.

   (a) For x_t to be integrated of order 1, what assumption do we need to make about the value of ρ?

   (b) Give expressions for the covariances $\mathsf{Cov}\left(\Delta x_t, \Delta x_{t-1}\right)$ and $\mathsf{Cov}\left(\Delta z_t, \Delta z_{t-1}\right)$ as functions of the model parameters.

   (c) Find an expression for the covariance beetween Δa_t and Δa_{t-1} assuming that ϵ_t and η_t are independent (and hence x_t and z_t also). Show that you can write, for $j \geq 1$

   $$\mathsf{Cov}\left(\Delta a_t, \Delta a_{t-j}\right) = \frac{\rho^j}{1-\rho^2}\sigma_\epsilon^2 + \frac{1-\rho}{1+\rho}\rho^{j-1}\sigma_\eta^2$$

(d) What is the value of $\mathsf{Cov}\left(\Delta a_t, \Delta a_{t-j}\right)$ for $j \geq 1$ when

$$\sigma_\epsilon^2 = \frac{(1-\rho)^2}{\rho}\sigma_\eta^2 \tag{10.8}$$

(e) We assume now that condition (10.8) holds throughout.

Let $u_t = \Delta a_t$, so by definition

$$a_t = a_{t-1} + u_t$$

using your answer to question (d) above, find an expression for the variance $\mathsf{Var}\left(a_t\right)$ as a function of t and $\sigma_u^2 = Var\left(u_t\right)$ under the assumption that $a_0 = 0$. Would you argue that a_t follows a random walk?

2. The authors show that in a one-sector model (i.e. output equals consumption, $y_t = c_t$), if agents aim to smooth their consumption, c_t over their infinite life span,

$$c_t = \mathsf{E}[c_{t+1}|\mathcal{I}_t]$$

where $\mathcal{I}_t$ is the consumers' information set at date t, then it can be shown that

$$c_t = \lim_{h \longrightarrow \infty} \mathsf{E}[a_{t+h}|\mathcal{I}_t] \tag{10.9}$$

(a) In view of your answers to all the questions above, comment on the latter equation (10.9). Is c_t stationary?

(b) Assuming that agents have access to perfect information about the permanent component x_t, (so that they observe it at any time), the authors solve the model to show that

$$c_t = c_{t-1} + \frac{1}{1-\rho}\epsilon_t$$
$$a_t = c_{t-1} + \rho\left(a_{t-1} - c_{t-1}\right) + \epsilon_t + \eta_t$$

i. Using a VAR representation, does a_t Granger-Cause c_t? does c_t Granger-Cause a_t?

ii. Is the vector $(c_t, a_t)'$ stationary?

iii. Find an AR(1) representation for the variable $w_t = a_t - c_t$.

iv. Comment on the cointegration properties of the variables c_t and a_t. (this is simple)

(c) The authors also show that when agents have no information about (x_t, z_t) (and they do not observe them), the model can be solved as

$$c_t = a_{t-1} + u_t$$
$$a_t = a_{t-1} + u_t$$

What can be said about the Granger-Causality properties of this system?

3. For 2.b/ and 2.c/ above, draw the impulse response functions of consumption to

   (a) a permanent shift in productivity ϵ_t. Does it make a difference whether the shift comes from ϵ_t or u_t? (try to answer this simply).

   (b) a permanent shift in the transitory shock η_t.

References

Beveridge, Stephen and Nelson, Charles R., (1981). "A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'," Journal of Monetary Economics, 7(2), 151-174.

Blanchard, O. J., J.-P. L'Huillier and G. Lorenzoni (2009). "News, Noise, and Fluctuations: An Empirical Exploration," NBER Working Paper No. 15015.

Morley, J., C.R. Nelson, and E. Zivot (2003). "Why are Beveridge-Nelson and Unobserved-Components Decompositions of GDP so Different?," The Review of Economics and Statistics, 85(2), 235-243.

Stock, J.H. and M.W. Watson (1988). "Variable Trends in Economic Time Series," Journal of Economics Perspectives, 2, 147-174.

Watson, M.W. (1986). "Univariate Detrending Methods with Stochastic Trends," Journal of Monetary Economics, 18, 49-75.

EXERCISE 16: The Phillips Curve

Consider the model

$$u_t = \overline{u} + \beta \left(\pi_t - \pi^e_{t|t-1} \right) + \epsilon^s_t$$

where π_t denotes the rate of inflation, $\pi^e_{t|t-1}$ denotes the inflationary expectation formed at time $t-1$, u_t denotes the unemployment rate, $\overline{u}$ is constant, and ϵ^s_t is an error term (which you can interpret as being a supply shock) satisfying $\mathsf{E}\left[\epsilon^s_t\right] = 0$, $\mathsf{Var}\left[\epsilon^s_t\right] = \sigma^2_\epsilon$ and $\mathsf{E}\left[\epsilon^s_t | \pi_{t-1}, \pi_{t-2}, ..., u_{t-1}, u_{t-2}, ...\right] = 0$.

1. (a) Give a simple economic interpretation to this model, and to $\overline{u}$. What would you conjecture about the sign of β? Explain.

   (b) Assume that inflationary expectations are formed adaptively, i.e., assume the partial-adjustment specification

$$\pi^e_{t+1|t} - \pi^e_{t|t-1} = (1-\mu) \left(\pi_t - \pi^e_{t|t-1} \right)$$

holds. Show that, under this updating scheme, the original model can also be written as

$$u_t = (1-\mu)\,\overline{u} + \mu u_{t-1} + \beta \Delta \pi_t + v_t \tag{10.10}$$

where v_t is the MA(1) process $v_t = \epsilon_t^s - \mu \epsilon_{t-1}^s$, $\Delta = 1 - L$ with L the lag operator.

(c) Assume in addition that inflation follows

$$\pi_t = \alpha_0 + \alpha_1 \pi_{t-1} + \alpha_2 u_{t-1} + \epsilon_t^d$$

where ϵ_t^d is inflation surprise, an iid shock that is independent of ϵ_{t-j}^s for all j. Derive the conditional-mean function

$$\mathsf{E}\left[\pi_t | \pi_{t-1}, \pi_{t-2}, ..., u_{t-1}, u_{t-2}, ...\right]$$

(d) Combine the representation for u_t under (10.10) with the inflation process here in a VAR representation.

(e) Compute the response of (u_{t+1}, π_{t+1}) to unit shocks $\left(\epsilon_t^s, \epsilon_t^d\right)$ (compute the responses to the two shocks independently).

(f) What are the long-run responses of (u_{t+j}, π_{t+j}), $j \to \infty$, to unit shocks $\left(\epsilon_t^s, \epsilon_t^d\right)$?

EXERCISE 17: Present-Value Model

Consider the Present Value Model

$$(1+r)\,P_t = \mathsf{E}\left[P_{t+1} + D_{t+1} | \mathcal{I}_t\right] \tag{10.11}$$

where P_t is the price of an asset, D_t the cash flow it yields between times $t-1$ and t and $\mathcal{I}_t$ denotes the information set available at time t to market participants.

1. (a) Show – using the law of iterated expectations – that expression (10.11) can be written (under a condition that you will precise) as the sum of expected future cash flows

$$P_t = \sum_{j=1}^{\infty} \frac{1}{(1+r)^j} \mathsf{E}\left[D_{t+j} | \mathcal{I}_t\right]$$

(b) How would you interpret the direction of causality between processes $(P_t)_{t\in\mathbb{Z}}$ and $(D_t)_{t\in\mathbb{Z}}$ ($\mathbb{Z}$ denotes the set of relative integers $-\infty, ..., -2, -1, 0, 1, 2, ..., \infty$).

(c) Assume now that D_t follows the process

$$D_t = d + u_t + \delta u_{t-1} + v_t$$

where u_t and v_t are independent Gaussian white noises. Assume also that the information set $\mathcal{I}_t$ consists of $(u_t, u_{t-1}, ...)$ and $(v_t, v_{t-1}, ...)$.
Compute the expected values $\mathsf{E}\left[D_{t+j} | \mathcal{I}_t\right]$ for $j = 1, 2, ...$

(d) Use your previous results to express the dynamics for P_t as a function of the model parameters and the processes u_t and v_t

(e) Combine your results to find a VAR(1) representation for the vector (P_t, D_t) (be careful with the definition of the innovations that drive the VAR) Comment on its Granger-Causality properties.