Birkbeck Economics

MSc Economics, PGCert Econometrics
MSc Financial Economics

Autumn 2009

# ECONOMETRICS

Ron Smith : R.Smith@bbk.ac.uk

Contents



## Notes

# 1. Introduction

## 1.1. Aims

This course provides an introduction to theoretical and applied econometrics which emphasises actually doing applied econometrics. This involves combining economic theory, statistical methods and an understanding of the data with the ability to use the appropriate software.

Econometrics now divides into time-series (often using macroeconomic or financial data) and microeconometrics (often using large cross-sections of data). The applications during the first term emphasise time-series methods, microeconometric methods are covered in more detail in the second term and the Advanced Econometrics option of the Economics MSc. However most of the basic results for the linear regression model apply to cross-section as well as time series data. The difference is that issues of temporal dependence, dynamics, are more important in time-series, while issues of non-linearity are more important in microeconometrics. Distinguishing correlations from causality is central to both.

## 1.2. Learning Outcomes

- Derive standard estimators (OLS, ML, GLS) and understand their properties
- Explain the basis for standard exact and asymptotic tests and use them in practice
- Develop and analyse basic univariate and multivariate time-series models for integrated and cointegrated data and know how to choose between alternative models
- Use standard econometrics packages and interpret their output
- Read understand and explain empirical articles in the literature of the sort that appear in the Economic Journal or American Economic Review
- Conduct and report on an independent piece of empirical research that uses advanced econometric techniques.

### 1.3. Your input

- To achieve the learning outcomes (and pass the exams) requires a lot of independent work by you. We will assume that you know how to learn and that there are things that we do not have to tell you because you can work them out or look them up for yourself.
- Read these notes.
- Ensure that you are familiar with matrix algebra, statistics and standard economic models.
- Ensure that you are familiar with a statistical package such as EViews or Stata.
- Try to attend all lectures and classes, if you have to miss them make sure that you know what they covered and get copies of notes from other students.
- Do the class exercises in advance. Continuously review the material in lectures, classes and these notes, working in groups if you can.
- Identify gaps in your knowledge and take action to fill them, by asking questions of lecturers or class teachers and by searching in text books. We are available to answer questions during office hours (posted on our doors) or by email.
- Do the applied exercise (section 2.3 of the notes) during the first term. We will assume that you have done it and base exam questions on it.
- Start thinking about a topic for your project during the first term and work on it during the second term.

### 1.4. Assessment

Assessment is based on two thirds examination, one third applied econometric project. The exam will contain seven questions: four in Section A, roughly based on topics covered in the first term; three in Section B, roughly based on topics covered in the second term. The division is rough because the topics overlap a lot. You will be required to do three questions, at least one from each section. The division into sections this year is slightly different from previous years. You will not be provided with statistical tables in the exam, any critical values needed

will be provided on the question paper. The exam will involve calculations and you may take a non-programmable calculator.

### 1.5. Structure of the Booklet

After this introduction, the rest of Section 1 sets out what is being covered in the course and recommended reading.

Section 2 exercises. These are (i) exercises for the weekly tutorial classes in the first term; (ii) some examples, mainly from old exams, with answer, (iii) an applied exercise to be done during the first term in your own time.

Section 3 contains advice on how to do your applied econometric project (which has to be handed in after Easter). Most of the techniques that you will need for your project will be taught in the first term.

The rest of the booklet contains notes which roughly follow the order of the lectures.

The course emphasises actually doing applied econometrics. The skill to combine economic theory, statistical methods and an understanding of the data with the ability to use the appropriate software is something that is learnt through experience rather than taught in lectures. The applied exercise in Section 2 is designed to give you that experience. It is essential that you start it as soon as possible. There will be classes in using econometric packages on the computer early in the Autumn term. The applied exercise contains a lot of information that we will assume that you have learned and the theory in the lectures will make a lot more sense if you have seen how it is used. It is essential that you know how to use an econometric package, either EViews or any other one you wish, by the middle of the Autumn term. We assume that you have done the applied exercise and you will notice that lots of questions using these data appear on past exam questions.

These notes cover the material taught in the Autumn term. In the Spring term there will be one lecture and a class each week covering more advanced and more applied topics.

### 1.6. Provisional Outline Autumn Term

By Weeks:

1. Least Squares and the Linear Regression Model (LRM).

2. Maximum Likelihood estimation of the LRM.
3. Test procedures, Asymptotic and Exact.
4. Specification and Diagnostic Tests.
5. Univariate Stochastic Processes: ARIMA and unit roots
6. ARDL models.
7. Cointegration.
8. Vector Autoregressions and Johansen estimation of Cointegrating Vectors.
9. Endogenous regressors and Instrumental Variable Estimation.
10. Applications and Revision.

### 1.7. Reading

The lectures do not follow any particular text, but you should use a text book to supplement the lectures and to cover issues in your project that are not in the lectures. There are a large number of good texts, choose one that uses matrix algbera. Use the index to locate topics, they are covered in different places in different texts. In most cases you do not need the latest edition of the text. Some are referred to below, but there are many other good texts. The book that is probably closest to this course is Verbeek, (2008). Students find Kennedy (2003) very useful. It is not a textbook, it leaves out all the derivations, but it has a brilliant explanation of what econometrics is all about. The 5th edition has an excellent chapter on doing applied econometrics. Greene (2008) is a very comprehensive text. If your current or future work involves econometrics, you will need a textbook for work and Greene is a good general reference. Ruud, P.A. (2000) *An Introduction to Classical Econometric Theory*, Oxford is another good text. If you have not done any econometrics before you might want to start with an introductory text that does not use matrix algebra, like Stock and *Watson* (2003) or many others.

Maddala and Kim, (1998) is a good introduction to time-series issues, but does not cover a range of other econometric topics that we will deal with. Enders (2005) is a more general applied time-series text and Patterson (2000) has many applications. Although a little dated, Hamilton (1994) remains the best

advanced time-series text. The approach to time-series in this course is loosely based on the Hendry methodology, Hendry (1995) provides the authorised version. An early exposition of the methodology is Spanos (1986), which is based on the econometrics course he gave here at Birkbeck.

Angrist and Pischke (2009) is an excellent accessible treatment of microeconometrics. Wooldridge (2002) covers cross-section and panel estimation. Favero (2001) and Canova (2007) are very good at linking the macroeconomic theory to the econometrics. Cuthbertson (1996) has good finance examples.

You should also read applied econometric articles. The Journal of Economic Literature, Journal of Economic Perspectives and the Journal of Economic Surveys are good places to look.

# References


Angrist, Joshua D. and Jorn-Steffen Pischke, (2009) *Mostly Harmless Econometrics*, Princeton.

Canova, Fabio (2007) *Methods for Applied Macroeconomic Research*, Princeton.

Cuthbertson, Keith (1996) *Quantitative Financial Economics*, Wiley.

Enders, W (2005) *Applied Econometric Time Series,* 2nd edition Wiley.

Favero, Carlo (2001) *Applied Macroeconometrics*, Oxford.

Greene, William (2008) *Econometric Analysis*, 6th edition, Prentice Hall.

Hamilton, James D (1994), *Time Series Analysis* (Princeton/Wiley).

Hendry, David (1995) *Dynamic Econometrics,* Oxford.

Kennedy, Peter (2003) *A Guide to Econometrics* 5th edition Blackwell.

Maddala G S and In-Moo Kim, (1998) *Unit Roots, Cointegration and Structural Change*, Cambridge.

Patterson, Kerry (2000) *An Introduction to Applied Econometrics, a time series approach* Macmillan.

Ruud, P.A. (2000) *An Introduction to Classical Econometric Theory*, Oxford.

Spanos, Aris (1986) *Statistical Foundations of Econometric Modelling*, Cambridge University Press.

Stock and *Watson (2003) Introduction to Econometrics,* Addison Wesley.

Verbeek, Marno (2008) *A Guide to Modern Econometrics*, 3rd edition , Wiley.

Wooldridge, Jeffrey (2002) *Econometric Analysis of Cross-section and panel data*, MIT Press.

## 2. Exercises

### 2.1. Class exercises

These exercises should be attempted before coming to class. Unless you try to do them yourself you will not know what problems you are having.

#### 2.1.1. Week 1, No class

#### 2.1.2. Week 2.

(a) Explain the following terms used to describe estimators:

Conditional Expectation; Least Squares; Maximum Likelihood; Unbiased; Minimum Variance; Consistent; Asymptotically Normal.

(b) Suppose we have the quadratic form $S = \beta' A \beta$ where $\beta$ is a $2 \times 1$ vector and $A$ a symmetric $2 \times 2$ matrix:

$$S = [\beta_1, \beta_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

expand the quadratic form in terms of scalars, take the derivatives with respect to $\beta_1$ and $\beta_2$ and show that $\partial S / \partial \beta = 2A\beta$.

(c) We will not do this in the lectures, but it can be found in any econometrics text. Consider the Linear Regression Model

$$y = X\beta + u$$

where $y$ is a $T \times 1$ vector of observations on a dependent variable; $X$ a $T \times k$ rank $k$ matrix of observations on non-stochastic exogenous variables; $u$ a $T \times 1$ vector of unobserved disturbances with $E(u) = 0$, $E(uu') = \sigma^2 I$, and $\beta$ a $k \times 1$ vector of unknown coefficients. The Least Squares estimator of $\beta$ is $\widehat{\beta} = (X'X)^{-1}X'y$. The Gauss-Markov theorem is that $\widehat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) of $\beta$. Prove this and explain what role each of the assumptions play in the proof. How would the derivation change if the exogenous variables $X$ were stochastic, but distributed independently of the errors.

#### 2.1.3. Week 3

Suppose for $t = 1, 2, ..., T$

$$Y_t = \alpha + u_t$$

where the observations are independent and the distribution of $Y_t$ is given by

$$f(Y_t) = (2\pi\sigma^2)^{-1/2} \exp -\frac{1}{2}(\frac{Y_t - \alpha}{\sigma})^2$$

(a) What is the log-likelihood function and the score vector? Derive the maximum likelihood estimators of $\theta = (\alpha, \sigma^2)$

(b) Derive the information matrix and the asymptotic variance-covariance matrix.

(c) How would you estimate the standard error of the maximum likelihood estimator of $\alpha$?

(d) Compare your derivations to the matrix form for the linear regression model in the notes.

### 2.1.4. Week 4.

Consider the Linear Regression Model

$$y = X\beta + u$$

where $y$ is a $T \times 1$ vector of observations on a dependent variable; $X$ a $T \times k$ rank $k$ matrix of observations on non-stochastic exogenous variables; $u$ a $T \times 1$ vector of unobserved disturbances with $E(u) = 0$, $E(uu') = \sigma^2 I$, and $\beta$ a $k \times 1$ vector of unknown coefficients.

(a) Derive the estimator $\widehat{\beta}$ that makes $X'\widehat{u} = 0$, where $\widehat{u} = y - X\widehat{\beta}$.

(b) Define $P_X = X(X'X)^{-1}X'$ and $M = I_T - P_X$. Show: (i) $MM = M$, (ii) $MP_X = 0$.

(c) Show that $\widehat{u} = My = Mu$.

(d) Show that $E(\widehat{u}'\widehat{u}) = (T - k)\sigma^2$.

(e) Suppose that $E(uu') = \sigma^2\Omega$. (i) What is $E(\widehat{\beta} - E(\widehat{\beta}))((\widehat{\beta} - E(\widehat{\beta}))'$? (ii) Derive the estimator $\widetilde{\beta}$ that makes $X'\Omega^{-1}\widetilde{u} = 0$, where $\widetilde{u} = y - X\widetilde{\beta}$.

### 2.1.5. Week 5

(a) In the Linear Regression Model in the week 2 exercise, suppose that the disturbances are also normally distributed and there are $k$ prior restrictions of the form $\beta - q = 0$, where $q$ is a known vector of order $k \times 1$. Derive a test statistic to test these restrictions. Explain how you would calculate the restricted and unrestricted sums of squares to carry out the test.

(b) The following equations were estimated on 24 observations 1918-1941, where $D_t$ is dividends and $E_t$ is earnings in year t. Standard errors are given in parentheses, SSR is sum of squared residuals, MLL is maximised log likelihood.

$$\begin{array}{lllll} D_t = & 0.59 & +0.40E_t & & SSR = 2.1849 \\ & (0.20) & (0.10) & & MLL = -5.297 \\ D_t = & -0.14 & +0.32E_t & -0.10E_{t-1} \quad +0.70D_{t-1} & SSR = 0.84821 \\ & (0.17) & (0.08) & (0.10) \quad (0.14) & MLL = 6.0576 \end{array}$$

Test the hypothesis that the coefficient of earnings in the first equation (i) equals one (ii) equals zero.

Test the hypotheses that the coefficients of lagged earnings and dividends in the second equation equal zero, (i) individually (ii) jointly. For the joint test use both F and LR tests.

Suppose the coefficients in the second equation are labelled $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$. Write the restrictions that the coefficients of lagged earnings and dividends equal zero in the form $R\beta - q = 0$.

### 2.1.6. Week 6

1. Explain what effect the following 'problems' have on the properties of the least squares estimates of the coefficients and their standard errors. How would you detect whether each problem was present:

(a) Heteroskedasticity.

(b) Serial correlation.

(d) Non-normality.

(e) Non-linearity.

(f) Exact multicollinearity.

2. Explain the following quote, from Angrist and Pischke (2009) p223. Do you agree with it? 'We prefer fixing OLS standard errors to GLS. GLS requires even stronger assumptions than OLS, and the resulting asymptotic efficiency gain is likely to be modest, while finite sample properties may be worse.'

### 2.1.7. Week 7

Consider the following models estimated over a sample $t = 1, 2, ..., T$. In each case $\varepsilon_t$ is white noise and $\rho$ and $\mu$ are less than one in absolute value.

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t \tag{1}$$

$$y_t = \alpha + y_{t-1} + \varepsilon_t \tag{2}$$

$$y_t = \alpha + \varepsilon_t + \mu\varepsilon_{t-1} \tag{3}$$

$$\Delta y_t = \alpha + \rho\Delta y_{t-1} + \varepsilon_t + \mu\varepsilon_{t-1} \tag{4}$$

(a) Suppose you had estimates of the parameters, how would you forecast $y_{T+1}$ and $y_{T+2}$ in each case, given data up to $y_T$?

(b) In cases (1) and (2) substitute back to remove the $y_{t-i}$.

(c) For cases (1) and (3) what is the expected value of $y_t$?

(d) For (1) and (3) derive the variance, covariances and autocorrelations of the series.

(e) For which of the models is $y_t$ I(1)?

### 2.1.8. Week 8

The Augmented Dickey Fuller Tests for non-stationarity uses a regression of the form:

$$\Delta y_t = \alpha + \beta y_{t-1} + \gamma t + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t$$

(a) What is the null hypothesis tested; what is the test statistic and what is its 95% critical value?

(b) Suppose all the $\delta_i = 0$. Substitute back to express $y_t$ in terms of $\varepsilon_{t-i}$, $t$ and $y_0$. Compare the cases $\beta < 0$ and $\beta = 0$.

(c) What is the rationale behind the $\sum_{i=1}^{p} \delta_i \Delta y_{t-i}$ term.

(d) using the Shiller data test whether the log price-earnings ratio, $\log(NSP/NE)$, has a unit root (i) over 1950-2000, (ii) over the whole sample; using intercept and no trend and using the AIC to choose $p$.

### 2.1.9. Week 9.

Consider the general model

$$d_t = \alpha_0 + \alpha_1 d_{t-1} + \beta_0 e_t + \beta_1 e_{t-1} + \gamma_0 p_t + \gamma_1 p_{t-1} + u_t$$

where $d_t$ is log nominal dividends, $e_t$ is log nominal earnings and $p_t$ is the log of the producer price index.

(a) How would you calculate the long-run elasticity of dividends to prices and earnings, $\theta_i$ in:

$$d_t^* = \theta_0 + \theta_1 e_t + \theta_2 p_t.$$

.

(b) How would you estimate the model subject to the restrictions that the long-run elasticity to earnings is unity and to prices zero.

(c) Suppose it was believed that the appropriate model was

$$d_t = \alpha + \beta e_t + \gamma p_t + v_t;\ v_t = \rho v_{t-1} + \varepsilon_t,$$

where $\rho$ is less than one in absolute value. Show that this is a restricted form of the general model and derive the two restrictions.

(d) Suppose it was believed that the adjustment to the long-run relationship was given by:

$$\Delta d_t = \lambda_1 \Delta d_t^* + \lambda_2 (d_{t-1}^* - d_{t-1}) + u_t$$

what restriction does this impose on the general model.

(e) Using the Shiller data, estimate the models over the period 1950-1986 and test the three sets of restrictions.

### 2.1.10. Week 10.

Consider the VAR

$$y_t = A_0 + A_1 y_{t-1} + \varepsilon_t;\ t = 1, 2, ..., T$$

where $y_t = (y_{1t}, y_{2t})'$, $A_0$ is a $2x1$ vector, $(a_o, a_1)'$,

$$A_1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \varepsilon_t \ \tilde{}\ N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix} \right)$$

(a) Write down each equation of the VAR and explain how you would estimate the coefficients and covariance matrix.

(b) Given estimates of the parameters, how would you forecast $y_{T+1}$ and $y_{T+2}$?

(c) What condition is required for $y_{1t}$ to be Granger non-causal with respect to $y_{2t}$?

(d) Write the VAR as

$$\Delta y_t = A_0 + \Pi y_{t-1} + \varepsilon_t;\ t = 1, 2, ..., T.$$

Explain the relation between $A_1$ and $\Pi$.

(e) What are the implications for $\Pi$ if $y_{it}$ are (i) I(0); (ii) I(1) and cointegrated; (iii) I(1) and not cointegrated? What restrictions does case (iii) put on $A_1$?

(f). Suppose the $y_{it}$ are I(1) with cointegrating relationship $z_t = y_{1t} - \beta y_{2t}$. Write out the restricted system and explain the restrictions this imposes on $\Pi$. Show that it has rank one.

(g) Derive the parameters of the ARDL(1,1) model:

$$y_{1t} = \alpha_0 + \beta_0 y_{2t} + \beta_1 y_{2t-1} + \alpha_1 y_{1t-1} + u_t$$

from the VAR. Hint note that $E(\varepsilon_{1t} \mid \varepsilon_{2t}) = (\omega_{12}/\omega_{22})\varepsilon_{2t}$ and use this in the first equation of the VAR substituting for $\varepsilon_{2t}$.

### 2.2. Example questions with answers (based on old exams).

#### 2.2.1. LRM

Consider the linear regression model

$$y = X\beta + u,$$

where $y$ is a $T \times 1$ vector of observations on a dependent variable; $X$ is a $T \times k$ full-rank matrix of observations on non-stochastic, exogenous variables; $u$ is a $T \times 1$ vector of unobserved disturbances with $E(u) = 0$ and $E(uu') = \sigma^2 I$; and $\beta$ is a $k \times 1$ vector of unknown coefficients.

(a) Derive the Ordinary Least Squares estimator of $\beta$, say $\widehat{\beta}$.

(b) Prove that $\widehat{\beta}$ has minimum variance in the class of linear unbiased estimators of $\beta$.

(c) Define the least squares residuals as $\widehat{u} = y - X\widehat{\beta}$. Show that $X'\widehat{u} = 0$.

(d) Show that $\widehat{u} = My = Mu$, where $M = I - X(X'X)^{-1}X'$.

**Answer**

(a) The sum of squared residuals is

$$\begin{aligned} u'u &= (y - X\beta)'(y - X\beta) \\ &= y'y + \beta' X'X\beta - 2\beta' X'y \end{aligned}$$

the first order condition is

$$\begin{aligned} \frac{\partial u'u}{\partial \beta} &= 2X'X\beta - 2X'y = 0 \\ X'X\beta &= X'y \\ \widehat{\beta} &= (X'X)^{-1}X'y \end{aligned}$$

(b) Consider another linear estimator

$$\begin{aligned}\widetilde{\beta} &= Ly = (C + (X'X)^{-1}X')y \\ &= (C + (X'X)^{-1}X')(X\beta + u) \\ &= CX\beta + \beta + Cu + (X'X)^{-1}X'u \\ E(\widetilde{\beta}) &= CX\beta + \beta\end{aligned}$$

so this will only be unbiased if $CX = X'C' = 0$ assume this is so. Its variance is

$$\begin{aligned}E(\widetilde{\beta} - \beta)(\widetilde{\beta} - \beta)' &= E\left\{(C + (X'X)^{-1}X')u\right\}\left\{(C + (X'X)^{-1}X')u\right\}' \\ &= E\left\{\begin{array}{c} Cu\ u'C' + (X'X)^{-1}X'u\ u'X(X'X)^{-1} \\ +(X'X)^{-1}X'u\ u'C' + Cuu'X(X'X)^{-1} \end{array}\right\} \\ &= \sigma^2\left\{CC' + (X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'C' + CX(X'X)^{-1}\right\} \\ &= \sigma^2\left\{CC' + (X'X)^{-1}\right\}\end{aligned}$$

since $CX = X'C' = 0$. $CC'$ is a positive definite matrix for $C \neq 0$ so $V(\widetilde{\beta}) > V(\widehat{\beta}) = \sigma^2(X'X)^{-1}$.

(c)

$$\begin{aligned}X'\widehat{u} &= X'(y - X(X'X)^{-1}X'y) \\ &= X'y - X'X(X'X)^{-1}X'y \\ &= X'y - X'y = 0\end{aligned}$$

(d)

$$\begin{aligned}\widehat{u} &= (y - X(X'X)^{-1}X'y) \\ &= (I - X(X'X)^{-1}X')y = My \\ &= (I - X(X'X)^{-1}X')(X\beta + u) \\ &= X\beta + u - X(X'X)^{-1}X'X\beta - X(X'X)^{-1}X'u \\ &= X\beta - X\beta + u - X(X'X)^{-1}X'u \\ &= (I - X(X'X)^{-1}X')u = Mu\end{aligned}$$

### 2.2.2. Diagnostic Tests

Consider the linear regression model

$$y_t = \beta' x_t + u_t, \quad t = 1, 2, ..., T,$$

where $y_t$ is an observation on a dependent variable at time $t$; $x_t$ is a $k \times 1$ vector of observations on some exogenous regressors; $u_t$ is an unobserved disturbance, and $\beta$ is a $k \times 1$ vector of unknown coefficients. For each of the 'problems' listed below: (i) explain the consequences of the problem for the properties of the least squares estimates of $\beta$ and their standard errors; (ii) explain how you would test for the presence of the problem.

(a) Fourth-order serial correlation of the disturbance $u_t$;

(b) Heteroskedasticity;

(c) Non-linearity;

(d) A shift in the variance of the errors at a known time $T1$, with $k < T1 < T - k$;

(e) A shift in the coefficients at a known time $T1$, with $k < T1 < T - k$

**Answer**

a) (i) $\widehat{\beta}$ is unbiased but not minimum variance, its standard errors are biased (ii) estimate either

$$\widehat{u}_t = \sum_{i=1}^{4} \rho_i \widehat{u}_{t-i} + b' x_t + e_t$$

and test $\rho_i = 0$ (with a Chi squared or F test) for up to fourth order serial correlation or estimate

$$\widehat{u}_t = \rho \widehat{u}_{t-4} + b' x_t + e_t$$

and test $\rho = 0$ using a t test for just fourth order serial correlation.

(b)(i) $\widehat{\beta}$ is unbiased but not minimum variance, its standard errors are biased (ii) it depends on the form of the heteroskedasticity, run

$$\widehat{u}_t^2 = a + b' z_t + e_t$$

and test $b = 0$. Possible choices for $z_t$ are the regressors, their squares, their squares and cross-products, powers of the fitted values.

(c) (i) $\widehat{\beta}$ is a biased and inconsistent estimator of the true parameters (ii) it depends on the form of the non-linearity but running

$$\widehat{u}_t = a + b\widehat{y}_t^2 + e_t$$

where $\widehat{y}_t = \widehat{\beta} x_t$ and testing $b = 0$ may have power against squares and interaction terms.

(d) This is a particular form of heteroskedasticity (i) $\widehat{\beta}$ is unbiased but not minimum variance, its standard errors are biased (ii) estimate the equation $t = 1, 2, ..., T1$ and get residuals $\widehat{u}_1$, then $T1 + 1, ..., T$ and get residuals $\widehat{u}_2$ then

$$\frac{\widehat{u}_1'\widehat{u}_1/(T1-k)}{\widehat{u}_2'\widehat{u}_2/(T2-k)} \sim F(T1-k, T2-k)$$

where $T2 = T - T1$.

(e) (i) $\widehat{\beta}$ is a biased and inconsistent estimator of the true changing parameters, (ii) assuming the variance is constant and using the estimates in (d) and calling the residuals from the estimates over the whole period $\widehat{u}$

$$\frac{(\widehat{u}'\widehat{u} - \widehat{u}_1'\widehat{u}_1 - \widehat{u}_2'\widehat{u}_2)/k}{(\widehat{u}_1'\widehat{u}_1 + \widehat{u}_2'\widehat{u}_2)/T - 2k} \sim F(k, T-2k)$$

### 2.2.3. Structural Stability

UK data 1979Q4-2003Q4 were used to explain the growth rate of GDP, $\Delta y_t$, by the growth of trade weighted foreign GDP, $\Delta y_t^*$, the lagged long-term interest rate, $r_{t-1}$, and lagged growth, $\Delta y_{t-1}$. This was estimated over the whole period and two sub-periods split at 1992Q3.

| | 79Q4-03Q4 | 79Q4-92Q3 | 92Q3-03Q4 |
|---|---|---|---|
| $int$ | 0.006 | 0.020 | 0.000 |
| $(se)$ | (0.002) | (0.007) | (0.002) |
| $\Delta^* y_t$ | 0.569 | 0.459 | 0.585 |
| $(se)$ | (0.143) | (0.221) | (0.142) |
| $r_{t-1}$ | -0.221 | -0.670 | 0.239 |
| $(se)$ | (0.073) | (0.222) | (0.151) |
| $\Delta y_{t-1}$ | 0.135 | 0.071 | -0.006 |
| $(se)$ | (0.088) | (0.118) | (0.147) |
| $SERx100$ | 0.559 | 0.686 | 0.282 |
| $SSRx100$ | 0.288 | 0.226 | 0.03253 |
| $MLL$ | 367.987 | 187.361 | 202.487 |
| $SC$ | 0.017 | 0.421 | 0.910 |
| $Het$ | 0.025 | 0.266 | 0.337 |
| $NObs$ | 97 | 52 | 45 |

Standard errors of coefficients are given in the $(se)$ rows, $SER$ is the standard error of regression, $SSR$ is the sum of squared residuals, $MLL$ is the Maximised

Log Likelihood, $SC$ and $Het$ give p values for tests for serial correlation and heteroskedasticity, $NObs$ is the number of observations.

(a) Test the hypothesis that the regression parameters are the same in both periods. The 5% critical value of F(4,89) is 2.47.

(b) Test the hypothesis that the variances are the same in the two periods. The 5% value of F(48,41) is 1.64.

(c) What light does your answer to (b) shed on your answer to (a).

(d) What evidence for misspecification is there in the three equations?

(e) Comment on the differences between the estimates for the two periods.

(f) [ Do you think that this equation could be interpreted as an IS curve?

**Answer**

(a) Chow test

$$\frac{(0.288 - (0.226 + 0.0325))/4}{(0.226 + 0.0325)/(97 - 8)} = 2.539$$

greater than the critical value, reject $H_0$ that the parameters are equal.

(b) Variance Ratio test

$$\left(\frac{0.686}{0.282}\right)^2 = 5.92$$

Reject $H_0$ that the variances are equal.

(c) The test in (a) is only valid if the variances are equal, thus the result may not be reliable.

(d) The whole period equation fails serial correlation and heteroskedasticity tests, those for the two sub periods do not. Thus it is likely that the serial correlation was induced by the parameter change and the heteroskedasticity by the change in variance in the two periods.

(e) Given the size of the standard errors and the similarity of the coefficients the effect of foreign output growth did not change between the periods. The coefficient of lagged interest rate went from being signficantly negative to being insignficantly positive. The coefficient of lagged growth is always insignificant. The big difference is that the variance is much smaller in the second period, during the 'Great moderation', the economy was much more stable.

(f) It is a sort of IS curve, but the IS curve usually has the level of output (or the output gap) rather than the growth rate as the dependent variable and the real interest rate rather than the nominal interest rate as the independent variable.

### 2.2.4. ARIMA

Call the logarithm of the real Standard and Poor Stock price index $y_t$. The following models were estimated over the period 1873-1986. Numbers in parentheses below coefficients are estimated standard errors and numbers in brackets are the maximised log-likelihoods, $\varepsilon_t$ is assumed to be a white-noise error in each case.

$$
\begin{array}{llcccccll}
A. & \Delta y_t & = & \underset{(0.0167)}{0.0162} & & & & + \varepsilon_t & [35.3726] \\
B. & \Delta y_t & = & \underset{(0.0181)}{0.0164} & & \underset{(0.83)}{+0.098\ \varepsilon_{t-1},} & & + \varepsilon_t & [35.7223] \\
C. & \Delta y_t & = & \underset{(0.0168)}{0.1523} & & & \underset{(0.0947)}{+0.0641\ \Delta y_{t-1}} & + \varepsilon_t & [35.6056] \\
D. & \Delta y_t & = & \underset{(0.0294)}{0.0270} & & \underset{(0.113)}{+0.822\ \varepsilon_{t-1}} & \underset{(0.149)}{-0.642\ \Delta y_{t-1}} & + \varepsilon_t & [37.8298] \\
E. & \Delta y_t & = & \underset{(0.102)}{-0.266} & \underset{(0.0008)}{+0.002\ t} & \underset{(0.046)}{-0.135\ y_{t-1}} & \underset{(0.095)}{+0.131\ \Delta y_{t-1}} & + \varepsilon_t & [39.9282]
\end{array}
$$

(a) Use model E to test for the presence of a unit root in $y_t$. The 5% critical value for the Augmented Dickey–Fuller test with trend is -3.4494.

(b) In model D, test at the 5% level whether the autoregressive and moving average coefficients are significant: (i) individually; (ii) jointly. Explain the conflict between the individual and joint results.

(c) Briefly indicate how model D can be used to make a two-period ahead forecast, given data up to 1986.

(d) Which of these models would you choose on the basis of the Akaike Information Criterion?

**Answer**

(a) The ADF statistic is -0.135/0.046=-2.93, Do not reject the null of a unit root.

(b) (i) individually: 0.822/0.113=7.27; -0.642/0.149=-4.31 so both are individually significant. (ii) jointly 2(37.8298-35.3726)=4.91 which is less than the 5% critical value of 5.99 so jointly they are not significant. If the true model is

$$\Delta y_t = \alpha + \varepsilon_t$$

and we multiply by $1 - \rho L$ where $L$ is the lag operator we get

$$\begin{aligned}(1-\rho L)\,\Delta y_t &= (1-\rho L)\,(\alpha+\varepsilon_t)\\ \Delta y_t &= (1-\rho)\,\alpha+\rho\Delta y_{t-1}+\varepsilon_t-\rho\varepsilon_{t-1}\end{aligned}$$

in this example the AR and MA coefficients are of opposite signs and similar order of magnitudes, so cancel out.

(c) Call 1986 $T$ then the forecasts are

$$\begin{aligned}\Delta y^f_{1987} &= 0.0270+0.822\widehat{\varepsilon}_{1986}-0.642\Delta y_{1986}\\ \Delta y^f_{1988} &= 0.0270-0.642\Delta y^f_{1987}\end{aligned}$$

(d) AIC $MLL_i - k_i$ chooses model E

| $Model$ | $MLL$ | $k$ | $AIC$ |
|---|---|---|---|
| $A$ | 35.3726 | 1 | 34.3726 |
| $B$ | 35.7223 | 2 | 33.7223 |
| $C$ | 35.6056 | 2 | 33.6056 |
| $D$ | 37.8298 | 3 | 34.8298 |
| $E$ | 39.9282 | 4 | 35.9282 |

### 2.2.5. VAR

Consider the first-order Vector Autoregression, VAR, for $\{(y_t, x_t); t = 1, 2, ..., T\}$:

$$\begin{aligned}y_t &= a_{10}+a_{11}y_{t-1}+a_{12}x_{t-1}+\varepsilon_{1t},\\ x_t &= a_{20}+a_{21}y_{t-1}+a_{22}x_{t-1}+\varepsilon_{2t},\end{aligned}$$

$$\begin{bmatrix}\varepsilon_{1t}\\ \varepsilon_{2t}\end{bmatrix}\sim \text{NID}\left(\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\right).$$

(a) How would you estimate the coefficients and the error variance–covariance matrix of the VAR?

(b) Under what restriction is $y_t$ Granger non-causal for $x_t$?

(c) Derive the parameters of the Autoregressive Distributed Lag, ARDL(1,1), model

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

from the parameters of the VAR. What is the variance of $u_t$?

(d) Suppose $y_t$ and $x_t$ were each I(1) and $z_t = y_t - \beta x_t$ was I(0). What implications does this have for the VAR?

**Answer**

(3) (a) Estimate each equation by OLS getting residuals $\widehat{\varepsilon}_{it}$ then

$$\widehat{\sigma}_{ij} = (T-3)^{-1} \sum \widehat{\varepsilon}_{it} \widehat{\varepsilon}_{jt}; i, j = 1, 2.$$

$T^{-1}$ also acceptable.

(b) If $a_{21} = 0$ $y_t$ is Granger non-causal for $x_t$.

(c)

$$\begin{aligned}
E(y_t \mid x_t, y_{t-1}, x_{t-1}) &= a_{10} + a_{11} y_{t-1} + a_{12} x_{t-1} + E(\varepsilon_{1t} \mid x_t, y_{t-1,} x_{t-1}) \\
E(\varepsilon_{1t} \mid x_t, y_{t-1,} x_{t-1}) &= \sigma_{12} \sigma_{22}^{-1} \varepsilon_{2t} \\
&= \sigma_{12} \sigma_{22}^{-1} \left( x_t - a_{20} - a_{21} y_{t-1} - a_{22} x_{t-1} \right) \\
E(y_t \mid x_t, y_{t-1}, x_{t-1}) &= a_{10} + a_{11} y_{t-1} + a_{12} x_{t-1} + \sigma_{12} \sigma_{22}^{-1} \left( x_t - a_{20} - a_{21} y_{t-1} - a_{22} x_{t-1} \right) \\
y_t &= \left( a_{10} - \sigma_{12} \sigma_{22}^{-1} a_{20} \right) + \sigma_{12} \sigma_{22}^{-1} x_t + \left( a_{11} - \sigma_{12} \sigma_{22}^{-1} a_{21} \right) y_{t-1} + \left( a_{12} - \sigma_{12} \sigma_{22}^{-1} a_{22} \right) x_{t-1} + u_t
\end{aligned}$$

$$\begin{aligned}
u_t &= \varepsilon_{1t} - \sigma_{12} \sigma_{22}^{-1} \varepsilon_{2t} \\
E(u_t^2) &= \sigma_{11} + \left( \sigma_{12} \sigma_{22}^{-1} \right)^2 \sigma_{22} - 2 \left( \sigma_{12} \sigma_{22}^{-1} \right) \sigma_{12}
\end{aligned}$$

(d) There would be a cross-equation restriction that comes from the cointegrating relationship and it would be estimated as a VECM

$$\begin{aligned}
\Delta y_t &= a_{10} + a_1 z_{t-1} + \varepsilon_{1t} \\
\Delta x_t &= a_{20} + a_2 z_{t-1} + \varepsilon_{2t}
\end{aligned}$$

### 2.2.6. Cointegration 1

A second-order cointegrating vector error-correction model (VECM), with unrestricted intercepts and restricted trends, was estimated on quarterly US data from 1947Q3 to 1988Q4. The variables included were the logarithm of real consumption ($c_t$), the logarithm of real investment ($i_t$), and the logarithm of real income ($y_t$). The Johansen maximal eigenvalue tests for, $r$, the number of cointegrating vectors, were:

| $H_o$ | $H_1$ | *Statistic* | 10%*CV* |
|---|---|---|---|
| $r = 0$ | $r = 1$ | 34.6 | 23.1 |
| $r \leq 1$ | $r = 2$ | 15.8 | 17.2 |
| $r \leq 2$ | $r = 3$ | 3.3 | 10.5 |

The Johansen Trace Tests were:

| $H_o$ | $H_1$ | *Statistic* | 10%*CV* |
|---|---|---|---|
| $r = 0$ | $r \geq 1$ | 53.7 | 39.3 |
| $r \leq 1$ | $r \geq 2$ | 19.1 | 23.1 |
| $r \leq 2$ | $r = 3$ | 3.3 | 10.5 |

Assuming that $r = 2$, the following two just-identified cointegrating vectors $Z1_t$ and $Z2_t$ (standard errors in parentheses) were estimated:

| $c$ | $i$ | $y$ | $t$ |
|---|---|---|---|
| 1 | 0 | $-1.13$ | 0.0003 |
| | | (0.16) | (0.0006) |
| 0 | 1 | $-1.14$ | 0.0007 |
| | | (0.26) | (0.001) |

The system maximised log-likelihood (MLL) was 1552.9. The system was then estimated subject to the over-identifying restrictions that: (i) both coefficients of income were unity, giving a MLL of 1552.3; and (ii) not only were the income coefficients unity, but that the trend coefficients were also zero, giving a MLL of 1548.1.

The Vector Error Correction Estimates [t statistics] for the just identified system (constants included but not reported) were

| | $\Delta c_t$ | $\Delta i_t$ | $\Delta y_t$ |
|---|---|---|---|
| $Z1_{t-1}$ | 0.075068 | 0.262958 | 0.192686 |
| | [2.74240] | [ 3.20914] | [ 4.63684] |
| $Z2_{t-1}$ | −0.011232 | −0.171416 | 0.009323 |
| | [−0.67114] | [−3.42157] | [ 0.36694] |
| $\Delta c_{t-1}$ | −0.209469 | −0.171819 | 0.094535 |
| | [−2.31259] | [−0.63368] | [ 0.68749] |
| $\Delta i_{t-1}$ | 0.022574 | 0.334330 | 0.156990 |
| | [0.72374] | [ 3.58069] | [ 3.31537] |
| $\Delta y_{t-1}$ | 0.212411 | 0.697502 | 0.126186 |
| | [3.17484] | [ 3.48267] | [ 1.24236] |
| $R^2$ | 0.146291 | 0.405637 | 0.320507 |
| $SER$ | 0.007527 | 0.022533 | 0.011427 |

(a) How many cointegrating vectors do the tests indicate?

(b) If there are $r$ cointegrating vectors, how many restrictions on each vector do you need to identify it.

(c) Interpret the just identifying restrictions used above.

(d) Test the two sets of overidentifying restrictions. 5% $\chi^2(2) = 5.99, \chi^2(4) = 9.49$.

(e) The VECM was estimated with unrestricted intercepts and restricted trends. What does this mean?

(f) Do you think investment is Granger Causal for Consumption.

**Answer**

(a) one (b) $r$

(c) Investment does not appear in the consumption function and consumption does not appear in the investment function.

(d) (i) 2(1552.9-1552.3)=1.2<$\chi^2(2)$, do not reject $H_0$ (ii) 2(1552.9-1548.1)=9.6>$\chi^2(4)$ reject $H_0$ .

(e) Write the VECM

$$\Delta y_t = \mu + \alpha(\beta y_{t-1} + \gamma t) + u_t$$

where the intercepts $\mu$ lie outside the error correction term and the trends $\gamma t$ are restricted to lie within it. If $y_t$ is a $mx1$ vector, whereas one estimates $m$ intercepts, one only estimates $r$ trend coefficients, giving $m - r$ restriction.

(f) The fact that both $Z2_{t-1}$ (which is a function of lagged investment) and $\Delta i_{t-1}$ are individually insignificant in the consumption equation suggests that

investment may be Granger non-causal for consumption, though the two terms could be jointly signficant.

### 2.2.7. Cointegration 2

Let $y_t$ be the logarithm of GDP, $r_t$ the short-term interest rate, and $p_t$ the logarithm of the price level. Using UK data from 1964Q3 to 1998Q2, Vector Autoregressions (VARs) of order zero to 4 were estimated for $\Delta y_t$, $\Delta r_t$ and $\Delta^2 p_t$, each equation including an intercept. The maximised log-likelihoods, $MLL_i$, for each order were 0:1591.4, 1:1612.6, 2:1623.0, 3:1635.3, 4:1644.5. For a fourth-order VAR, likelihood ratio tests of the non-causality of $\Delta r_t$ with respect to the other two variables gave a test statistic of 31.9 and of the non-causality of $\Delta^2 p_t$ with respect to the other two variables 8.4. For a fourth-order VAR, with unrestricted intercepts and no trends, the Johansen trace test statistics for the number of cointegrating vectors, $r$, and their 95% critical values in parentheses were: $r = 0 : 61\ (21)$; $r \leq 1 : 37\ (15)$; $r \leq 2 : 22\ (8)$.

(a) On the basis of the Akaike Information Criterion, what would be the preferred order of the VAR?

(b) Explain how the Granger non-causality tests are constructed. What are the degrees of freedom of the tests? Interpret the results.

(c) How many cointegrating vectors do there seem to be in this data? Interpret your conclusion. Explain what unrestricted intercept means.

(d) Interpret the variables in the VAR and explain what relationship you would expect between them.

**Answer**

(a) AIC $MLL_i - k_i$ chooses 4

| $Lag$ | $MLL$ | $k$ | $AIC$ |
|---|---|---|---|
| 0 | 1591.4 | 3 | 1588.4 |
| 1 | 1612.6 | 12 | 1600.6 |
| 2 | 1623.0 | 21 | 1602.0 |
| 3 | 1635.3 | 30 | 1605.3 |
| 4 | 1644.5 | 39 | 1605.5 |

(b) To test Granger causality of $\Delta r_t$ with respect to the other two variables, the restricted model sets all the coefficients of $\Delta r_{t-i}$ to zero in the equations for the other two variables $\Delta y_t$ and $\Delta^2 p_t$. This is four restrictions in each of the two equations so each test has 8 degrees of freedom. Similarly for $\Delta^2 p_t$. 31.9 is greater

than 15.51 so reject the hypothesis that $\Delta r_t$ in Granger non causal with respect to the other two. 8.4 is less than the critical value so do not reject the hypothesis that $\Delta^2 p_t$ is non causal. Changes in interest rates help predict the change in inflation and growth, the change in inflation does not help predict change in interest rates and growth.

(c) There are three cointegrating vectors, each of the variables is $I(0)$ as you would expect, given that they are changes in possibly I(1) variables. The unrestricted intercept means that it is not restricted to lie within the cointegrating vector.

(d) Raising interest rates should have a negative effect on output growth (IS curve), output growth should have a positive effect on the change in inflation (Phillips curve), the change in inflation should raise interest rates (Fisher effect), though this does not seem to be happening here.

### 2.2.8. VECM 1

A first order VECM was estimated on US annual data for log dividends, $ld_t$, and log earnings, $le_t$ $t = 1872 - 1999$ assuming unrestricted intercept and no trend. The system is

$$\begin{aligned} \Delta d_t &= a_{10} + \alpha_1(d_{t-1} - \beta e_{t-1}) + u_{1t} \\ \Delta e_t &= a_{20} + \alpha 2(d_{t-1} - \beta e_{t-1}) + u_{2t} \end{aligned}$$

with $E(u_{it}u_{jt}) = \sigma_{ij}$, $i, j = 1, 2$. The test statistics, $TS$, and 5% critical values, $CV$,for the Johansen trace cointegration tests were

| $H_o$ | $H_1$ | $TS$ | $CV$ |
|---|---|---|---|
| $r = 0$ | $r \geq 1$ | 79.14 | 15.49 |
| $r \leq 1$ | $r = 2$ | 0.32 | 3.84 |

The estimates (with standard errors in parentheses) are $\beta = 0.914$, (0.014), $\alpha_1 = -0.333$ (0.045), $\alpha_2 = 0.112$ (0.102).

An ECM equation was also estimated

$$\begin{array}{lllllllll} \Delta ld_t = & -0.142 & +0.276 & \Delta le_t & -0.363 & ld_{t-1} & +0.332 & le_{t-1} & +\varepsilon_t \\ & (0.018) & (0.031) & & (0.037) & & (0.033) & & \end{array}.$$

(a) How many cointegrating vectors do the trace tests indicate?

(b) What just identifying restriction is imposed on the cointegrating vector?

(c) Test the hypothesis that the long-run elasticity of dividends to earnings in the VECM is unity.

(d) Explain the concept of weak exogeneity. Does the VECM suggest that earnings are weakly exogenous for $\beta$?

(e) Assuming weak exogeneity of earnings, how do the parameters of the ECM relate to those of the VECM.

(f) Compare the VECM and ECM estimates of the long run elasticity of earnings to dividends.

**Answer**

(a) One (b) The coefficient of $d_t$ equals unity (c) (0.914-1)/0.014=-6.1 Reject $H_0 : \beta = 1$. (d) A variable is weakly exogenous for a parameter if the parameters of interest are just functions of the parameters of the conditional distribution and the parameters of the conditional and marginal distribution are variation free (no cross-equation restrictions). In this case the estimate of $\alpha_2$ is not significantly different from zero, t=0.112/0.102=1.1, so there is no information in the earnings equation about $\beta$. (e) Conditional on weak exogeneity the system is

$$\begin{aligned} \Delta d_t &= a_{10} + \alpha_1(d_{t-1} - \beta e_{t-1}) + u_{1t} \\ \Delta e_t &= a_{20} + u_{2t} \end{aligned}$$

the ECM is

$$\Delta d_t = (a_{10} - \phi a_{20}) + \phi \Delta e_t + \alpha_1(d_{t-1} - \beta e_{t-1}) + v_t$$

where $\phi = \sigma_{12}/\sigma_{22}$. (f) The long-run estimate from the ECM is 0.332/0.363=0.914, the same as from the VECM.

### 2.2.9. VECM 2

Suppose the observed vector-valued time series $\{\mathbf{y}_t, t = 1, \ldots, T\}$, where $\mathbf{y}_t$ is an $n \times 1$ vector of random variables for all $t$, is generated as

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \mathbf{u}_t, \quad t = 2, \ldots, T,$$

where $\mathbf{u}_t$ is, in turn, generated by

$$\mathbf{u}_t = \mathbf{B}\mathbf{u}_{t-1} + \epsilon_t, \quad t = 1, \ldots, T,$$

and $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{B}$ are $n \times n$ matrices of coefficients and $\epsilon_t$ are unobserved i.i.d. random disturbances with $E[\epsilon_t] = \mathbf{0}$ and $\text{var}(\epsilon_t) = \Omega$ for all $t$, where $\Omega$ is an $n \times n$ positive definite symmetric matrix.

(a) Show that this model can be rewritten as

$$\Delta \mathbf{y}_t = \mathbf{C}\mathbf{y}_{t-1} + \mathbf{D}\Delta \mathbf{y}_{t-1} + \mathbf{E}\Delta \mathbf{y}_{t-2} + \epsilon_t, \quad t = 3, \ldots, T,$$

where $\mathbf{C} = -(\mathbf{B} - \mathbf{I}_n)(\mathbf{A}_1 + \mathbf{A}_2 - \mathbf{I}_n)$, $\mathbf{D} = (\mathbf{B} - \mathbf{I}_n)\mathbf{A}_2 + \mathbf{B}\mathbf{A}_1$ and $\mathbf{E} = \mathbf{B}\mathbf{A}_2$.

**Answer** This follows from:

$$\begin{aligned}
\mathbf{y}_t - \mathbf{B}\mathbf{y}_{t-1} - \mathbf{y}_{t-1} &= \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} \\
&\quad -\mathbf{B}(\mathbf{A}_1\mathbf{y}_{t-2} + \mathbf{A}_2\mathbf{y}_{t-3} + \mathbf{u}_{t-1}) + \mathbf{u}_t - \mathbf{y}_{t-1} \\
&= \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} - \mathbf{B}\mathbf{A}_1\mathbf{y}_{t-2} - \mathbf{B}\mathbf{A}_2\mathbf{y}_{t-3} \\
&\quad +\epsilon_t - \mathbf{y}_{t-1} \\
&\quad \pm\mathbf{A}_2\mathbf{y}_{t-1} \pm \mathbf{B}\mathbf{A}_1\mathbf{y}_{t-1} \pm \mathbf{B}\mathbf{A}_2\mathbf{y}_{t-1} \\
\Rightarrow \Delta \mathbf{y}_t &= -(\mathbf{B} - \mathbf{I}_n)(\mathbf{A}_1 + \mathbf{A}_2 - \mathbf{I}_n)\mathbf{y}_{t-1} \\
&\quad +((\mathbf{B} - \mathbf{I}_n)\mathbf{A}_2 + \mathbf{B}\mathbf{A}_1)\Delta \mathbf{y}_{t-1} + \mathbf{B}\mathbf{A}_2\Delta \mathbf{y}_{t-2} \\
&\quad +\epsilon_t,
\end{aligned}$$

i.e. $\mathbf{C} = -(\mathbf{B} - \mathbf{I}_n)(\mathbf{A}_1 + \mathbf{A}_2 - \mathbf{I}_n)$, $\mathbf{D} = (\mathbf{B} - \mathbf{I}_n)\mathbf{A}_2 + \mathbf{B}\mathbf{A}_1$ and $\mathbf{E} = \mathbf{B}\mathbf{A}_2$.

(b) Show that $\mathbf{y}_t$ is

(i) $I(2)$ if $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{I}_n$ and $\mathbf{B} = \mathbf{I}_n$;

(ii) $I(1)$ if $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{I}_n$ and $|\mathbf{I}_n - \mathbf{B}z| = 0$ has all roots outside the unit circle;and

$I(1)$ if $|\mathbf{I}_n - \mathbf{A}_1 z - \mathbf{A}_2 z^2| = 0$ has all roots outside the unit circle and $\mathbf{B} = \mathbf{I}_n$.

**Answer**(i) It follows from (a) that, if $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{I}_n$ and $\mathbf{B} = \mathbf{I}_n$, then

$$\begin{aligned}
\Delta \mathbf{y}_t &= \mathbf{A}_1\Delta \mathbf{y}_{t-1} + \mathbf{A}_2\Delta \mathbf{y}_{t-2} + \epsilon_t \\
\Leftrightarrow \quad \epsilon_t &= (\mathbf{I}_n - \mathbf{A}_1 L - \mathbf{A}_2 L^2)\Delta \mathbf{y}_t,
\end{aligned}$$

and $|\mathbf{I}_n - \mathbf{A}_1 z - \mathbf{A}_2 z^2| = 0$ has a unit root (i.e. $z = 1$ is a solution). Hence, $\Delta \mathbf{y}_t$ is $I(1)$, or equivalently $\mathbf{y}_t$ is $I(2)$. (ii) If $|\mathbf{I}_n - \mathbf{B}z| = 0$ has roots outside the unit circle, then $\mathbf{u}_t$ is stationary; then, $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{I}_n$ implies that $\mathbf{y}_t$ is $I(1)$. (iii) If $\mathbf{B} = \mathbf{I}_n$, then it follows from (a) that

$$\Delta \mathbf{y}_t = \mathbf{A}_1\Delta \mathbf{y}_{t-1} + \mathbf{A}_2\Delta \mathbf{y}_{t-1} + \epsilon_t,$$

so that $\Delta \mathbf{y}_t$ is stationary, provided that $|\mathbf{I}_n - \mathbf{A}_1 z - \mathbf{A}_2 z^2| = 0$ has all roots outside the unit circle. Hence, under these conditions $\mathbf{y}_t$ is $I(1)$.

(c) Is there a way to test whether the process satisfies case (ii) as opposed to (iii) of part (b)? If so, how can this be determined? If not, why not?

**Answer**Case (iii) yields a stationary AR(2) for $\Delta \mathbf{y}_t$. Case (ii) implies

$$\begin{aligned} \Delta \mathbf{y}_t &= (\mathbf{B} - \mathbf{A}_2)\Delta \mathbf{y}_{t-1} + \mathbf{B}\mathbf{A}_2 \Delta \mathbf{y}_{t-2} + \epsilon_t \\ &= \tilde{\mathbf{A}}_1 \Delta \mathbf{y}_{t-1} + \tilde{\mathbf{A}}_2 \Delta \mathbf{y}_{t-2} + \epsilon_t, \end{aligned}$$

where $\tilde{\mathbf{A}}_1 = \mathbf{B} - \mathbf{A}_2$ and $\tilde{\mathbf{A}}_2 = \mathbf{B}\mathbf{A}_2$. This is also an AR(2), and stationarity of $\Delta \mathbf{y}_t$ implies that the roots of its characteristic polynomial $|I_n - \tilde{\mathbf{A}}_1 z - \tilde{\mathbf{A}}_2 z^2| = 0$ lie outside the unit circle. Hence, the two processes are indistinguishable.

### 2.2.10. Method of Moments 1

Consider the linear regression model

$$y = X\beta + u.$$

$y$ is a $T \times 1$ vector of observed dependent variables. $X$ is a $T \times k$ full rank matrix of right hand side variables. $\beta$ is a $k \times 1$ vector of unknown parameters. $u$ is a $T \times 1$ vector of unobserved disturbances, with $E(u) = 0$, $E(u\ u') = \sigma^2 I_T$.

(a) Suppose $E(X'u) = 0$ derive the method of moments estimator of $\beta$ and its variance covariance matrix.

(b) Suppose $E(X'u) \neq 0$, but there are is a $T \times k$ matrix of instruments, $W$ such that $E(W'u) = 0$ and $E(X'W)$ is a full rank matrix derive the method of moments estimator of $\beta$ and its variance covariance matrix.

(c) Briefly compare the method of moments and maximum likelihood approach to estimation.

**ANSWER**

Method of moments replaces population moments by sample equivalents. (a) define $\widehat{u} = y - X\widehat{\beta}$ then

$$\begin{aligned} X'\widehat{u} &= X'(y - X\widehat{\beta}) = X'y - X'X\widehat{\beta} = 0 \\ \widehat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + u) \\ &= \beta + (X'X)^{-1}X'u \end{aligned}$$

since $E(\widehat{\beta}) = \beta$

$$\begin{aligned} V(\widehat{\beta}) &= E(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)' = E((X'X)^{-1}X'u)((X'X)^{-1}X'u)' \\ &= E((X'X)^{-1}X'u\ u(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned}$$

(b) Define $\widetilde{u} = y - X\widetilde{\beta}$ then

$$\begin{aligned} W'\widetilde{u} &= W'(y - X\widetilde{\beta}) = W'y - W'X\widetilde{\beta} = 0 \\ \widetilde{\beta} &= (W'X)^{-1}W'y \\ &= (W'X)^{-1}W'(X\beta + u) \\ &= \beta + (W'X)^{-1}W'u \\ V(\widetilde{\beta}) &= E((W'X)^{-1}W'u\ u'W(W'X)^{-1}) \\ &= \sigma^2(W'X)^{-1}W'W(W'X)^{-1} \end{aligned}$$

(c) ML requires specifying the distribution of the errors, if this is correctly specified then ML is fully efficient but may not be robust when the distribution is incorrectly specified. MM does not require a distribution to be specified and may be more robust when then distribution is not known.

### 2.2.11. SEM 1

Consider the simultaneous equation model with structural form

$$\begin{bmatrix} 1 & 0 \\ a & 1 \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} x_t = \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}, \ t = 1, \cdots, T,$$

where $x_t$ is exogenous, and $a, \gamma_1$ and $\gamma_2$ are parameters; and assume $(\epsilon_{1t}, \epsilon_{2t})'$ is independent across $t$, with

$$E \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} = \mathbf{0}, \ \text{var} \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} = \Sigma \ \forall\, t,$$

where $\Sigma$ is a positive definite and symmetric matrix.

(a) Derive the reduced form of this model, i.e. express the parameters and residuals of

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} x_t + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}, \ E \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = \mu, \ , \text{var} \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} = \Omega$$

in terms of the parameters and residuals of the structural form.

**Answer** The first equation is already a reduced form (RF) equation, so that $b_1 = -\gamma_1$ and $u_{1t} \equiv \epsilon_{1t}$. Substituting the first equation into the second yields

the reduced form coefficient for the second equation, $b_2 = -(\gamma_2 - a\gamma_1)$, and $u_{2t} = \epsilon_{2t} - a\epsilon_{1t}$. It follows that $E[\mathbf{u}_t] = \mathbf{0}$, and

$$\Omega = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} - a\Sigma_{11} \\ \Sigma_{12} - a\Sigma_{11} & \Sigma_{22} + a^2\Sigma_{11} - 2a\Sigma_{12} \end{bmatrix}.$$

(b) Is the structural form of this simultaneous equation model identifiable?

**Answer** The structural form is not identifiable because identification requires at least 4 restrictions on the parameters of the system. The structural form imposes only three: two normalization restrictions (the coefficient of $y_{1t}$ in the first equation, and of $y_{2t}$ in the second, is normalized to 1) and one exclusion restriction ($y_{2t}$ is excluded from the first equation). Hence, identification requires one further restriction.

(c) Suppose that $\Sigma_{12} = \text{cov}(\epsilon_{1t}, \epsilon_{2t}) = 0$, and you run the OLS regression of $y_{2t}$ onto $y_{1t}$ and $x_t$. Show that the resulting coefficient estimates are unbiased for $-a$ and $-\gamma_2$.

**Answer** The right-hand-side variable $y_{1t}$ is uncorrelated with $\epsilon_{2t}$ as a consequence of $\Sigma_{12} = 0$ and $x_t$ is exogenous by hypothesis, i.e. also uncorrelated with $\epsilon_{2t}$. Hence, the regression equation satisfies Gauss-Markov assumptions, and unbiasedness then follows from the Gauss-Markov Theorem.

(d) Continuing with the assumption that $\Sigma_{12} = \text{cov}(\epsilon_{1t}, \epsilon_{2t}) = 0$, show how to use indirect least squares to obtain estimates of the structural form parameters.

**Answer** OLS estimation of the first (RF) equation yields estimates of $\gamma_1$ and $\Sigma_{11}$. OLS estimates of the second RF equation yields estimates of the RF parameters in part (a). The covariance estimate, obtained from the OLS residuals of both equations, yields an estimate of $a$. Under the assumption that $\Sigma_{12} = 0$, the variance estimate of the second equation then yields an estimate of $\Sigma_{22}$, while the estimate of the coefficient on $x_t$ yields an estimate of $\gamma_2$.

### 2.2.12. SEM 2

Consider the simultaneous equations model:

$$\begin{aligned} y_{1t} &= \beta_{12}y_{2t} + \gamma_{11}x_{1t} + u_{1t}, \\ y_{2t} &= \beta_{21}y_{1t} + \gamma_{22}x_{2t} + \gamma_{23}x_{3t} + u_{2t}, \end{aligned}$$

where $y_{1t}$ and $y_{2t}$ are endogenous variables, $x_{1t}$, $x_{2t}$ and $x_{3t}$ are exogenous variables, and $(u_{1t}, u_{2t})$ are NID$(0, \Sigma)$ random disturbances.

(a) Discuss the identifiability of each equation of the system in terms of the order and rank conditions for identification.

(b) Explain why the Ordinary Least Squares estimator of $(\beta_{12}, \gamma_{11})$ is inconsistent.

(c) What are the Two-Stage Least Squares estimators of the coefficients in the two equations? Describe the procedure step by step.

(d) How would you conduct a Wu–Hausman test for the exogeneity of $y_{2t}$ in the first equation?

**Answer**

(a) The system is

$$\begin{array}{ccccc} 1 & -\beta_{12} & -\gamma_{11} & 0 & 0 \\ -\beta_{21} & 1 & 0 & -\gamma_{22} & -\gamma_{23} \end{array}$$

Order condition is that the number of restrictions on each equation, $d \geq m$ where $m$ is the number of endogenous variables, here 2. (There are other acceptable ways of expressing the order condition). There are 3 restrictions on the first equation ($\beta_{11} = 1, \gamma_{12} = 0, \gamma_{22} = 0$) so it is overidentified. There are 2 restrictions on the second equation ($\beta_{22} = 1, \gamma_{21} = 0$) so it is just identified. The matrix corresponding to the exclusions in the first equation $-\gamma_{22} \quad -\gamma_{23}$ has rank one, as does the matrix corresponding to the exclusions in the second equation $-\gamma_{11}$, so the rank condition holds.

(b) $y_{2t}$ is correlated with $u_{1t}$ because, $u_{1t}$ determines $y_{1t}$ and $y_{1t}$ determines $y_{2t}$.

(c) First estimate the reduced form by OLS

$$\begin{aligned} y_{1t} &= \pi_{11}x_{1t} + \pi_{12}x_{2t} + \pi_{13}x_{3t} + e_{1t} \\ y_{2t} &= \pi_{21}x_{1t} + \pi_{22}x_{2t} + \pi_{23}x_{3t} + e_{2t} \end{aligned}$$

and use this to obtain the predicted values $\widehat{y}_{1t}$ and $\widehat{y}_{2t}$ then estimate by OLS using these fitted values.

$$\begin{aligned} y_{1t} &= \beta_{12}\widehat{y}_{2t} + \gamma_{11}x_{1t} + v_{1t} \\ y_{2t} &= \beta_{21}\widehat{y}_{1t} + \gamma_{22}x_{2t} + \gamma_{23}x_{3t} + v_{2t} \end{aligned}$$

(d) From the reduced form obtain the residuals from the $y_{2t}$ equation $\widehat{e}_{2t}$ then estimate

$$y_{1t} = \beta_{12}y_{2t} + \gamma_{11}x_{1t} + \delta\widehat{e}_{2t} + v_{1t}$$

test $\delta = 0$. If you cannot reject $\delta = 0$ then you can conclude that $y_{2t}$ is exogenous.

### 2.2.13. GARCH 1

Suppose that the rate of return on a stock $(x_t)$ is generated according to the model:

$$\begin{aligned} x_t &= \beta + \varepsilon_t, \\ \varepsilon_t &= u_t(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)^{1/2}, \quad u_t \sim \text{NID}(0,1) \end{aligned}$$

where $\alpha_0 > 0$ and $0 \leqslant \alpha_1 < 1$.

(a) Calculate the mean and variance of $x_t$, and show that $\{x_t\}$ is a serially uncorrelated process.

(b) What properties does the ordinary least squares estimator of $\beta$ have? Explain how to obtain asymptotically efficient estimates of $(\beta, \alpha_0, \alpha_1)$.

(c) Explain why the sample autocorrelation function of $\{x_t^2\}$ may be useful in the evaluation of the validity of the first-order ARCH assumption about $\varepsilon_t$.

(d) Explain how to test the hypothesis $\alpha_1 = 0$ using the Lagrange multiplier principle.

**Answer**

(a) $E(\varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, ...) = 0$ so $E(x_t) = \beta$.

$$V(x_t) = E(x_t - \beta)^2 = E(\varepsilon_t^2) = E({u_t}^2(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) = \alpha_0/(1 - \alpha_1)$$

$$E(\varepsilon_t \varepsilon_{t-k} \mid \varepsilon_{t-1}, \varepsilon_{t-2}, ...) = \varepsilon_{t-k} E(\varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, ...) = 0$$

Hence $x_t$ is a serially uncorrelated process with constant variance and mean $\beta$.

(b) Above we showed that all the properties required for the Gauss-Markov theorem hold therefore it is minimum variance in the class of linear unbiased estimators. Non-linear estimators may be more efficient and the MLE will be asymptotically efficient, which chooses the parameters to maximise

$$\begin{aligned} \log L &= -\frac{T}{2} \log 2\pi - \frac{1}{2} \sum_1^T \log h_t - \frac{1}{2} \sum_1^T \log \frac{\varepsilon_t^2}{h_t} \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned}$$

(c) Let $v_t = \varepsilon_t^2 - h_t$, then

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + v_t$$

Using law of iterated expectation, show that $v_t$ is a zero mean, constant variance, white noise, so $x_t^2$ follows an AR1. If there are significant higher order autocorrelations, this would suggest that this model is wrong.

(d) LM test is $TR^2$ based on OLS regression

$$\widehat{\varepsilon}_t^2 = \alpha_0 + \alpha_0 \widehat{\varepsilon}_{t-1}^2 + v_t$$

is Chi-squared one.

### 2.2.14. Method of Moments 2

An intertemporal utility maximisation problem yields the following first-order condition:

$$E\left[\left\{\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}(1+R_{t+1})\right\} z_t\right] = 0,$$

where $C_t$ denotes consumption in period $t$, $R_{t+1}$ is the return on financial wealth, $\delta$ is the discount rate, $\gamma$ is the coefficient of relative risk aversion, and $z_t$ is an $m \times 1$ vector of valid instruments.

(a) Assuming you have a time series of $n$ observations on consumption, the rate of return, and the instruments, explain how the moment conditions above can be exploited in order to estimate $\gamma$ and $\delta$ consistently.

(b) What is the minimum number of moment conditions that is required in (a)? Are there any gains to be made by having more moment conditions than the minimum?

(c) Outline a method for obtaining asymptotically efficient estimates of $\gamma$ and $\delta$.

(d) Explain how a test for the validity of the moment conditions can be carried out.

**Answer**

(a) $\theta = (\gamma, \delta)'$ can be estimated consistently by GMM choose the $\theta$ that minimises

$$g_n(\theta)' W_n g_n(\theta)$$

where $W_n$ is a symmetric positive definite weighting matrix and $g_n(\theta)$ is the $m \times 1$ vector of empirical moment conditions

$$g_n(\theta) = n^{-1} \sum_{t=1}^{n} \left[\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}(1+R_{t+1})\right] z_t = 0$$

(b) Need $m \geq 2$. There may be efficiency gains from using more than the minimum number of moment conditions.

(c) When $m = 2$, $\widehat{\theta}$ is the unique solution of $g_n(\widehat{\theta}) = 0$. When $m > 2$, obtain an initial estimate $\widehat{\theta}_1$ using an arbitrary $W_n$. Using $\widehat{\theta}_1$ estimate consistently

$$S = var\left[\sqrt{n}g_n(\theta)\right]$$

Obtain second round GMM estimate $\widehat{\theta}_2$ by setting $W_n = S^{-1}$. $\widehat{\theta}_2$ is aymptotically efficient, but the procedure may be interated further.

(d)

$$J = ng_n(\widehat{\theta})'S^{-1}g_n(\widehat{\theta}) \sim a\ \chi^2(m-2)$$

Under $H_0$ : valid moment conditions.

### 2.2.15. Non-parametric

Suppose the data $\{(y_i, x_i), i = 1, \ldots, n\}$, where both $y_i$ and $x_i$ are scalars, are generated as

$$y_i = f(x_i) + \epsilon_i,\ \ i = 1, \ldots, n,$$

where $f(\cdot)$ is an unknown, twice differentiable function, $E[\epsilon_i|x_i] = 0$ and $\text{var}(\epsilon_i|x_i) = \sigma^2$ almost surely, for all $i$, so that $E[y_i|x_i] = f(x_i)$. For any $x_i$, denote the neighborhood of $x_i$ consisting of the $k$ nearest neighbors on both sides of $x_i$ by $N_k(x_i)$. Note: For the purpose of the computations you are asked to carry out below, ignore the largest and smallest $x_i$ for which these neighborhoods may not contain $k$ points on one side. Consider the nonparametric $k$ nearest neighbor estimator of $f(x_i)$,

$$\hat{f}(x_i) = \frac{1}{2k+1}\sum_{j=1}^{n} y_j 1_{\{x_j \in N_k(x_i)\}},$$

where $1_{\{x_j \in N_k(x_i)\}} = 1$ if $x_j \in N_k(x_i)$ and 0 otherwise.

(a) Letting $\mathbf{x} = (x_1, \ldots, x_n)'$, show that

$$\begin{aligned} E\left[\hat{f}(x_i)\middle|\mathbf{x}\right] &= \frac{1}{2k+1}\sum_{j=1}^{n} f(x_j)1_{\{x_j \in N_k(x_i)\}} \\ \text{var}\left(\hat{f}(x_i)\middle|\mathbf{x}\right) &= \frac{\sigma^2}{2k+1}. \end{aligned}$$

**Answer** The first results follows from the conditional mean assumption $E[y_j|x_j] = f(x_j)$, and the second results follows from the i.i.d. and conditional homoskedastic-

ity assumptions and the fact that the sum contains only $2k+1$ non-zero elements:

$$\begin{aligned} E\left[\hat{f}(x_i)\middle|\mathbf{x}\right] &= \frac{1}{2k+1}\sum_{j=1}^{n} E[y_j|x_j]1_{\{x_j\in N_k(x_i)\}} \\ \text{var}\left(\hat{f}(x_i)\middle|\mathbf{x}\right) &= \frac{1}{(2k+1)^2}\sum_{x_j\in N_k(x_i)} \text{var}(y_j|x_j) = \frac{\sigma^2}{2k+1}. \end{aligned}$$

(b) Suppose $|x_i - x_j| = \Delta > 0$ for all neighbors $x_i$ and $x_j$, i.e. the distance between all $x$ values is the same. Using a Taylor's series expansion of $f(x_i)$ about $x_j$, i.e.

$$f(x_i) \approx f(x_j) + f'(x_i)(x_i - x_j) + \frac{1}{2}f''(x_j)(x_i - x_j)^2,$$

and the fact that $1 + \ldots + (k-1)^2 + k^2 \approx \frac{1}{3}(k^3 - 1)$, show that the bias of $\hat{f}(x_i)$, conditional on $\mathbf{x}$, is

$$E[\hat{f}(x_i)|\mathbf{x}] - f(x_i) \approx cf''(x_i)\Delta^2 k^2,$$

for some constant $c$ (which you do not have to specify). **Answer**

$$\begin{aligned} E\left[\hat{f}(x_i)|\mathbf{x}\right] &= \frac{1}{2k+1}\sum_{x_j\in N_k(x_i)} f(x_j)1_{\{x_j\in N_k(x_i)\}} \\ &\approx \frac{1}{2k+1}\sum_{x_j\in N_k(x_i)}\left[f(x_i) + f'(x_i)(x_i - x_j) + \frac{1}{2}f''(x_i)(x_i - x_j)^2\right] \\ &= f(x_i) + \frac{1}{2k+1}f'(x_i)\sum_{x_j\in N_k(x_i)}(x_i - x_j) \\ &\quad + \frac{1}{2k+1}\frac{1}{2}f''(x_i)\sum_{x_j\in N_k(x_i)}(x_i - x_j)^2 \\ &= f(x_i) + \frac{1}{2k+1}f'(x_i)\Delta\left[-k - (k-1) - \ldots + k\right] \\ &\quad + \frac{1}{2k+1}\frac{1}{2}f''(x_i)\Delta^2 2\left[1 + \ldots + (k-1)^2 + k^2\right] \\ &\approx f(x_i) + \frac{1}{2k+1}f''(x_i)\Delta^2\frac{1}{3}(k^3 - 1) \\ \Rightarrow E\left[\hat{f}(x_i)|\mathbf{x}\right] - f(x_i) &= cf''(x_i)k^2\Delta^2, \end{aligned}$$

for some constant $c$.

(c) Show that if $\Delta$ is proportional to $\frac{1}{n}$, then the value of $k$ that minimizes the conditional mean squared error of $\hat{f}(x_i)$ is proportional to $n^{\frac{4}{5}}$.

**Answer** The conditional MSE is equal to the sum of conditional variance and squared conditional bias, so that

$$\text{MSE}\left(\hat{f}(x_i)\middle|\mathbf{x}\right) = \frac{\sigma^2}{2k+1} + \left[cf''(x_i)\right]^2 k^4\Delta^4 \quad \to \min_k !$$

Except for irrelevant constants, this is the same as

$$\frac{c_1}{k} + c_2 k^4\Delta^4 = \frac{c_1}{k} + c_2 k^4 n^{-4} \;\to\; \min_k !,$$

for some constants $c_1$ and $c_2$. This yields $k^\star \propto n^{\frac{4}{5}}$.

(d) At which rate (expressed as a function of $n$) does the conditional mean squared error converge for the minimizing value of $k$?

**Answer** It follows from (c) that the conditional MSE is proportional to $n^{-\frac{4}{5}}$.

### 2.2.16. GARCH 2

Using US annual data $1873 - 2000$ on log stock prices, $ls_t$ an ARIMA (1,1,0) model was estimated

$$\Delta ls_t = \mu + \rho\Delta ls_{t-1} + \varepsilon_t$$

and a test for second order ARCH gave a p value of 0.0472. The equation was re-estimated assuming GARCH(1.1) errors, i.e. assuming that $\varepsilon_t \sim N(0, h_t)$ with

$$h_t = \varpi + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}.$$

The estimates (standard errors in parentheses) are $\mu = 0.043$ (0.016), $\rho = 0.071$ (0.109), $\varpi = 0.007$ (0.007), $\alpha = 0.0149$ (0.092), $\beta = 0.625$ (0.0291).

(a) Explain how a test for second order ARCH is conducted.

(b) Explain how the GARCH(1,1) model is estimated

(c) What are the conditions required for the variances to be non-negative? Are they satisfied in this case?

(d) Comment on the significance of the ARIMA and GARCH coefficients.

**Answer**

(a) estimate the ARIMA model and get residuals $\widehat{\varepsilon}_t$ the estimate

$$\widehat{\varepsilon}_t^2 = a + b_1\widehat{\varepsilon}_{t-1}^2 + b_2\widehat{\varepsilon}_{t-2}^2 + v_t$$

and test the joint hypothesis $b_1 = b_2 = 0$. (b) It can be estimated by maximum likelihood, since the log likelihood function is

$$MLL = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln h_t - \frac{1}{2}\sum_{t=1}^{T}\frac{\varepsilon_t^2}{h_t}$$

and this can be maximised with respect to the unknown parameters determining $\varepsilon_t$ and $h_t$.(c) Requires $\varpi$, $\alpha$ and $\beta$ all to be positive, this is satisfied in this case. (d) In the expected value equation, the drift is significant the AR1 coefficient is not in the variance equation the ARCH term is not significant but the GARCH term is very significant and quite large; so the variance is a very persistent process.

## 2.3. Applied Exercise.

### 2.3.1. Data

This exercise is designed to teach you to use a variety of different estimators in EViews and interpret the output. Although we give instructions for EViews, but you can use any program you wish to analyse the data.

The data is in an Excel file: Shiller.xls; on
http://www.econ.bbk.ac.uk/faculty/smith
under courses, data for exercises.

If you are in the workstation room, the file is Shiller.xls in G:ems data; eviews courses.

The file contains annual US data from 1871 to 2000 on

ND nominal dividends for the year
NE nominal earnings for the year
NSP nominal standard and poors stock price index, January value
PPI producer price index, January Value
R average interest rate for the year

The data is updated from Robert J Shiller 'Market Volatility', MIT Press 1989 and we will use it to re-examine the hypotheses in a famous paper J Lintner 'Distribution of Income of Corporations among Dividends, Retained Earnings and Taxes', American Economic Review May 1956.

### 2.3.2. Getting Started in EViews

Click on EViews icon.

Click on File, New, Workfile.

In the dialog box specify annual data, and put 1871 2000 in the boxes for beginning and end.

You will then get a box with two variables, C for the constant and Resid for the residuals.

Choose File, Import, Read Text Lotus Excel.

In the dialog box where it asks for Names or numbers type 5, and OK. Notice that it will start reading data at B2. This is correct, column A has years, which it already knows and row 1 has names, which it will read as names You will see the five variables in the workfile box.

You can save this file and any changes to it to disk, if you want to work at home.

### 2.3.3. Data Analysis

Highlight ND and NE in the box. Click Quick, Graph, Line Graph, OK the two variables and you will get a graph of dividends and earnings. Were there occasions when firms paid out more in dividends than they earned? After looking at it, close it and delete it. You could save it if you want.

In looking at data, it is often useful to form ratios. Graph NSP. Notice how the trend dominates the data. Three useful financial ratios are the Price-Earnings Ratio, $PE = NSP/NE$; the dividend yield, $DY = ND/NSP$ and the payout ratio, $PO = ND/NE$. All of these remove the common trend in the variables and are more stationary.

**Generate transformations of the data to create new series**

Type Quick, Generate Series and type into box

PE=NSP/NE

Press OK. Do the same for DY and PO and graph them. On which series can you see the effects of World War I, the 1929 crash, World War II, the Oil shock of 1973? Plot the three ratios. Are there shifts in the average levels of the three ratios? Look at the Price Earnings Ratio, Payout Ratio and Dividend Yield for the 1990s.

Get summary statistics on the ratios. Click on the series name, choose view, descriptive statistics, histogram and stats. This will give minimum and maximum values (check these are sensible), mean, median, skewness (which is zero for a normal distribution) and kurtosis (which is 3 for a normal distribution) and the JB test for the null hypothesis that the distribution is normal.

Always graph the data and transformations of it and look at the descriptive statistics before starting any empirical work. Make sure series are in comparable units before putting them on the same graph.

### 2.3.4. Regression

We are going to work with the logarithms of the data. Type Quick, Generate Series and type into box

LD=LOG(ND)

And OK. You will get a new series in the box LD. Similarly generate

LE=LOG(NE)

LSP=LOG(NSP)

**Run a static regression**

Click, Quick, Estimate an Equation, Type in

LD C LE

Always look at the dialogue window and note the options. Notice the default estimation method is LS- Least Squares (NLS and ARMA). NLS is non-linear least squares, arma, autoregressive moving average. We use these below. If you click the arrow on the right of LS, you will see that there are other methods you could choose: including Two stage Least Squares and GARCH, which we will use below. There is an option tag at the top, which you can use to get Heteroskedasticity and Autocorrelation Consistent (HAC) Standard Errors. You could also have entered Log(ND) C Log(NE).

Click OK and you will get the following output

Dependent Variable: LD
Method: Least Squares
Date: 07/29/05 Time: 13:26
Sample (adjusted): 1871 1999
Included observations: 129 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -0.434900 | 0.018699 | -23.25741 | 0.0000 |
| LE | 0.874196 | 0.011825 | 73.92920 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.977291 | Mean dependent var | 0.026989 |
| Adjusted R-squared | 0.977112 | S.D. dependent var | 1.323178 |
| S.E. of regression | 0.200179 | Akaike info criterion | -0.363824 |
| Sum squared resid | 5.089111 | Schwarz criterion | -0.319486 |
| Log likelihood | 25.46668 | F-statistic | 5465.527 |
| Durbin-Watson stat | 0.874621 | Prob(F-statistic) | 0.000000 |

Both the constant and the coefficient of LE are very significant, t ratios much bigger than 2 in absolute values and p values (Prob) of zero. The P value gives you the probability that the null hypothesis (in this case that the coefficient is zero) is true. It is conventional to reject the null hypothesis if the p value is less than 0.05. However the Durbin Watson Statistic (which should be close to 2) of 0.87 indicates severe serial correlation that suggests dynamic misspecification.

Type View; Actual Fitted Residual; Actual Fitted Residual Graph and you will get a graph of the residuals in blue and the actual in red and the fitted in

green. Notice that there are two toolbars, an inner one on the equation box and an outer one on the workfile box. To copy, highlight what you want to copy; use the edit button on the outer workfile box and click copy. You can then paste this into a Word File. The graph shows that the residuals are not random, there are quite long runs where the actual is above or below the fitted and there are some big spikes, larger positive residuals than one would expect, where the actual is much higher than the fitted. These were cases where earnings dropped sharply, but dividends did not respond, because dividends were smoothed relative to earnings.

### 2.3.5. Regression Output

Programs will give you:

Estimates of the coefficients, their standard errors, t ratios (ratio of coefficient to the standard error which is the test statistic for testing the null hypothesis that the coefficient is zero) and perhaps p values for the null hypothesis that the coefficients are zero.

Various descriptive statistics such as the mean of the dependent variable $\overline{y} = \sum_{t=1}^{T} y_t/T$, and its standard deviation

$$s_y = \sqrt{\sum_{t=1}^{T}(y_t - \overline{y})^2/(T-1)}$$

The Sum of Squared Residuals:

$$\sum_{t=1}^{T} \widehat{u}_t^2$$

the standard error of regression:

$$s = \sqrt{\sum_{t=1}^{T} \widehat{u}_t^2/(T-k)}$$

where $k$ is the number of regressors. The ordinary coefficient of determination and the version corrected for degrees of freedom (R bar squared):

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \widehat{u}_t^2}{\sum_{t=1}^{T}(y_t - \overline{y})^2};\ \overline{R}^2 = 1 - \frac{\sum_{t=1}^{T} \widehat{u}_t^2/(T-k)}{\sum_{t=1}^{T}(y_t - \overline{y})^2/(T-1)}$$

$R^2$ measures the proportion of variation in the dependent variable $y_t$ explained by the regression. An F test for the hypothesis that all the slope coefficients (i.e. other than the intercept) are equal to zero:

$$\frac{[\sum_{t=1}^{T}(y_t-\overline{y})^2-\sum_{t=1}^{T}\widehat{u}_t^2]/(k-1)}{\sum_{t=1}^{T}\widehat{u}_t^2/(T-k)} \sim F(k-1,T-k)$$

The Maximised Log-likelihood and some model selection criteria such as Schwarz or Akaike are often given (see theoretical notes). Note that EViews and Mfit use different formuale for these model selection criteria. In Eviews you choose the model with the smallest value, in Mfit you choose the model with the largest value.

The Durbin Watson statistic is

$$DW=\frac{\sum_{t=2}^{T}\Delta\widehat{u}_t^2}{\sum_{t=1}^{T}\widehat{u}_t^2}$$

This measures serial correlation in the residuals. It should be about 2 and is roughly equal to $2(1-\rho)$ where $\rho$ is the serial correlation coefficient. It is only appropriate for first order serial correlation when there are no lagged dependent variables in the equation. Use an LM test otherwise.

### 2.3.6. Dynamic Linear Regression

Given that the serial correlation in the original regression suggested dynamic misspecification, we add lagged values, denoted by (-1) in EViews.

Click, Quick, estimate equation and type in

LD C LE LE(-1) LD(-1) @TREND

@trend, is a variable that goes 1,2,3, etc. You will get estimates of the equation, these are rounded,

$$\begin{array}{rccccc} d_t = & -0.19 & +0.27e_t & +0.62d_{t-1} & +0.06e_{t-1} & +0.0007t \\ & (0.04) & (0.03) & (0.04) & (0.04) & (0.0006) \end{array}$$

with sample 1872-1999 (one observation was lost for lags) $R^2=0.996$ and $SER=0.083$. This fits much better, average error of 8.3%, rather than 20% above. All but lagged earnings and the trend are individually significant. The Durbin Watson is much better at 1.765.

Click View on the equation box; then Actual Fitted Residual; then Actual Fitted Residual Graph. The estimates of the residuals still show some outliers, big errors.

**Misspecification/Diagnostic tests** Click View on the equation box, choose Residual Tests, Serial Correlation LM tests, and accept the default number of lags to include 2. You will get the LM serial correlation test. Note that neither lagged residual is individually significant (t value less than 2, p value >0.05) nor are they jointly significant F stat p value is 0.19. So we do not have a serial correlation problem with this equation. On diagnostic tests, the null hypothesis is that they are well specified, p values below 0.05 indicate that there is a problem.

Click View, Residual tests, histogram- normality test. You will get the histogram and in bottom right the JB test of 56.59 and a p value of 0.0000. There is clearly a failure of normality, caused by the outliers.

Click View, residual, White Heteroskedasticity (no cross terms) p value is 0.24977, so no indication of heteroskedasticity, i.e. we do not reject the hypothesis of constant variance.

There are a range of other diagnostic tests available, e.g. for structural breaks, we will look at below. Diagnostic tests examine whether the assumptions made about the errors hold. The null is always that the model is well-specified: the assumption, e.g. normality, holds.

**Specification tests** Given that the model is well-specified, questionable in this case, we can test restrictions on the coefficients of the model to develop a new specification.

Click View, coefficient tests, redundant variables, enter

LE(-1) @trend

OK, you will get and F statistic and p value and a Likelihood Ratio test and p value, both the p values are over 0.2, so we cannot reject the joint hypothesis that the coefficients of both these variables are zero. Thus we can exclude them from the equation. It gives the equation with them excluded, note the estimates from this.

Click View, coefficient tests, Wald and type in to the box: C(3)=0,C(5)=0. This is exactly the same restriction as we tested above and we get the same answer from the Wald, that we can accept the two restrictions, with a Chi-squared p value of 0.2231.

Wald can also handle non-linear restrictions. Click View, coefficient tests, Wald again and type in: (C(2)+C(3))/(1-C(4))-1=0. This tests that the long-run coefficient on log earnings equals unity. Click OK. The hypothesis is clearly rejected with Chi-squared p value of 0.0017. Wald tests are not invariant to how you write non-linear restrictions. We could have written the same restriction:

C(2)+C(3)+C(4)-1=0. This gives a Chi-squared p value of 0.0068, so we still reject. But there are cases where writing the restriction one way leads to rejection and another way to acceptance.

### 2.3.7. Theoretical background.

Lintner suggested that there was a target or long-run dividend pay-out ratio, say, $\Theta$, such that $D_t^* = \Theta E_t$. We will take logs of this relationship, using lower case letters for logs, e.g. $d_t = \log(D_t)$, etc. Notice natural logs are almost universally used. Taking logs we get $d_t^* = \log(\Theta) + e_t$. This can be written in an unrestricted form as $d_t^* = \theta_0 + \theta_1 e_t$, where his theory suggests that $\theta_1 = 1$ and $\theta_0 = \log(\Theta)$. To this he added a 'Partial Adjustment Model' (PAM) and a random error

$$\begin{aligned} \Delta d_t &= \lambda(d_t^* - d_{t-1}) + u_t \\ d_t &= \lambda\theta_0 + \lambda\theta_1 e_t + (1-\lambda)d_{t-1} + u_t. \\ d_t &= b_0 + b_1 e_t + b_2 d_{t-1} + u_t \end{aligned}$$

The PAM can be justified if, for instance, firms smooth dividends, not adjusting them completely to short term variations in earnings. We estimate the $b_i$, as in the equation above, and then work out the theoretical parameters from them:

$$\lambda = 1 - b_2;\ \theta_1 = b_1/(1-b_2);\ \theta_0 = b_0/(1-b_2).$$

Over the period 1872-99 $b_1 = 0.30$, $b_2 = 0.67$ as we saw from the restricted model above, so $\lambda = 0.33$ and $\theta_1 = 0.30/0.33 = 0.91$. Notice that if we impose the further restriction $\theta_1 = 1$ the model can be estimated after creating a new variable earnings minus lagged dividends as:

$$\Delta d_t = a_0 + a_1(e_t - d_{t-1}) + v_t.$$

More general adjustment processes such as Error Correction Models (ECM) can be used and other variables, e.g. inflation and stock market prices included. The model above is in terms of nominal dividends and earnings, it could be done in terms of real dividends and earnings with $rd_t = d_t - p_t$, and written:

$$d_t - p_t = b_0 + b_1(e_t - p_t) + b_2(d_{t-1} - p_{t-1}) + u_t$$

removing the restrictions gives the unrestricted model:

$$d_t = c_0 + c_1 e_t + c_2 p_t + c_3 d_{t-1} + c_4 p_{t-1} + v_t.$$

The restricted model has three parameters, the unrestricted model has five parameters, so there are two homogeneity restrictions: $c_1 + c_2 = 1; c_3 + c_4 = 0$. Check this by working out the $c_i$ in terms of the $b_i$.

### 2.3.8. Misspecification tests on the Partial Adjustment Model

Re-estimate the partial adjustment model: enter D(LD) C LE LD(-1) in the equation box. You should get the estimates (standard errors), rounded, for 1872-1999

$$\begin{array}{rccc} \Delta d_t = & -0.129 & +0.302 e_t & -0.330 d_{t-1} \\ & (0.015) & (0.0234) & (0.027) \end{array}$$

with $R^2 = 0.567$, $SER = 0.083$, $DW = 1.847$, with long run coefficient $\theta_1 = -0.302/0.330 = 0.91$.

On the equation box, click View, then Residual tests, there will be a range of choices. Click Histogram-normality. You will get the histogram, which shows outliers, plus a range of statistics, with at the bottom the $J-B = 73.65$, $p = 0.000$. This indicates that the hypothesis that the residuals are normally distributed is rejected.

Click View, residual tests, serial correlation LM test, accept the default of two lags, you will get $p = 0.194$ on the F version, so we do not reject the null of no serial correlation. Notice the regression used to conduct the test is given below. RESID is the residuals.

Click View, residual tests, ARCH-LM, accept default one lag, $p = 0.258$ on F version. So we accept the null of no Autoregressive Conditional Heteroskedasticity. The regression to conduct the test is given below.

Click View, residual tests, White Heteroskedasticity (no cross-terms), we get $p = 0.0396$. So we reject the null of homoskedasticity (no heteroskedasticity). The regression to conduct the test is given below. The t on LE is significant.

Note that the ARCH and White tests have the same null, homoskedasticity, but different alternatives (different forms of heteroskedasticity). The form of the test is the same in regressing squared residuals on a set of regressors, the regressors differ between the two tests.

Click View, stability tests, Ramsey Reset tests, put the number of fitted terms at 2, $p = 0.951$. So we do not reject the null of linearity. Look at the regression below.

If you wanted to test for a change in the parameters at a particular date, you would use the Chow Stability tests, specifying the date at which you thought the relationship changed. The Breakpoint tests for equality of the regression coefficients before and after the break, assuming the variances in the two periods are constant. The Forecast tests whether the estimates for the first period forecast the second period.

Click View, stability tests, recursive estimates and choose CUSUM. This gives you a graph, which looks OK, the test statistics do not cross the confidence intervals, so no evidence against structural stability.

Click View, stability tests, recursive estimates and choose CUSUM of squares. This gives you a graph, which where the test statistics do cross the confidence intervals, so this suggests that there is evidence against structural stability.

Diagnostic tests for the same null hypothesis (e.g. homoskedasticity or structural stability) can give conflicting results because they are testing against different alternative hypotheses.

### 2.3.9. Estimate the Partial Adjustment Model by Non-linear Least Squares

Close the equation, you could name it and save it, and click, quick, estimate an equation again, type in

D(LD)=c(1)*(c(2)+c(3)*LE-LD(-1))

The D(...) first differences the data on LD. This estimates the partial adjustment model giving estimates of the long-run parameters and speed of adjustment directly: $c(1) = \lambda$, $c(2) = \theta_0$, $c(3) = \theta_1$, . You should get parameter estimates for the speed of adjustment and long-run parameters identical to those derived from the restricted linear model, $\lambda = 0.33$ and $\theta_1 = 0.91$.

The R squared is lower (0.56) because here we are explaining the change in log dividends (the growth rate of dividends) not the level of log dividends. The long-run elasticity of dividends to earnings is 0.91 and significantly different from unity, the speed of adjustment is 33% a year. This is the same equation as we had above, with exactly the same standard error of regression.

However you may get completely different estimates. This is because the likelihood function has multiple maxima and when the starting values are set at c(1)=0 c(2)=0 c(3)=0, it goes to this local maximum. It has $MLL = 88.24$, compared to the other maximum of 137.7, the parameters large negative numbers and a $R^2 = 0.06$. To get the global maximum we need to set other starting values. The local maximum is quite close to zero, so even starting the coefficients a little bit positive will solve the problem. To do this type

param c(1) 0.05 c(2) 0.0 c(3) 0.05

in the command window at the top under the toolbar. With these starting values it will get to the global maximum. When doing non-linear estimation, try to start with sensible starting values, using the economic interpretation or preliminary OLS regressions to give you sensible values. Also experiment with

different starting values.

### 2.3.10. Estimate the Partial Adjustment model allowing for non-normality and ARCH.

Above we estimated the model

$$d_t = b_0 + b_1 e_t + b_2 d_{t-1} + u_t$$

on the assumption that $u_t \sim IN(0, \sigma^2)$. But there was evidence that the errors were non-normal and this was mainly caused by excess kurtosis. Now we are going to assume that $u_t \sim It(0, h_t, \nu)$, the errors are independent with a student t distribution, expected value zero, a time varying variance $E(u_t^2) = h_t$, and degrees of freedom $\nu$. The degrees of freedom determine how thick the tails of the distribution are. If $\nu$ is small, the tails are much fatter than the normal distribution, if $v$ is around 30, it is very similar to the normal. The form of time varying variance we will use is $GARCH(1,1)$

$$h_t = c_0 + c_1 \widehat{u}_{t-1}^2 + c_2 h_{t-1} + \varepsilon_t$$

Close or save any equations. Click quick, estimate an equation, enter D(LD) C LE LD(-1) and then change method from LS to ARCH using the arrow on the right of the method box. You will now get a GARCH box. Change error distribution from Normal to Student's t. Accept the other defaults, click OK. The estimates of the equation are

$$\begin{array}{rccc} \Delta d_t = & -0.074 & +0.203 e_t & -0.220 d_{t-1} \\ & (0.011) & (0.016) & (0.018) \end{array}$$

the estimates of the variance equation are

$$\begin{array}{rccc} h_t = & 0.0002 & +0.304 \widehat{u}_{t-1}^2 & +0.740 h_{t-1} \\ & (0.0002) & (0.198) & (0.105) \end{array}$$

with $v = 3.24$.

The short-run coefficients are rather different from the OLS estimates, indicating much slower adjustment and a smaller short-run elasticity of dividends to earnings. But the long run estimate of the elasticity of dividends to earnings $\theta_1 = -0.203/0.220 = 0.92$ is very similar to our earlier estimate 0.91. The ARCH term $c_1$ is not significant $t = 1.5$, but the GARCH term $c_2$ is very signficant

$t = 7.06$. The ARCH(1)-LM test that we did above sometimes does not have very much power against GARCH. The estimate of $v$ is very small, close to the lower limit, indicating very fat tails relative to a normal distribution.

### 2.3.11. Estimate ARIMA models

Estimate a random walk model for log stock prices, up till 1990; then an ARIMA model and use it to forecast. Use quick estimate an equation, set the sample to 1873-1990 and type in

D(LSP) C.

You should get an estimate of the drift (C ) 0.036, MLL=37.42, s=0.177.

Estimate an ARIMA(1,1,1) model for log stock prices.

Click estimate on the equation equation bos, check the sample is 1873 1990 and type in

D(LSP) C AR(1) MA(1)

You will get estimates with MLL=41.19, s=0.173. Notice that both the AR (t=-3.06) and MA (t=5.50) terms are significant. Click forecast on the equation box. Set the forecast period to 1990 2000 look at the graph. It will save the forecast as LSPF. Close the equation and graph LSP and LSPF. This is clearly a terrible forecast, you will see that the actual and predicted steadily diverge over the 1990s.

Although the AR and MA terms are individually significant, they do not reduce the standard error of the regression very much relative to a random walk, and on a likelihood ratio test they are jointly significant, 2(41.19-37.42) LR=7.54 compared to a $\chi^2(2)$ at the 5% of 5.99 but not at the 1% of 9.21. This may be a common factor which cancels out.

### 2.3.12. Testing for Unit Roots

Click on LSP, then view, then unit root tests. You will get a dialogue box. Leave the test as Augmented Dickey Fuller (there are lots of other alternatives), choose level, choose intercept and trend, choose Akaike, leave maximum lags at 12. Choose OK. You will get the ADF test results. The ADF statistic is -0.026, much greater than the 5% critical value of -3.446 (given on the program output). Below is given the regression that was run to get the results. Notice that the lag length is 5 and that the test statistic is just the t ratio on LSP(-1) in the regression.

Repeat the process (choose view, unit root test) set if for first difference rather than level, choose just intercept. The lag length chosen is 3. The ADF is -6.09 which is much smaller than the 5% critical value of -2.88. Note that the critical values are different depending whether or not you have a trend.

We cannot reject a unit root for LSP but we can for the first difference of LSP, therefore LSP is clearly I(1). In practice, unit root tests are not always as clear-cut as this.

### 2.3.13. VAR, cointegration and VECM

Use Quick, estimate VAR and you will get a dialogue box. Enter LD LE as endogenous variables. In the list of exogenous variables add @trend to C. Accept the defaults for everything else. This will give you a second order unrestricted VAR with intercept and trend. Notice: both trends are significant; the second lags of both variables are insignificant; LD(-1) is insignificant in the LE equation. Click View, lag structure, lag length criteria and accept the default maximum lag of 8. You will get a table which shows that everything except the LR indicates that one lag is optimal. The optimal value has a star beside it. Choose View, lag structure, Granger Causality test. LE is clearly Granger causal for LD, but LD is not Granger Causal for LE p=0.330. This fits with the individual tests, LD(-1) and LD(-2) are both insignificant in the LE equation.

Choose estimate from the equation box and replace 1 2 by 1 1 in the lag intervals box  Look at the new estimates. Choose View, impulse responses, click the impulse definition tab at the top, choose generalised impulses. These graphs show the effect of a shock to each variable on itself and on the other variable. The generalised impulse response function assumes that the shocks have the estimated correlation in the sample, the Choleski uses an assumed causal structure for the shocks. LD shows a humped shaped response to a shock to LE, which remains significantly positive for 10 years. LE shows an immediate response to a shock to LD, through the contemporaneous covariance matrix, but it declines to zero.

Click, View, lag structure, AR roots, graph. It shows two inverse roots within the unit circle. Both roots are real (no complex component shown on the y axis). If a trend is not included in the VAR, it shows one on the unit circle and one within, which suggests that there may be one stochastic trend and one cointegrating vector. Click View, cointegration tests, and click the bottom button, option 6, summarise all 5 sets of assumptions and exclude @trend from the list of exogenous variables (the assumptions about the deterministic components allow

for a trend). Make sure that the lag order is 1 1. There are two tests, Trace and Maximal Eigenvalue, and 5 sets of assumptions about the determinist structure. Except for the case with a quadratic trend, all tests and assumptions about the deterministic elements indicate one cointegrating Vector. The information criteria are given below. From the stars, you can see that Akaike chooses one cointegrating vector (equation) and quadratic trend, Scwarz chooses one cointegrating vector, linear, intercept, no trend. Notice that EViews uses different conventions for lag length in the VAR and VECM case.

Suppose we maintain our PAM model for dividends,

$$\Delta d_t = \lambda(\theta_0 + \theta_1 e_t - d_{t-1}) + u_t.$$

There are two possible models for earnings, given that lagged dividends do not influence earnings: trend stationary, with long run trend growth $\eta_1$

$$\Delta e_t = a(\eta_1 t - e_{t-1} + e_o) + \varepsilon_{1t}$$

or difference stationary, with long run trend growth $\eta_2$ :

$$\Delta e_t = \eta_2 + \varepsilon_{2t}$$

Assuming $\lambda > 0$ as seems the case, with a trend stationary model and $a > 0$, both roots lie outside the unit circle. With a difference stationary model, there is one root on the unit circle, one stochastic trend, and one root outside the unit circle, one cointegrating vector. With the difference stationary model the VECM is

$$\begin{aligned} \Delta e_t &= \eta_1 + \varepsilon_{2t} \\ \Delta d_t &= \lambda(\theta_0 + \theta_1\eta_2) + \lambda(\theta_1 e_{t-1} - d_{t-1}) + [u_t + \lambda\theta_1\varepsilon_{2t}] \end{aligned}$$

Click Estimate on the equation box, choose Vector Error Correction rather than VAR, click the cointegration tab at the top and choose option 3, set the lag length to 0 0 and press OK. VECM 0 0, corresponds to VAR 11, in Eviews. You could also choose the number of cointegrating vectors, but leave it at the default of one. You will get the VECM estimates. The coefficient of earnings $\theta_1 = 0.91$ is very similar to what we got with the partial adjustment model, but this may be coincidence. You would clearly reject the hypothesis that the long-run elasticity was unity, $t = (0.914 - 1)/0.014 = -6.1$. View, Cointegration Graph, will give you a plot of the cointegrating relation, a measure of the deviation from equilibrium: $d_t - \theta_0 - \theta_1 e_t$.

You can impose restrictions on the cointegrating vectors and adjustment coefficients using the tab at the top marked VEC restrictions. Click impose restrictions and then type B(1,1)=1,B(1,2)=-1. This imposes a long-run unit coefficient of unity on earnings. Click OK. You will get the restricted estimates and a Likelihood ratio test that indicates that the restriction is rejected, as we determined above.

### 2.3.14. Two stage Least squares & Wu-Hausman test

Above we estimated a partial adjustment model by regressing log dividends on log earnings and lagged log dividends. The evidence of the VAR suggests that earnings may be treated as exogenous, since there was little feedback from lagged dividends to earnings. However, if $E(u_t\varepsilon_t) \neq 0$, this may cause $e_t$ to be correlated with $u_t$. We now investigate this.

First re-estimate the PAM by OLS, i.e. run LD C LE LD(-1) using LS over the period 1871-1999, it will use 1872-1999, since one observation is lost for the lag. The coefficient on LE is 0.301918 with a standard error of 0.023779.

Click estimate on the equation-box toolbar, change the method from LS to TSLS, you will get a new dialogue box with two windows. Leave the upper equation one the same and in the lower one for instrument list enter: C LE(-1) LD(-1) @trend. Click OK and you will get the TSLS estimates. The coefficient of LE is 0.344658 with a standard error of 0.035905. Thus the OLS and TSLS estimates do not look significantly different, the TSLS estimate $\pm 2$ standard errors covers the LS estimate.

We can test this formally with a Wu-Hausman test. Estimate by OLS: LE C LE(-1) LD(-1) @TREND. This gives us the same estimates as we got for the LE equation from the VAR 1 with intercept and trend. The VAR is the reduced form. This is the first stage of two stage least squares. You should always check this first stage, to see whether the instruments explain the endogenous variable, in this case E(-1) and @TREND are very significant. Close the equation, use Quick, Generate and define ULE=RESID. This saves the residuals from the first stage (reduced form equation for LE) as ULE. Then use OLS to estimate LD C LE LD(-1) ULE. The coefficient on ULE is -0.0777 with a t statistic of -1.63, so we do not reject the hypothesis that we can treat LE as exogenous.

We could also use two stage least squares to estimate a rational expectations model, in which dividends are determined by expected earnings in the next period, the expectations based on information in the current period. Click estimate,

choose TSLS, and type into the equation box: LD C LE(1) LD(-1) and into the instrument box: C LE LD(-1) @TREND. You will get a coefficient on future earnings of 0.33. Notice that the sample is 1872 1998, we have lost one observation at the end of the period, because of the future variable on the right hand side.

# 3. Doing your econometric project.

## 3.1. Introduction

Your applied project accounts for a third of the econometrics marks and past experience shows that doing the project helps people understand the theoretical material and pays off in better performance in the exam. Being able to show potential employers a well presented piece of empirical research also often helps get a job. You can also extend your project for your MSc dissertation, if you wish. These notes provide advice on doing the project, ignoring this advice may result in you being penalised when the project is marked.

During the first term you must learn how to produce and interpret regression output, by doing the applied exercise. At the beginning of the second term you should you should submit a 100 word outline. The outline should give a preliminary title, say where you are going to get the data and indicate the sort of model you will use in terms of dependent and independent variables. You have no formal supervisor but you should talk with one of the teachers at least once to discuss what you are doing.

Try to write up what you are doing as you go along and build in an 'audit trail'. It is very easy to make mistakes and to forget what you did, having records and documentation is essential. Make notes of exact references when you read articles, trying to search through the library for the reference at the last moment is a waste of time. Date your drafts so you can identify them and do not submit an old version by mistake.

**Make sure that you keep multiple back-ups of your data and drafts (e.g. on College server, your hard drive and a USB stick). You can lose computer files in lots of different ways and it is a lot of work redoing everything. We hear about a lot of computer disasters as the deadline approaches and we will not be sympathetic if you are having trouble completing because you did not back-up properly.**

### 3.1.1. Format

- **The completed project must be handed in at the end of the first week of the third term. Staple it together but do not put it in a binder or folder.**
- **The maximum length of the project is 5,000 words, we stop reading after that. Tables, graphs and technical appendices are not**

**included in this total.**

- **The data you used should be provided on disk, CD or memory stick, with your name on it, in an envelope, which also has your name on it, stapled to the back of the project. You can use any program, but the data must be provided in a way that allows us to replicate your results.**

- **The first page of the project should have;**

**title,**
**abstract,**
**your name,**
**the programme you are following (MSc Econ, PGCert Econometrics, etc),**
**the computer programs you used (e.g. Microfit, EViews, Stata, Excel) and**
**a word count.**

- **The pages should have page numbers.**

- **The project should be divided into properly titled sections, with an introduction and conclusion.**

- **The project should have a clear description of the data: sources and definitions, measurement units, graphs, etc.**

- **The project should look and read like and academic article with references to the literature presented in a standard academic form and listed in a bibliography.**

- **Do discuss the project with other students, people at work, etc. But the project must be your own work. Plagiarism is heavily penalised and more difficult to get away with than you may think. If in doubt, ask advice from the staff. We interview a selection of students to check that their project is their own work.**

### 3.1.2. Advice

Applied Econometrics involves a synthesis of data, economic theory and statistical methods to develop a model which can be used for some substantive purpose (e.g. to forecast, design policy, test theory, etc). Therefore you need to know how to: choose a topic; get the data; develop a theoretical model; use the statistical techniques on the computer; and write up the results in a way that convinces the reader that you have done something useful. You do not have to get a 'good' model, that is often a matter of luck. Do not be surprised if you cannot get a good model (right signs, passing misspecification tests) for a standard equation. The published results are often based on a lot of data mining (experimenting with different samples, estimation methods, variables including dummy variables) and the ones that do not work even after all this do not get published.

Do not paste program output into the text of the project, summarise it properly in equations or tables. Program output can be included in an appendix if necessary. Read empirical articles and try to copy their style and approach.

We do not assess you on the product, the quality of the final model, but on the process. We do this under five categories: data, economic (or other) theory, econometric analysis, presentation, written style. There is some substitution, e.g. if you put a lot of work into collecting or constructing the data we will give you extra credit for that, but there are diminishing marginal products, so make sure you put effort into all five elements. Remember that when we mark the project, we have the data you used. It is very easy for us to check what you did and what you missed doing and we do check.

## 3.2. Stages of doing your project

### 3.2.1. Choose a topic and get data

You can choose any topic, it does not have to be directly economic. Your choice of topic will be mainly constrained by availability of data. You cannot do applied econometrics without data, therefore your first objective is to find the data.

Part-time students often do work-related projects, where the data comes from their employer. We keep the data confidential, the project is not seen by anyone except the examiners. Note that it is very likely that the style you write the project in will be very different from the style you would use at work. If you do a work related project make sure you explain any abbreviations or technical terms that we might not be familiar with.

Foreign students often do projects replicating standard models estimated for the US or UK on data for their own country. Replicating published articles is a good source for projects, however you should note that it is often difficult to replicate published articles; see Dewald et al. (1986). Be aware that there are often mistakes in published articles. You could do the same model for a different time-period, country or industry and try and improve on the published model. Use the search engines to track down papers on subjects you are interested in.

There are a large number of possible data sources on the web, use Google or another search engine. Good sources are IMF International Financial Statistics; the World Bank data base: US data at the St Louis Fed or Bureauc of Economic Analysis or UK data at ONS. We also have a range of other databases on the Birkbeck eLibrary, including Datastream available in the Library.

Give the source and exact definition of any data you use and ensure you understand it. Just defining a variable as inflation is unacceptable; the reader needs to know at the minimum whether it is a percentage change or the change in the logarithm, at which frequency, whether it is the CPI, WPI, GDP deflator, etc.. Learn national accounts terminology, e.g. 'Investment' is usually Gross Domestic Fixed Capital Formation. Are the variables in current or constant prices? what currency? are they seasonally adjusted? what is the coverage (e.g. UK or Great Britain)? have the definitions (or coverage e.g. German reunification) changed over the sample period? Remember the ratio of a current price series to a constant price series is a price index.

In general the more observations the better, but 25 is a minimum. Be cautious about using large data sets, over 1,000 observations, unless you are familiar with handling them. Keep an eye on degrees of freedom as a rough rule of thumb the maximum number of parameters you should try and estimate is one third of the number of observations. With 25 observations this would be a maximum of 8 and preferrably fewer.

Do not agonise too long over choosing a topic and once you have started and got the data together do not be tempted to switch.

The topic should involve explaining at least one variable by some others. The relationship can be a standard economic one:

explaining a country's imports by the level of demand, GDP, domestic prices, import prices and the exchange rate;

explaining consumption by income, inflation and wealth remembering that the PIH says its a random walk;

explaining money demand by income and interest rates.

Many past students have looked at non-economic relationships:

explaining attendance at football matches by ticket prices, the quality of the teams, hooliganism and the weather;

explaining $CO_2$ concentrations by industrial production and sea temperature;

explaining London Marathon running times by age and sex.

If you do a standard economic topic like consumption and demand for money functions make sure that you are familiar with the recent literature, do not rely on the treatment in elementary text-books. If you do a non-standard topic there may be no literature and you have to provide the relevant context.

**Try to identify some precise questions that you will try to answer from this data.**

### 3.2.2. Getting to know your data.

Once you have got your data and are clear on the exact definitions you may have to adjust them in various ways: interpolate to deal with missing observations, splice data that are on different base years or convert them into different currencies. Published data are not infallible, look for possible mistakes in the data. Learn about the context, e.g. relevant history and institutions, to interpret the data.

Whenever you load new data, you should spend a lot of time graphing it, using line graphs, histograms and scatter diagrams as appropriate. Look for mistakes (missing observations or errors in the data, for instance). Establish the order of magnitudes of the data; what values would you expect them to take? Note whether variables like interest rate or growth rates are proportions or percent and whether they are at per annum rates. Make notes of the main features of the data. Things to look for in time-series are: trends, cycles, seasonal patterns, outliers (try and find what caused unusual observations, e.g. wars or devaluations). For trended time-series look for patterns in growth rates and ratios, which are often stable in economic time-series. For seasonally unadjusted quarterly data also look at the year on year growth rate to remove seasonality. Theory often tells you the ratios to look at: average propensity to consume, velocity of circulation, real exchange rate, real interest rate, etc. Look at the relative variance of different series, using histograms and standard deviations, again usually for growth rates and ratios; the variance of the level is usually dominated by trends in time-series. In cross-section, look at the distributions of the variables (histograms), look for outliers, use scatter diagrams to try and establish the shape of the relationships between variables to determine functional form and identify heteroskedasticity. The scatter diagrams

will help you develop the econometric models later.

Notice that although we discuss the theory below, you should be using the theory to analyse the data at this stage. The theory will tell you what transformations of the original data are likely to be appropriate, e.g. to construct real exchange rates or real interest rates from the original data. Very many economic models use logarithms of the original variables because: economic variables such as prices and quantities tend to be non-negative; the change in the logarithm is approximately the growth rate; variances are more likely to be constant (errors are proportional rather than additive); coefficients can be interpreted as elasticities; scale does not matter (multiplicative constants are picked up in the intercept) and many interesting economic restrictions (e.g. homogeneity of degree zero in prices) can be expressed as linear restrictions on logarithmic models.

As part of the data description for time-series you should check the order of integration of the variables, using unit root tests. Do not do unit root tests on cross-section data.

Discuss how your data relate to the theoretical concepts they are supposed to be measuring. Comment on the quality of the data, much economic data is very bad. In some case you can create your own data with dummy variables. Where there are a number of possible measures for a theoretical concept, try to use each of them and test which is the best.

### 3.2.3. Develop a Theory

Theory should be interpreted very widely here: what do we know about the process that might have generated the data? Standard economic theory may tell you what variables are likely to be relevant: in a demand function: income and own and other prices will appear on the right hand side. Thus imports would depend on GDP, domestic prices, foreign prices and the exchange rate. It may tell you restrictions that can be tested: the demand function should be homogenous of degree zero in prices. In the case of import demand, this means that the relevant regressor is the real exchange rate. Theory may tell you about functional form, but usually does not. If the variables are always positive start with logarithmic transformations. If the dependent variable is a proportion, say p, lying between zero and one, consider a logistic transformation: log{p/(1-p)}.

It is often quite difficult to translate pure economic theory into the form of an equation that can be estimated, but wider theory is often useful in giving you a starting point. Many economic variables (particularly asset prices like stock

market prices and foreign exchange rates) should be random walks, and this is a good starting point. Many theories imply that certain ratios should be constant in the long run, and this provides a starting point. In time series data it is useful to think of the theory in terms of three components: a long-run equilibrium, an adjustment process and a process for forming expectations. Always look for seasonality in data of frequency less than annual.

A central issue is distinguishing correlation from causality. Angrist and Pischke (2009) are very good on this. Theory should help with this distinction. Sometimes you do not want to make causal statements, just forecast. Use your common sense to develop the theory, ask is this a sensible way of explaining the data, and try to identify the interesting questions.

### 3.2.4. Example

It is often useful to use theory to set up a general model which nests the alternatives. Purchasing Power Parity (PPP) says the spot exchange rate, S, (measured in units of domestic currency per unit of foreign currency) should equal the ratio of domestic, P, to foreign, P*, prices:

$$S = \frac{P}{P^*}$$

(Note the way this is written depends on the way the exchange rate is defined, domestic/foreign or foreign/domestic). In most cases we do not have a cross section of actual prices (an exception is the Economist Big Mac data), but time-series on price indexes; P* is a foreign price index and P is a domestic price index .

$$S = R\frac{P}{P^*}$$

R, the real exchange rate, depends on the units of measurement of the price indices. Using lower case for logarithms, assuming time-series data, and adding a random error term; we can write this as:

$$s_t = r + p_t - p_t^* + v_t$$

This is the restricted equation. Notice that we can estimate $v_t$ from a regression of the log real exchange rate, $r_t = s_t - p_t + p_t^*$ on a constant and use the Sum of Squared Residuals from this regression as our RSSR. An unrestricted equation is

$$s_t = \alpha + \beta_1 p_t + \beta_2 p_t^* + u_t$$

and we can get our USSR from this and test the two hypotheses $\beta_1 = 1; \beta_2 = -1;$ with an F test.

Suppose we do not know what price index to use. We could either use the Consumer Price Index, $p_{1t}$ or the Wholesale Price Index, $p_{2t}$. We could decide by constructing a more general unrestricted model

$$s_t = \alpha + \beta_1 p^*_{1t} + \beta_2 p_{1t} + \gamma_1 p^*_{2t} + \gamma_2 p_{2t} + e_t$$

Then we could choose between the two measures by comparing the restrictions $\beta_1 = 0; \beta_2 = 0;$ which implies the WPI is the right measure, with $\gamma_1 = 0; \gamma_2 = 0$ which implies the CPI is the right measure. Hopefully, the F tests will accept one hypothesis and reject the other. Of course, we might reject both or accept both, which means that we have to think about the problem a bit more.

Try to use the theory to formulate specific questions you want to ask of the data and organise your write-up around them: does PPP hold? is the CPI or WPI the right measure to use? etc. In practice, it would be better to ask whether PPP held in the long-run. This implies that $r_t = s_t - p_t + p^*_t$ should be $I(0)$. It could also be tested using a dynamic ARDL model. The unrestricted model would be

$$\Delta s_t = \alpha_0 + \alpha s_{t-1} + \beta p_{t-1} + \gamma p^*_{t-1} + \sum_{i=1}^{p} a_i \Delta s_{t-i} + \sum_{i=0}^{p} b_i \Delta p_{t-i} + \sum_{i=0}^{p} c_i \Delta p^*_{t-i} + u_t$$

the restricted model would be:

$$\Delta s_t = \alpha_0 + \alpha r_{t-1} + \sum_{i=1}^{p} a_i \Delta s_{t-i} + \sum_{i=0}^{p} b_i \Delta p_{t-i} + \sum_{i=0}^{p} c_i \Delta p^*_{t-i} + u_t$$

This assumes that $p_t$ and $p^*_t$ are weakly exogenous for the long-run parameters, which could be tested in the context of a VAR.

### 3.2.5. Estimate some equations

Your examination of the data and review of the theory should have given you some ideas about designing the models that you will estimate: the variables that you include, the functional form that you use, questions that you need to answer with hypothesis tests, the sign and magnitude of the coefficients you expect. Kennedy (2003, chapter 5 and 21) is good on general specification issues and what to do when you get "wrong" signs.

**You must organise the estimation process**. It is very easy to make mistakes, being organised makes this less likely. It is very easy to lose files: make back-ups on separate disks and establish a system for naming files and variables. It is very easy to get buried in vast piles of regression output: organise your estimation to stop this happening. Getting lost is particularly easy when for each regression you also calculate diagnostic tests for normality, structural stability, etc. You do not have to do these diagnostic tests for every equation, but they are often informative. You need to provide appropriate diagnostics for the equations you report, but the equations you report will only be a small subset of those you run. Always look at the graph of fitted values and residuals. This should be an automatic check to see if there is any suggestion of problems like omitted variables. In time-series your first concern should be about dynamic specification and structural stability; in cross-sections functional form and heteroskedasticity. Remember one form of misspecification can show up in the diagnostic test for another form of misspecification. If there is evidence of misspecification, think about how you should respecify the model.

Look at the magnitude of the coefficients, short-run and long-run, are they sensible? What values would you expect from theory? An effect may be statistically significant, but so small as to be economically unimportant. Or large in economic terms, but imprecisely estimated so not statistically significant. Remember that our conventional significance levels, like 5% are just conventions, other levels may be appropriate.

The first stage in getting organised is to write up as you go along. Before you start estimating anything you should have written the first draft of the data analysis and theory sections, with predictions for likely values of the estimated coefficients. The second stage of getting organised is to be systematic in your estimation. Design a sequence of estimates in the light of your questions and record the results in summary tables. It is very easy to forget which model the picture of the residuals and fitted values corresponds to. Go to the computer with a plan of what you are going to do organised around the crucial questions and a draft table to summarise the results. If you have designed a table, you can just put a cross in the box (or the p value) if it fails the normality test for instance.

Use the simplest estimation method appropriate. If you use complex estimation methods, make sure that you know what the software is doing and can interpret the output. Report how sensitive your results are to estimation method, sample used, variables included, etc.

### 3.2.6. Write up the results.

The final project should read like an academic economic article, not a part of your autobiography. All the problems you had finding the data; making the computer work; trying to understand the articles; your personal crises; do not belong in the project. Read academic articles to see how they present their results, but bear in mind that the way research is presented in the literature is not a good guide to how it was actually carried out. Becker (1986) and McCloskey (1987) have good advice on writing. Do deals with fellow students to read drafts of each others projects to see if they are clear and to correct errors. Worry about spelling, grammar, construction of sentences and paragraphs and try to make the writing lively and interesting.

The project should have a Title, Your name, An abstract (about 100-200 words), a word count. The pages should be numbered. It should be divided into sections. Possible sections are:

1. **Introduction.** This should motivate the questions you are asking, provide some background and explain why the issue is interesting.

2. **Theory.** Provide a brief link to the literature, set up the model and any hypotheses you want to test. Set out the questions you are going to try and answer.

3. **Data.** Give exact definitions, units and sources; discuss any measurement problems, how the variables relate to the theoretical concepts; characterise the basic properties of the data, identify any trends, seasonals, cycles, outliers etc; provide any relevant history. You must give some graphs which summarise the main features of the data. If you miss something which would be very obvious had you graphed the data we will penalise heavily. For time-series discuss the order of integration of the data. You may want to include a separate data appendix with more detail.

4. **Statistical Model.** Briefly discuss the estimation methods you are going to use and why they are appropriate. This involves justifying the assumptions that you made about the distribution of the error terms. In most cases you will use Ordinary Least Squares, but you need to justify your choice of estimator. Do not put text-book material in your project, e.g. proofs that OLS is BLUE. Just give a reference to the text-book. The project should

contain all information that we do not know but need to know, not things we know.

5. **Results.** Make sure that you organise the presentation of your results clearly, bringing out the important ones and referring to the others in passing. For instance, if you tried a variable that proved insignificant, just say you tried it and it proved insignificant, you do not have to go into detail. Think carefully about how you want to present the numerical results, either as equations in the text or tables. What information do you need to convey? This will include the estimated coefficients, their standard errors (or their t ratios or p values, but only one of the three), the standard error of regression and some diagnostic tests. Make sure you explain your results, e.g. say whether you give standard errors or t ratios. Look at empirical economics articles and see how they convey the information. You can put program (EViews, Mfit, etc) output as an appendix, but the main text should convey the crucial results in a comprehensible form. Make sure that you interpret the results in terms of the substantive issues and consider both the size of coefficients and their significance.

6. **Conclusions.** What did we learn from this project? How were the questions posed earlier answered? What is their relevance for practical questions of forecasting, policy etc? Are the answers consistent with theory or institutional information? Is the model statistically well specified?

7. **References.** Make sure you follow the standard economic style of referencing, as in these notes.

8. **Appendices.** More detailed output or technical derivations which are not in standard sources can be put as appendices.

### 3.3. Commandments

These ten commandments of applied econometrics are given by Kennedy (2003).

1. Thou shalt use common sense and economic theory.

2. Thou shalt ask the right question.

3. Thou shalt know the context.

4. Thou shalt inspect the data.
5. Thou shalt not worship complexity.
6. Thou shalt look long and hard at thy results.
7. Thou shalt beware the costs of data mining.
8. Thou shalt be willing to compromise.
9. Thou shalt not confuse (statistical) significance with substance.
10. Thou shalt confess in the presence of sensitivity.

### 3.4. General advice

Read, write and think. Keep reading and relate your problem to what is in the literature. Mankiw Romer, Weil (1992) use the theory very effectively to organise their results. Fair (1996) is good on forecasting issues. Read more applied articles and see how the authors did it. Make your project look professional, something that might get published, many past projects have been published. Potential employers often ask to see projects, so keep a copy for yourself. Try and follow the examples of professional writing in the literature. Start writing early and keep rewriting. Organise your empirical investigation, so you do not get lost. Back up your computer files. Focus on particular questions. Explain why these are interesting questions. Try to make the project clear, brief and interesting. Make sure you have got the references right, make sure that when you quote from another paper you put it in quotation marks and give the exact source. Keep rewriting it to achieve those goals. Get other students to read it and comment on it. Remember you are not being marked on how good the final model is, you are being marked on how you went about it and how you reported what you did. You will be penalised if you have not taken account of advice in these notes. Keep a copy of your project, we will not return it. Try and enjoy the process, it can be fun discovering new things.

# References

Becker, H S (1996) **Writing for Social Scientists**, University of Chicago Press.

Dewald, W G, J G Thursby & R G Anderson (1986) Replication in Empirical Econometrics, **American Economic Review** Sept p587-603.

Fair, R C (1996), Econometrics and Presidential Elections, **Journal of Economic Perspectives** Summer, p89-102.l.

Kennedy P (2003) **A Guide to Econometrics** 5th edition, Blackwell.

Mankiw N G, D Romer and D N Weil (1992) A Contribution to the Empirics of Economic Growth, **Quarterly Journal of Economics**, p407-437.

McCloskey D N (1987) **The Writing of Economics**, Macmillan.

# 4. Notes, The Linear Regression Model, LRM

## 4.1. Notation

Econometrics is mainly about estimating linear regression models. The bivariate regression model is of the form:

$$y_t = \beta_1 + \beta_2 x_t + u_t$$

for $t = 1, 2, ..., T$. This is a set of $T$ equations which explain observations on a dependent variable, $y_t$, by an independent variable $x_t$ (which may be a non-linear function of some other variable) and errors, are $u_t = y_t - \beta_1 - \beta_2 x_t$. Least squares chooses $\beta_i$ to minimise $\sum u_t^2$. Multiple regression, with k explanatory variables takes the form

$$y_t = \beta_1 + \beta_2 x_{2t} + ... + \beta_k x_{kt} + u_t$$

where $x_{1t} = 1$ all $t$. This can be written in vector form as:

$$y_t = \beta' x_t + u_t$$

where $\beta$ and $x_t$ are $k \times 1$ vectors. Or in matrix form as

$$y = X\beta + u$$

where $y$ is a $T \times 1$ vector and $X$ is a $T \times k$ matrix. For the bivariate regression, this is

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ .. & .. \\ 1 & x_T \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ .. \\ u_T \end{bmatrix}$$

with

$$(X'X) = \begin{bmatrix} T & \sum x_t \\ \sum x_t & \sum x_t^2 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{T\sum x_t^2 - (\sum x_t)^2} \begin{bmatrix} \sum x_t^2 & -\sum x_t \\ -\sum x_t & T \end{bmatrix}$$

The sum of squared residuals $u'u = \sum_{t=1}^{T} u_t^2$ is

$$\begin{aligned} u'u &= (y - X\beta)'(y - X\beta) \\ &= y'y + \beta' X'X\beta - 2\beta' X'y \end{aligned}$$

If $A$ is a $n \times m$ matrix, and $B$ is an $m \times k$ matrix the transpose of the product $(AB)'$ is $B'A'$ the product of a $k \times m$ matrix with a $m \times n$ matrix, $A'B'$ is not conformable. $y'X\beta = \beta'X'y$ because both are scalars ($1 \times 1$ matrices). Scalars are always equal to their transpose. The term $\beta'X'X\beta$ is a quadratic form, i.e. of the form $x'Ax$. Quadratic forms play a big role in econometrics. Matrix, $A$, is positive definite if for any $a$, $a'Aa > 0$. Matrices with the structure $X'X$ are always positive definite, since they can be written as a sum of squares. Define $z = Xa$, then $z'z = a'X'Xa$ is the sum of the squared elements of $z$.

Writing $u'u = \sum_{t=1}^{T} u_t^2$ out explicitly we get the three terms

$$\sum y_t^2 + [\beta_1^2 T + \beta_2^2 \sum x_t^2 + 2\beta_1\beta_2 \sum x_t] - 2(\beta_1 \sum y_t + \beta_2 \sum x_t y_t) \qquad (4.1)$$

you can see that the middle term $\beta'X'X\beta$ in [..] is a quadratic.

### 4.1.1. Differentiation with vectors and matrices

To minimise $u'u$, which is a function of the $k$ elements $\beta$, we will need to take derivatives, getting $k$ derivatives with respect to each element of $\beta$. Consider the equation:

$$P = \underset{1\times n}{x'} \; \underset{n\times 1}{a}$$

Then the derivatives of $P$ with respect to $x$ and $x'$ are defined as :

$$\frac{dP}{dx} = a \text{ and } \frac{dP}{dx'} = a'$$

For $n = 2$:

$$\begin{aligned} P &= [x_1, x_2] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \\ &= x_1 a_1 + x_2 a_2 \end{aligned}$$

Then

$$\frac{dP}{dx_1} = a_1 \text{ and } \frac{dP}{dx_2} = a_2$$

So

$$\frac{dP}{dx} = \begin{bmatrix} \frac{dP}{dx_1} \\ \frac{dP}{dx_2} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = a$$

and

$$\frac{dP}{dx'} = \left[\frac{dP}{dx_1}, \frac{dP}{dx_2}\right] = [a_1, a_2] = a'$$

Consider the quadratic form:

$$Q = \underset{1\text{x}n}{x'} \underset{n\text{x}n}{A} \underset{n\text{x}1}{x}$$

Then the derivative of Q with respect to $x$ or $x'$ is defined as :

$$\frac{dQ}{dx} = 2Ax \text{ and } \frac{dQ}{dx'} = 2x'A$$

For $n = 2$, assuming $A$ is symmetric for simplicity,:

$$Q = [x_1, x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{aligned} Q &= [x_1, x_2] \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{12}x_1 + a_{22}x_2 \end{bmatrix} \\ &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \end{aligned}$$

So:

$$\frac{dQ}{dx_1} = 2a_{11}x_1 + 2a_{12}x_2 \text{ and } \frac{dQ}{dx_2} = 2a_{12}x_1 + 2a_{22}x_2$$

Then

$$\begin{aligned} \frac{dQ}{dx} &= \begin{bmatrix} \frac{dQ}{dx_1} \\ \frac{dQ}{dx_2} \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + 2a_{12}x_2 \\ 2a_{12}x_1 + 2a_{22}x_2 \end{bmatrix} = 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \underset{2\text{x}2}{2A} \underset{2\text{x}1}{x} \end{aligned}$$

and

$$\begin{aligned} \frac{dQ}{dx'} &= \left[\frac{dQ}{dx_1}, \frac{dQ}{dx_2}\right] = [2a_{11}x_1 + 2a_{12}x_2, 2a_{12}x_1 + 2a_{22}x_2] \\ &= 2\,[x_1, x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \\ &= \underset{1\text{x}2}{2x'} \underset{2\text{x}2}{A} \end{aligned}$$

### 4.1.2. Least Squares in the LRM

Consider the linear regression model

$$\underset{T\times 1}{y} = \underset{T\times k}{X}\underset{k\times 1}{\beta} + \underset{T\times 1}{u}$$

The problem is to minimize the sum of squared residuals with respect to $\beta$

$$\begin{aligned} u'u &= (y - X\beta)'(y - X\beta) \\ &= (y' - \beta'X')(y - X\beta) \\ &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta. \end{aligned}$$

Since these are all scalars, the second and third terms are equal and we can write

$$u'u = y'y - 2\beta'X'y + \beta'X'X\beta$$

The second term is:

$$P = \underset{1\times k}{2\,\beta\,'}\underset{k\times 1}{(X'y)}$$

From above we know that if $P = x'a$, $\frac{dP}{dx} = a$ so

$$\frac{dP}{d\beta} = 2X'y$$

The third term is a quadratic form

$$Q = \underset{1\times k}{\beta\,'}\underset{k\times k}{(X'X)}\underset{k\times 1}{\beta}$$

From above we know that if $Q = x'Ax$, $\frac{dQ}{dx} = 2Ax$ so:

$$\frac{dQ}{d\beta} = 2X'X\beta$$

And the first order condition (FOC) is

$$0 = -2X'y + 2X'X\widehat{\beta}$$

so

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

To find the least squares estimates, we cancel the $2s$ and solve for $\widehat{\beta}$, the least squares estimator that makes the FOC equal zero. This requires that $X'X$ is non-singular, so an inverse exists. The second derivative is $2X'X$ which is a positive definite matrix, so this is a minimum.

In the bivariate model to minimise $u'u$ we have to differentiate the sum of squred residuals, (4.1) above, twice, with respect to $\beta_1$ and $\beta_2$, to get the $2 \times 1$ vector of derivatives and set them equal to zero. The two elements of the vector are

$$\frac{\partial u'u}{\partial \beta_1} = 2\widehat{\beta}_1 T + 2\widehat{\beta}_2 \sum x_t - 2\sum y_t = 0 \tag{4.2}$$

$$\frac{\partial u'u}{\partial \beta_2} = 2\widehat{\beta}_2 \sum x_t^2 + 2\widehat{\beta}_1 \sum x_t - 2\sum x_t y_t = 0 \tag{4.3}$$

Check that this corresponds to the matrix formula. We can also write these as

$$\begin{aligned} -2\sum (y_t - [\widehat{\beta}_1 + \widehat{\beta}_2 x_t]) &= -2\sum \widehat{u}_t = 0 \\ -2\sum x_t (y_t - [\widehat{\beta}_1 + \widehat{\beta}_2 x_t]) &= -2\sum x_t \widehat{u}_t = 0 \end{aligned}$$

The least squares estimates make the residuals (estimates of the errors) uncorrelated with the regressors. Our least squares estimate of $\beta$ is denoted $\widehat{\beta}$ and is a $2 \times 1$ vector.

$$(X'X)^{-1} X'y = \frac{1}{T\sum x_t^2 - (\sum x_t)^2} \begin{bmatrix} \sum x_t^2 & -\sum x_t \\ -\sum x_t & T \end{bmatrix} \begin{bmatrix} \sum y_t \\ \sum x_t y_t \end{bmatrix}$$

$$\widehat{\beta}_1 = \frac{\sum x_t^2 \sum y_t - \sum x_t \sum x_t y_t}{T\sum x_t^2 - (\sum x_t)^2}$$

$$\widehat{\beta}_2 = \frac{-\sum x_t \sum y_t + T\sum x_t y_t}{T\sum x_t^2 - (\sum x_t)^2}$$

These can be expressed in more intuitive form. From the first equation (4.2)

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum y_t}{T} - \widehat{\beta}_2 \frac{\sum x_t}{T} \\ &= \overline{y} - \beta_2 \overline{x} \end{aligned}$$

substituting for $\widehat{\beta}_1$ in the second equation (4.3) can be written

$$\begin{aligned} \widehat{\beta}_2 \sum x_t^2 + (\overline{y} - \widehat{\beta}_2 \overline{x}) \sum x_t - \sum x_t y_t &= 0 \\ \widehat{\beta}_2 \sum x_t (x_t - \overline{x}) - \sum x_t (y_t - \overline{y}) &= 0 \end{aligned}$$

$$\widehat{\beta}_2 = \frac{\sum x_t(y_t - \overline{y})}{\sum x_t(x_t - \overline{x})} = \frac{\sum (x_t - \overline{x})(y_t - \overline{y})}{\sum (x_t - \overline{x})^2}$$

Dividing top and bottom by T, this is the ratio of the estimated covariance of $x_t$ and $y_t$ to the estimated variance of $x_t$.

Note that

$$\begin{aligned}
\sum (x_t - \overline{x})(y_t - \overline{y}) &= \sum x_t y_t + T\overline{yx} - \sum x_t \overline{y} - \sum y_t \overline{x} \\
&= \sum x_t y_t + T\frac{\sum x_t}{T}\frac{\sum y_t}{T} - \sum x_t \frac{\sum y_t}{T} - \sum y_t \frac{\sum x_t}{T} \\
&= \sum x_t(y_t - \overline{y})
\end{aligned}$$

### 4.2. Statistical properties of the LRM.

Least squares is a purely arithmetic procedure, to establish its statistical properties we need to make some statistical assumptions. Suppose we have a sample of data of observations on random variables $y_t$ a scalar and $x_t$ a $k \times 1$ vector. The joint distribution of the random variables, $y_t, x_t$, can be written as the product of the distribution of $y_t$ conditional on $x_t$ and the marginal distribution of $x_t$ :

$$D_j(y_t, x_t; \theta_j) = D_c(y_t \mid x_t; \theta_c) D_m(x_t; \theta_m) \tag{4.4}$$

$\theta_j$ is a vector of parameters of the joint distribution, $\theta_c$ of the conditional distribution, $\theta_m$ of the marginal. The distribution that we will be interested in is the distribution of $y_t$ conditional on $x_t$ and the parameters that we will be interested in are the parameters of the conditional distribution $\theta_c$ which we will usually denote by $\theta$. We will assume that the $x$ is exogenous, which means that there is no information in the marginal distribution for $x$ about the parameters of the conditional distribution that we are interested in. Usually we are only interested in the first two moments of the distribution, the conditional expectation (the regression function) and the conditional variance. If $y_t$ and $x_t$ are jointly Normally distributed, say:

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} \sim N\left( \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \right)$$

then the conditional expectation of $y_t$ is a linear function of $x_t$ :

$$E(y_t \mid x_t) = \mu_y + \left[\Sigma_{yx}\Sigma_{xx}^{-1}\right](x_t - \mu_x)$$

We can decompose $y_t$ into two components, the systematic part given by the conditional expectation and the unsystematic part, the error. The error is:

$$u_t = y_t - E(y_t \mid x_t) = y_t - \beta' x_t$$

so:

$$y_t = \beta' x_t + u_t;\ t = 1, 2, ..., T. \tag{4.5}$$

If the random variables are jointly normally distributed and the observations are independent, the conditional variance is a constant:

$$E(y_t - E(y_t \mid x_t))^2 = E(u_t^2) = \sigma^2 = \sigma_y^2 - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{4.6}$$

The parameters of the conditional distribution which we will want to estimate are $\theta_c = \theta = (\beta, \sigma^2)$.

In matrix form the $T \times T$ conditional variance covariance matrix of $y$ is:

$$E(y - E(y \mid X))(y - E(y \mid X))' = E(uu') = \sigma^2 I_T.$$

This is a $T \times T$ matrix with $\sigma^2$ on the diagonal and zeros on the off-diagonals. Distinguish $uu'$ a $T \times T$ matrix and $u'u$ the scalar sum of squared errors.

If the joint distribution of $y_t$ and $x_t$ is normal, the conditional distribution is also normal, and if the sample is independent we can write the distribution for an observation:

$$\begin{aligned} D_c(y_t \quad &| \quad x_t; \theta) \sim IN(\beta' x_t, \sigma^2) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}(\frac{y_t - \beta' x_t}{\sigma})^2\right\} \end{aligned}$$

or in matrix form:

$$D_c(y \quad | \quad X; \theta) \sim N(X\beta, \sigma^2 I) \tag{4.7}$$

$$= (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}. \tag{4.8}$$

Notice that we do not need to specify conditional independence in the matrix form, the fact that the variance covariance matrix is $\sigma^2 I$ implies that the conditional covariances between $y_t$ and $y_{t-i}$ are zero.

We need to make some assumptions about $X$. First it should be of full rank k, there should be no exact linear dependences between the columns of $X$, the various right hand side variables. This is required for $(X'X)^{-1}$ to exist. Secondly, the right hand side variables should be exogenous, roughly independent of the errors. Exogeneity is discussed in more detail below in the notes for week 9.

### 4.3. Maximum Likelihood estimation

#### 4.3.1. Introduction

Suppose we have a random variable $y$ with a known probability density function $f(y, \theta)$, where $\theta$ is a vector of parameters (e.g. mean (expected value) and variance). We can use this formula to tell us the probability of particular values of $y$, given known parameters. For instance, given that a coin has a probability of getting a head of $p = 0.5$, what is the probability of observing 10 heads in a row? Answer $(0.5)^{10}$. Alternatively, we can use the same formula to tell us the likelihood of particular values of the parameters, given that we have observed a sample of realisations of $y$, say $y_1, y_2, .., y_T$: Given that we observe ten heads in a row, how likely is it that this sample would be generated by an unbiased coin (i.e $p = 0.5$)? Again the answer is $(0.5)^{10}$. In the first case we interpret $f(y, \theta)$ as a function of $y$ given $\theta$. In the second case we interpret $f(y, \theta)$ as a function of $\theta$ given $y$. The maximum likelihood (ML) procedure estimates $\widehat{\theta}$ as the value most likely to have generated the observed sample. In the coin example, $p = 0.5$ is very unlikely to have generated the observed sample of 10 heads. If the sample is random, the observations are independent and we can just multiply the probabilities for each observation together as we did in the coin example and write the Likelihood as:

$$L(\theta) = f(y_1, \theta) f(y_2, \theta) ... f(y_T, \theta)$$

We then choose $\theta$ that maximises this value for our observed sample $y_1, y_2, ..., y_T$. It is more convenient to work with the logarithm of the likelihood function. Since logs are a monotonic function the value of $\theta$ that maximises the log-likelihood will also maximise the likelihood. Thus the log-likelihood is:

$$LL(\theta) = \sum_{t=1}^{T} \log f(y_t, \theta).$$

To find the maximum we take the derivatives of $LL(\theta)$, and set them to zero:

$$S(\widehat{\theta}) = \frac{\partial LL(\widehat{\theta})}{\partial \theta} = \frac{\partial \sum \log f(y_t, \widehat{\theta})}{\partial \theta} = 0$$

then solve for the value of $\theta, \widehat{\theta}$ that makes the derivatives equal to zero. Notice that $LL(\theta)$ is a scalar function of $\theta$, and if $\theta$ is a $k \times 1$ vector, $\frac{\partial LL(\theta)}{\partial \theta}$ will be a $k \times 1$ vector of derivatives. $S(\widehat{\theta})$ is often called the Score vector. For simple

examples, like the LRM below we can solve these equations analytically, for more complicated examples we solve them numerically. To check that we have found a maximum, we need to check the second order conditions and calculate the kxk matrix of second derivatives:

$$\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'},$$

evaluated at the true $\theta$. For a maximum this matrix should be negative definite. The information in observation $t$ is the negative of the expected value of the matrix of second derivatives:

$$I_t(\theta) = -E(\frac{\partial^2 LL_t(\theta)}{\partial\theta\ \partial\theta'})$$

which is a symmetric $k \times k$ matrix. The average information matrix in the sample of size $T$ is:

$$I_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} I_t(\theta) = -E(\frac{1}{T}\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'}).$$

A useful result is that for any unbiased estimator (in small samples) or consistent estimator (asymptotically when $T \to \infty$) the inverse of the information matrix provides a lower bound (the Cramer-Rao lower bound) on the variance covariance matrix of the estimator

$$V(\widehat{\theta}) \geq I(\widehat{\theta})^{-1}.$$

### 4.3.2. General properties of ML estimators

Under certain conditions (which usually hold in economic examples) the ML estimator $\widehat{\theta}$ is consistent, that is for some small number $\epsilon > 0$

$$\lim_{T \to \infty} \Pr(\mid \widehat{\theta}_T - \theta \mid > \epsilon) = 0.$$

The ML estimator is asymptotically normally distributed and asymptotically attains the Cramer-Rao lower bound (i.e. it is efficient), it is asymptotically $N(\theta, I(\theta)^{-1})$. $I(\widehat{\theta})^{-1}$ is often used to provide estimates of the asymptotic variance covariance matrix of the ML estimator. When we evaluate asymptotic distributions we look at $\sqrt{T}(\widehat{\theta}-\theta)$ as $T \to \infty$, because since it is consistent the distribution of $\widehat{\theta}$ collapses to a point and scale the information matrix by $T$.

$(\sqrt{T})^{-1}S(\theta)$ is also asymptotically normal $N(0, I(\theta))$. We will use these two asymptotic normality properties in testing. In addition, $E(S(\theta)S(\theta)') = T \times I(\theta)$.

ML estimators are also invariant in that for any function of $\theta$, say $g(\theta)$, the ML estimator of $g(\theta)$ is $g(\widehat{\theta})$. Partly because of this ML estimators are not necessarily unbiased. Some are, many are not.

### 4.3.3. ML estimation of the LRM

For the LRM, the likelihood of the sample is given by (4.7) above, but now interpreted as a function of $\theta = (\beta, \sigma^2)$, the unknown parameters:

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}.$$

The Log-likelihood function is :

$$LL(\beta, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

and to find the estimates that maximise this we differentiate it with respect to $\beta$ and $\sigma^2$ and set the derivatives equal zero. Notice that

$$u'u = (y - X\beta)'(y - X\beta) = y'y + \beta'X'X\beta - 2\beta'X'y.$$

When we transpose we reverse the order to maintain the correct dimensions and $\beta'X'y = y'X\beta$ because both are scalars. Thus:

$$\frac{\partial LL(\beta, \sigma^2)}{\partial \beta} = -\frac{1}{2\sigma^2}(2X'X\beta - 2X'y) \tag{4.9}$$

and

$$\frac{\partial LL(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}u'u. \tag{4.10}$$

The derivative with respect to $\sigma^2$ of $\log(\sigma^2)$ is $1/\sigma^2$ and of $-1/2\sigma^2 = -(2\sigma^2)^{-1}$ is $(-1)(-(2\sigma^2)^{-2})$.

Setting (4.9) equal to zero gives one First Order Conditions, FOC

$$\begin{aligned} -\frac{1}{2\widehat{\sigma}^2}(2X'X\widehat{\beta} - 2X'y) &= 0 \\ \frac{1}{\widehat{\sigma}^2}(X'y - X'X\widehat{\beta}) &= 0 \end{aligned}$$

where the hats denote that these are the values of $\beta$ and $\sigma^2$ that make the FOCs equal to zero. Notice that this can be written

$$\frac{1}{\widehat{\sigma}^2}X'(y - X\widehat{\beta}) = \frac{1}{\widehat{\sigma}^2}X'\widehat{u} = 0 \tag{4.11}$$

the first order conditions choose $\beta$ that makes the estimated residuals, $\widehat{u} = y - X\widehat{\beta}$, uncorrelated with (orthogonal to) the explanatory variables. This estimate is

$$\widehat{\beta} = (X'X)^{-1}X'y.$$

Notice that we need $X$ to be of full rank for the inverse of $(X'X)$ to exist. If $(X'X)$ is singular, $\widehat{\beta}$ is not defined. This is called exact multicollinearity.

Setting (4.10) equal to zero gives

$$-\frac{T}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\widehat{u}'\widehat{u} = 0$$

multiply through by $2\widehat{\sigma}^4$

$$-T\widehat{\sigma}^2 + \widehat{u}'\widehat{u} = 0$$

so our maximum likelihood estimator of the variance is:

$$\widehat{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{T}.$$

The ML estimator is biased and we usually use the unbiased estimator $s^2 = \widehat{u}'\widehat{u}/(T-k)$.

To check second order conditions and construct the information matrix we take derivates of (4.9) and (4.10)

$$\frac{\partial^2 LL(\beta, \sigma^2)}{\partial\beta\ \partial\beta'} = -\frac{1}{\sigma^2}X'X \tag{4.12}$$

$$\frac{\partial LL(\beta, \sigma^2)}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4}X'u. \tag{4.13}$$

Notice the derivative of $(\sigma^2)^{-1}X'u$ is $-(\sigma^2)^{-2}X'u$. Finally

$$\frac{\partial^2 LL(\beta, \sigma^2)}{\partial(\sigma^2)^2} = \frac{T}{2\sigma^4} - \frac{u'u}{\sigma^6}. \tag{4.14}$$

To get the information matrix we take the negative of the expected value of the second derivative matrix. Notice that $E(X'u) = 0,\ E(u'u) = T\sigma^2$ so the expected value of the final second derivative can be written:

$$\frac{T}{2\sigma^4} - \frac{T\sigma^2}{\sigma^6} = \frac{T}{2\sigma^4} - \frac{T}{\sigma^4} = -\frac{T}{2\sigma^4}$$

$$I(\theta) = -E(\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'}) = \begin{bmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix}$$

$$I(\beta, \sigma^2)^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{T} \end{bmatrix}.$$

This gives the lower bound for the Variance-covariance matrix for estimators of $\beta, \sigma^2$. Notice that the estimators of $\beta$ and $\sigma^2$ are independent, their covariances are zero. But there will be non-zero covariances between the elements of $\widehat{\beta}$.

We can put the ML estimates into the Log-likelihood function, to get the Maximised Log-Likelihood, MLL, reported by most programs

$$\begin{aligned} MLL &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2}\widehat{u}'\widehat{u} \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\widehat{\sigma}^2) - \frac{T\widehat{\sigma}^2}{2\widehat{\sigma}^2} \\ &= -\frac{T}{2}(\log(2\pi) + 1) - \frac{T}{2}\log(\widehat{\sigma}^2) \end{aligned}$$

apart from the constant this is just the negative of half the sample size times the log of the ML estimate of the variance. This can be negative or positive.

### 4.3.4. Properties of the ML estimators in the LRM

General asymptotic properties of ML estimators were discussed above, to derive the specific small sample properties of the LRM estimators we will use two results repeatedly.

Firstly, linear functions of normally distributed variables are normally distributed. If $y$ is $IN(\mu, \sigma^2)$ then $a + by$ is $N(a + b\mu, b^2\sigma^2)$. The multivariate generalisation of this is that if the $T \times 1$ vector $Y \sim N(M, \Sigma)$, where $M$ is $T \times 1$, $\Sigma$ is a $T \times T$ variance covariance matrix. Then for given $A$ and $B$ of order $K \times 1$ and $K \times T$ :

$$A + BY \sim N(A + BM, B\Sigma B'). \tag{4.15}$$

Secondly, quadratic forms (sums of squares) of standardised normally distributed variables are Chi-squared. So

$$\sum_{t=1}^{T}\left(\frac{y_i-\mu}{\sigma}\right)^2 \sim \chi^2(T)$$

and for the $T \times 1$ vector $Y \sim N(M, \Sigma)$, then:

$$(Y-M)'\Sigma^{-1}(Y-M) \sim \chi^2(T) \tag{4.16}$$

distributed as Chi-squared with $T$ degrees of freedom.

**Regression Coefficients** To return to the LRM, where $y$ is a $T \times 1$ vector and $X$ is a $T \times k$ full rank matrix of exogenous variables, then conditional on $X$, since

$$y \sim N(X\beta, \sigma^2 I)$$

and the ML estimator is

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

a linear function of $y$, $\widehat{\beta}$ is normally distributed using (4.15):

$$\begin{aligned}\widehat{\beta} &\sim N\{(X'X)^{-1}X'X\beta,\ (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1}\}\\ &\sim N\{\beta, \sigma^2(X'X)^{-1}\}\end{aligned}$$

This indicates (1) $\widehat{\beta}$ is unbiased, $E(\widehat{\beta}) = \beta$, (2) it is fully efficient, its variance covariance matrix attains the lower bound obtained above $\sigma^2(X'X)^{-1}$. We generally estimate the variance covariance matrix by $s^2(X'X)^{-1}$, where $s^2 = \widehat{u}'\widehat{u}/(T-k)$, the unbiased estimator. The square roots of the diagonal elements of this matrix give the standard errors of the individual regression coefficients, e.g. $\beta_i$ and the off diagonal elements give the covariances between regression coefficients, e.g. $Cov(\beta_i, \beta_j)$.

**Residuals** The estimated residuals are uncorrelated with the explanatory variables by construction:

$X'\widehat{u} = X'(y - X\widehat{\beta}) = X'(y - X(X'X)^{-1}X'y) = X'y - X'y = 0.$

$X'\widehat{u}$ is a set of k equations of the form:

$$\sum_{t=1}^{T}\widehat{u}_t = 0; \sum_{t=1}^{T}x_{2t}\widehat{u}_t = 0; ...; \sum_{t=1}^{T}x_{kt}\widehat{u}_t = 0.$$

In addition:

$$\widehat{u} = y - X\widehat{\beta} = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = (I - P_x)y = My.$$

$P_x$ is a 'projection matrix' it is symmetric and idempotent ($P_xP_x = P_x$) and orthogonal to $M$, ($P_xM = 0$), which is also symmetric and idempotent. So

$$y = P_xy + My$$

it is split into two orthogonal components, the projection of $y$ on $X$ and the orthogonal remainder.

Notice that the estimated residuals are a transformation of the true disturbances:

$$\begin{aligned} \widehat{u} &= (I - X(X'X)^{-1}X')y = (I - X(X'X)^{-1}X')(X\beta + u) \\ &= (I - X(X'X)^{-1}X')u = Mu. \end{aligned}$$

We cannot recover the true disturbances from this equation since M is singular, rank T-k. The sum of squared residuals is:

$$\sum_{t=1}^{T} \widehat{u}_t^2 = \widehat{u}'\widehat{u} = u'M'Mu = u'Mu.$$

To calculate the expected value of the sum of squared residuals (strictly conditional on $X$ which has not been made explicit), note that $\widehat{u}'\widehat{u}$ is a scalar, thus equal to its trace, the sum of its diagonal elements. Thus using the properties of traces we can write

$$\begin{aligned} E(\widehat{u}'\widehat{u}) &= E(u'Mu) = E(tr(u'Mu)) = E(tr(Muu')) \\ &= tr(M\sigma^2 I) = \sigma^2 tr(M) = \sigma^2(T-k). \end{aligned}$$

Thus the unbiased estimate of $\sigma^2$ is $s2 = \widehat{u}'\widehat{u}/(T-k)$. The last step uses the fact that the Trace of $M$ is

$$\begin{aligned} tr\left[I_T - X(X'X)^{-1}X'\right] &= tr(I_T) - tr\left(X(X'X)^{-1}X'\right) \\ &= tr(I_T) - tr\left((X'X)^{-1}X'X\right) \\ &= tr(I_T) - tr(I_K) = T - k \end{aligned}$$

The sum of squared standardised original disturbances $u'u/\sigma^2$ are distributed as $\chi^2(T)$, but the sum of squared standardised residuals $\widehat{u}'\widehat{u}/\sigma^2 = u'Mu/\sigma^2$ are $\chi^2(rankM) = \chi^2(T-k)$. Alternatively

$$(T-k)(\frac{s^2}{\sigma^2}) \sim \chi^2(T-k).$$

**4.4. What happens when assumptions fail.**

(a) If $X$ is not of full rank k, because there is an exact linear dependency between some of the variables, the OLS/ML estimates of $\beta$ are not defined and there is said to be exact multicollinearity. The model should be respecified to remove the exact dependency. When there is high, though not perfect, correlation between some of the variables there is said to be multicollinearity. This does not involve a failure of any assumption.

(b) If the $X$ are not strictly exogenous the estimates of $\beta$ are biased, though if the $X$ are predetermined (e.g. lagged dependent variables) and the disturbance term is not serially correlated, they will remain consistent. Otherwise, they will be inconsistent. In certain circumstances failure of the exogeneity assumptions can be dealt with by the method of Instrumental Variables discussed below.

(c). If normality does not hold and the form of the distribution is not known the Least Squares estimator, $\widehat{\beta} = (X'X)^{-1}X'y$, is no longer the Maximum Likelihood estimator and is not fully efficient, but it is the minimum variance estimator in the class of linear unbiased estimators (biased or non-linear estimators may have smaller variances). In small samples, the tests below will not have the stated distributions, though asymptotically they will be normal. If the form of the distribution is known (e.g. a t distribution) maximum likelihood estimators can be derived for that particular distribution and they will be different from the OLS estimators. EViews and Microfit will estimate model with errors distributed as t, under the GARCH options. For small degrees of freedom, the t has fatter tails, when the degrees of freedom are aound 30 it is close to normal.

(d) If $y \sim N(X\beta, \sigma^2\Omega)$, that is its variance covariance matrix is not $\sigma^2 I$, there are two possible problems: the variances (diagonal terms of the matrix) are not constant and equal to $\sigma^2$ (heteroskedasticity) and/or the off diagonal terms, the covariances, are not equal to zero (failure of independence, serial correlation, autocorrelation). Under these circumstances, $\widehat{\beta}$ remains unbiased but is not minimum variance (efficient). Its variance-covariance matrix is not $\sigma^2(X'X)^{-1}$, but $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$. Corrected variance-covariance matrices are available in most packages (White Heteroskedasticity consistent covariance matrices or Newey-West autocorrelation consistent ones). These use estimates of $X'\Omega X$ in the formula. Use Options on the equation menu in EViews to get HAC (Heteroskedasticy and Autorcorrelation Consistent) standard errors. Notice that residual serial correlation or heteroskedasticity may indicate not that there is some covariances between the true disturbances but that the model is wrongly specified, e.g. vari-

ables are omitted, see below. When it is appropriate to model the disturbance structure in terms of $\Omega$, Generalised Least Squares, discussed below, can be used. Often residual serial correlation or heteroskedasticity should lead you to respecify the model rather than to use Generalised Least Squares.

(e) Omitted variables. Suppose the data are generated by

$$y_t = \beta' x_t + \gamma' z_t + u_t \tag{4.17}$$

and you omit $z_t$, an hx1 vector and estimate

$$y_t = \alpha' x_t + v_t. \tag{4.18}$$

What is the relationship between the estimates? Suppose we describe the relation between the omitted and included right hand side variables by the multivariate regression model:

$$z_t = B x_t + w_t \tag{4.19}$$

where $B$ is an hxk matrix. This is just a set of h regressions in which each $z_t$ is regressed on all $k$ $x_t$. If you replace $z_t$ in (4.17) by the right hand side of (4.19) you get:

$$y_t = \beta' x_t + \gamma'(B x_t + w_t) + u_t$$

$$y_t = (\beta' + \gamma' B) x_t + (\gamma' w_t + u_t).$$

Thus $\alpha = (\beta' + \gamma' B)$ and $v_t = (\gamma' w_t + u_t)$. The coefficient of $x_t$ in (4.18) will only be an unbiased estimator of $\beta$, the coefficient of $x_t$ in (4.17) if either $\gamma = 0$ ($z_t$ really has no effect on $y_t$) or $B = 0$, (there is no correlation between the included and omitted variables). Notice that $v_t$ also contains the part of $z_t$ that is not correlated with $x_t$, $w_t$, and there is no reason to expect $w_t$ to be serially uncorrelated or homoskedastic. Thus misspecification, omission of $z_t$, may cause the estimated residuals to show these problems.

### 4.4.1. Generalised Least Squares

If $y \sim N(X\beta, \sigma^2\Omega)$ its distribution is given by:

$$2\pi^{-T/2} \mid \sigma^2\Omega \mid^{-1/2} \exp\left\{-\frac{1}{2}(y - X\beta)'(\sigma^2\Omega)^{-1}(y - X\beta)\right\}.$$

Notice that when $\Omega = I$, then the term in the determinant, $\mid \sigma^2\Omega \mid^{-1/2}$ is just $(\sigma^2)^{-T/2}$.

If $\Omega$ is a known matrix the Maximum Likelihood Estimator is the Generalised Least Squares estimator $\beta^{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, with variance-covariance matrix $V(\beta^{GLS}) = \sigma^2(X'\Omega^{-1}X)^{-1}$. Whereas the OLS estimator chooses $\beta$ to make $(\sigma^2)^{-1}X'\widehat{u} = 0$, the GLS estimator chooses $\beta$ to make $(\sigma^2)^{-1}X'\Omega^{-1}\widetilde{u} = 0$, where $\widetilde{u} = y - X\beta^{GLS}$. In practice GLS is implemented by finding a 'transformation matrix' $P$ such that $P'P = \Omega^{-1}$ and $P\Omega P' = I$. This can always be done since $\Omega$ must be a positive-definite symmetric matrix. You then transform the data by premultiplying the equation by $P$

$$Py = PX\beta + Pu$$

$$y^* = X^*\beta + u^*$$

where $y^* = Py$, etc. OLS is then applied to the transformed data, which is fully efficient since

$$E(u^*u^{*\prime}) = E(Puu'P') = PE(uu')P' = P(\sigma^2\Omega)P' = \sigma^2 P\Omega P' = \sigma^2 I.$$

In practice, $\Omega$ is rarely known completely, but it may be known up to a few unknown parameters. These can be estimated and used to form an estimate of $\Omega$, and $P$. This is known as the Feasible or Estimated GLS estimator. It generally differs from the exact ML estimator. The text books give large number of examples of FGLS estimators, differing in the assumed structure of $\Omega$. But in many cases it is better to respecify the model or correct the standard errors than to apply FGLS to try and fix problems with the residuals.

# 5. Testing

## 5.1. Introduction

Suppose that we have prior information on $\theta$, which suggests that elements of the parameter vector take specified values, such as zero or one or are linked by other restrictions, and we wish to test this hypothesis. A test involves:

(a) a null hypothesis usually called $H_0$; e.g. for a scalar parameter: $H_0 : \beta = 1$;

(b) an alternative hypothesis, e.g. $H_1 : \beta \neq 1$, this is a two sided alternative, a one sided alternative would be $\beta < 1$;

(c) a test statistic, which does not depend on the true value of the parameters (is pivotal), (e.g. $(\widehat{\beta} - 1)/SE(\widehat{\beta})$, where $SE(\widehat{\beta})$ is the estimated standard error

of $\widehat{\beta}$) with a known distribution when the null hypothesis is true (e.g. a central t distribution);

(d) a specified size $\alpha$, the chosen probability of Type I error (rejecting $H_0$ when it is true) usually 0.05;

(e) critical values so that if the null hypothesis is true the probability of lying outside the critical values is $\alpha$;

(f) a power function which gives the probability of rejecting the null as a function of the true (unknown) value of $\beta$. The power of a test is the probability of rejecting $H_0$ when it is false (one minus the probability of type two error).

The procedure is: to not reject $H_0$ if the test statistic lies within the critical values and to reject $H_0$ if the test statistic lies outside the critical values. Notice that rejecting the null does not mean accepting the alternative. The results can also be presented as p values, which can be thought of as giving the probability that the hypothesis is true. If the p value is small, less than the chosen size (probability of rejecting null when true), e.g. 0.05, then the null hypothesis is rejected.

The test asks whether the difference of the estimate from the null hypothesis could have arisen by chance, it does not tell you whether the difference is important, therefore you should distinguish substantive (economic) importance from statistical significance. A coefficient may be statistically significant because it is very precisely estimated but so small as to be of no economic importance. Conversely the coefficient may be large in economic terms but have large standard errors so not be statistically significant. It is also useful to think of a test as informing a decision, accepting or rejecting the null and considering the costs of the two sorts of mistakes. The costs can be embodied in some form of loss function or utility function.

### 5.2. Exact Tests

In the LRM with linear restrictions we can derive exact small sample tests. Suppose, our null hypothesis is a set of m linear restrictions of the form $R\beta = q$ or $R\beta - q = 0$, where $R$ and $q$ are known and of order $m \times k$ and $m \times 1$ respectively. The unrestricted model has k parameters, the restricted model k-m, each restriction reduces the number of parameters we estimate. In the case where m=k, all the parameters are specified, $R$ is an identity matrix and the restrictions are $\beta = q$.

Since

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

and the restrictions are linear

$$(R\widehat{\beta} - q) \sim N(R\beta - q, \sigma^2 R(X'X)^{-1}R')$$

Under $H_0 : R\beta - q = 0$

$$(R\widehat{\beta} - q) \sim N(0, \sigma^2 R(X'X)^{-1}R')$$

and

$$(R\widehat{\beta} - q)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\widehat{\beta} - q) \sim \chi^2(m).$$

Notice that this is a special case of the Wald test statistic below and is of the same form. This is not yet a test statistic because it depends on the unknown $\sigma^2$, but we know $(T-k)s^2/\sigma^2 \sim \chi^2(T-k)$ and that for independent Chi-squares:

$$\frac{\chi^2(m)/m}{\chi^2(T-k)/T-k} \sim F(m, T-k)$$

so

$$\frac{(R\widehat{\beta} - q)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\widehat{\beta} - q)/m}{[(T-k)s^2/\sigma^2]/(T-k)} \sim F(m, T-k)$$

or since the two unknown $\sigma^2$ cancel:

$$\frac{(R\widehat{\beta} - q)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta} - q)/m}{s^2} \sim F(m, T-k).$$

This provides us with a test statistic. In practice it is easier to calculate it from another way of writing this formula. Define the unrestricted and restricted estimated equations as

$$y = X\widehat{\beta} + \widehat{u}; \quad \text{and} \quad y = X\beta^* + u^*$$

then

$$\frac{(u^{*\prime}u^* - \widehat{u}'\widehat{u})/m}{\widehat{u}'\widehat{u}/(T-k)} \sim F(m, T-k),$$

the ratio of (a) the difference between the restricted and unrestricted sum of squared residuals divided by the number of restrictions to (b) the unbiased estimate of the unrestricted variance. Computer programs automatically print out a test for the hypothesis that all the slope coefficients in a linear regression are zero, this is $F(k-1, T-k)$.

### 5.3. Asymptotic Tests

Often we cannot derive exact, small sample tests and use aymptotic approximations. We saw above that the ML estimates are those which maximise $LL(\theta)$, i.e. the $\widehat{\theta}$, which make

$$\frac{\partial LL(\theta)}{\partial \theta} = S(\widehat{\theta}) = 0$$

where $S(\widehat{\theta})$ is the score vector, the derivatives of the LL with respect to each of the k elements of the vector $\theta$ evaluated at the values, $\widehat{\theta}$, which make $S(\theta) = 0$. We will call these the unrestricted estimates and the value of the Log-likelihood at $\widehat{\theta}$, $LL(\widehat{\theta})$.

Suppose theory suggests $m \leq k$ prior restrictions (possibly non-linear) of the form $R(\theta) = 0$, where $R(\theta)$ is an $m \times 1$ vector. If $m = k$, theory specifies all the parameters and there are none to estimate. The restricted estimates maximises

$$\pounds = LL(\theta) - \lambda' R(\theta)$$

where $\lambda$ is a $m \times 1$ vector of Lagrange Multipliers. The first order condition, FOC, is

$$\frac{\partial \pounds}{\partial \theta} = \frac{\partial LL(\theta)}{\partial \theta} - \frac{\partial R(\theta)}{\partial \theta}\lambda = 0$$

Write the $kx1$ vector $\partial LL(\theta)/\partial\theta$ as $S(\theta)$ and the $k \times m$ matrix $\partial R(\theta)/\partial\theta$ as $F(\theta)$ then at the restricted estimate $\theta^*$, which makes the FOC hold

$$S(\theta^*) - F(\theta^*)\lambda^* = 0$$

Notice that at $\theta^*$ the derivative of the Log-likelihood function with respect to the parameters is not equal to zero but to $F(\theta^*)\lambda^*$. The value of the Log-likelihood at $\theta^*$ is $LL(\theta^*)$ which is less than or equal to $LL(\widehat{\theta})$.

If the hypotheses (restrictions) are true:

(a) the two log-likelihoods should be similar, i.e. $LL(\widehat{\theta}) - LL(\theta^*)$ should be close to zero;

(b) the unrestricted estimates should satisfy the restrictions $R(\widehat{\theta})$ should be close to zero (note $R(\theta^*)$ is exactly zero by construction);

(c) the restricted score, $S(\theta^*)$, should be close to zero (note $S(\widehat{\theta})$ is exactly zero by construction) or equivalently the Lagrange Multipliers $\lambda^*$ should be close to zero, the restrictions should not be binding.

These implications are used as the basis for three types of test procedures. The issue is how to judge 'close to zero'? To judge this we use the asymptotic

equivalents of the linear distributional results used above in the discussion of the properties of the LRM. Asymptotically the ML estimator is normal

$$\widehat{\theta} \ \widetilde{a} \ N(\theta, I(\theta)^{-1})$$

asymptotically the scalar quadratic form is chi-squared

$$(\widehat{\theta} - \theta)' I(\theta)(\widehat{\theta} - \theta) \ \widetilde{a} \ \chi^2(k).$$

and asymptotically $R(\widehat{\theta})$ is also normal

$$R(\widehat{\theta}) \ \widetilde{a} \ \ N(R(\theta), F(\theta)' I(\theta)^{-1} F(\theta))$$

where $F(\theta) = \partial R(\theta)/\partial\theta$. This gives us three procedures for generating asymptotic test statistics for the m restictions $H_0 : R(\theta) = 0$; each of which are asymptotically distributed $\chi^2(m)$, when the null hypothesis is true:

(a) Likelihood Ratio Tests

$$LR = 2(LL(\widehat{\theta}) - LL(\theta^*)) \sim \chi^2(m)$$

(b) Wald Tests

$$W = R(\widehat{\theta})'[F(\theta)' I(\theta)^{-1} F(\theta)]^{-1} R(\widehat{\theta}) \sim \chi^2(m)$$

where the term in [...] is an estimate of the variance of $R(\widehat{\theta})$ and $F(\theta) = \partial R(\theta)/\partial\theta$.

(c) Lagrange Multiplier (or Efficient Score) Tests where $\partial LL(\theta)/\partial\theta = S(\theta)$

$$LM = S(\theta^*)' I(\theta^*)^{-1} S(\theta^*) \sim \chi^2(m).$$

The Likelihood ratio test is straightforward to calculate when both the restricted and unrestricted models have been estimated. The Wald test only requires the unrestricted estimates. The Lagrange Multiplier test only requires the restricted estimates. For the LRM, the inequality W>LR>LM holds, so you are more likely to reject using W. In the LRM, the LM test is usually calculated using regression residuals as is discussed below. The Wald test is not invariant to how you write non-linear restrictions. Suppose $m = 1$, and $R(\theta)$ is $\theta_1\theta_2 - \theta_3 = 0$. This could also be written $\theta_1 - \theta_3/\theta_2 = 0$ and these would give different values of the test statistic. The former form, using multiplication rather than division, is usually better.

### 5.4. Model Selection Procedures

Hypothesis tests require the two models being compared to be 'nested': one model (the restricted model) must be a special case of the other (the unrestricted or maintained model). In many cases we want to compare 'non-nested' models, e.g.

$$\begin{aligned} M_1 &: y_t = a_1 + b_1 x_t + u_{1t} \\ M_2 &: y_t = a_2 + c_2 z_t + u_{2t} \end{aligned}$$

where $x_t$ and $z_t$ are different scalar variables. There are no restrictions on $M_1$ that will give $M_2$ and vice-versa. We could nest them both in a general model:

$$M_3 : y_t = a_3 + b_3 x_t + c_3 z_t + u_{3t}.$$

The restriction $c_3 = 0$ gives $M_1$; so rejecting the restriction $c_3 = 0$ rejects $M_1$
The restriction $b_3 = 0$ gives $M_2$; so rejecting the restriction $b_3 = 0$ rejects $M_2$.
This gives four possible outcomes:

1. Reject $M_1$, do not reject $M_2 : c_3 \neq 0; b_3 = 0;$
2. Reject $M_2$, do not reject $M_1 : b_3 \neq 0; c_3 = 0;$
3. Reject both; $b_3 \neq 0; c_3 \neq 0;$
4. Do not reject either: $b_3 = 0; c_3 = 0.$

There are a range of other non-nested tests available (Microfit has a large selection) but they all give rise to the same four possibilities. If $x_t$ and $z_t$ are highly correlated case 4 is quite likely. Notice that these are based on individual tests (t tests), joint tests may give conflicting answers. On individual tests we could reject both the hpothesis $b_3 = 0$; and $c_3 = 0$, i.e. both have significant t ratios, but we could not reject the joint hypothesis that they are both equal to zero. Conversely, they could be individually insignificant but jointly significant.

An alternative approach is not to test but to choose the 'best' model on some 'model selection' criterion. As with British newspapers, the most popular are the worst. The most popular are $R^2$ and $\overline{R}^2$. Treat them with the scepticism you would give to a story in the Sun.

Better criteria for choosing between various models are the Akaike Information Criterion ($AIC_i = MLL_i - k_i$); and the Schwarz Bayesian Information Criterion or Posterior Odds Criterion ($SBC = MLL_i - 0.5k_i \log T$); where $MLL_i$ is the

maximised log likelihood of model i, $k_i$ is the number of parameters estimated in model i, and $T$ is the sample size. You choose the model with the largest value. The SBC tends to choose a more parsimonious model (fewer parameters).

About half of statistics programs (including Microfit) define the AIC in terms of the MLL, in which case you choose the model with the largest value. The other half (including EViews) define it equivalently in terms of the sum of squared residuals, in which case you choose the model with the smallest value. Be careful, which way they are defined.

## 6. Diagnostic Tests

The estimates one gets of a model are only valid if a number of assumptions hold and it is important to test those assumptions. Such tests are called diagnostic or misspecification tests. Failure on a particular diagnostic test (rejection of the null that the model is well specified) only indicates that the model is sick, it does not tell you what the illness is. For instance, if you have chosen the wrong functional form you may fail tests for serial correlation. Apart from the structural stability tests most of these tests are Lagrange Multiplier tests which involve auxilliary regressions using the residuals from the first stage regressions. These tests ask whether the residuals have the properties we would expect if the assumptions were true. The null hypothesis is always that the assumptions are true, the model is well specified. Thus if the p value for the test is greater than 0.05, you can accept the hypothesis that the model is well specified at the 5% level.

There are a very large numbers of these tests for serial correlation and non-linearity, which use the residuals as the dependent variable; for heteroskedasticity, which use the squared residuals as the dependent variable; and for normality which check that the third and fourth moments of the residuals have the values they should have under normality. For each null, e.g. constant variance (homoskedasticity) there are a large number of different alternatives (ways that the variance changes) thus lots of different tests for heteroskedasticity of different forms. Although the justification of these tests is asymptotic, versions which use F tests and degrees of freedom adjustment seem to work well in practice. In Microfit four of these tests are provided automatically in EViews they are available on the View menu after a regression. See the applied exercise for details. Always inspect graphs of actual and predicted values and residuals.

### 6.1. Structural stability

The assumption that the parameters are constant over the sample is crucial and there are a variety of tests for constancy. Two are special cases of the F test for linear restrictions above.

Suppose that we have a sample of data for $t = 1, 2, ..., T$ and we believe that the relationship may have shifted at period $T_1$ within the sample, and both sub-samples have more than $k$ observations. The unrestricted model estimates separate regressions for each sub period $t = 1, 2, ..., T_1$ and for $t = T_1 + 1, T_1 + 2, ..., T$; define $T_2 = T - T_1$: $X_1$ a $T_1 \times k$ matrix, $X_2$ a $T_2 \times k$ matrix, etc. Then the models for the two subperiods are:

$$\begin{aligned} y_1 &= X_1\beta_1 + u_1 \\ y_2 &= X_2\beta_2 + u_2 \end{aligned}$$

where we assume $u_i \sim IN(0, \sigma^2)$, $i = 1, 2$; the variances are the same in both periods. The unrestricted residual sum of squares is $(\widehat{u}_1'\widehat{u}_1 + \widehat{u}_2'\widehat{u}_2)$ with degrees of freedom $T - 2k$. The restricted model is

$$y = X\beta + u$$

where $X$ is a $T \times k$ matrix. The restricted residual sum of squares is $\widehat{u}'\widehat{u}$ with degrees of freedom $T - k$. The null hypothesis is that $\beta_1 = \beta_2$, $k$ restrictions and the test statistic is

$$\frac{[\widehat{u}'\widehat{u} - (\widehat{u}_1'\widehat{u}_1 + \widehat{u}_2'\widehat{u}_2)]/k}{(\widehat{u}_1'\widehat{u}_1 + \widehat{u}_2'\widehat{u}_2)/(T - 2k)} \sim F(k, T - 2k).$$

This is known as Chow's first or breakpoint test. He also suggested a second 'predictive failure' or forecast test for the case where there $T_2 < k$ though it can be used whether or not there are enough observations to estimate the second period model. The test statistic is:

$$\frac{[\widehat{u}'\widehat{u} - \widehat{u}_1'\widehat{u}_1)]/T_2}{\widehat{u}_1'\widehat{u}_1/(T_1 - k)} \sim F(T_2, T_1 - k).$$

This tests the hypothesis that in

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ X_2 & I \end{bmatrix} \begin{bmatrix} \beta_1 \\ \delta \end{bmatrix} + \begin{bmatrix} u_1 \\ 0 \end{bmatrix}$$

$\delta$ the $T_1 \times 1$ vector of forecast errors are not significantly different from zero. This has a dummy variable for each observation in the second period.

Chow's first test assumes that the variances in the two periods are the same. This can be tested using the Variance Ratio or 'Goldfeld-Quandt' test:

$$\frac{s_1^2}{s_2^2} = \frac{\widehat{u}_1'\widehat{u}_1/(T_1-k)}{\widehat{u}_2'\widehat{u}_2/(T_2-k)} \sim F(T_1-k, T_2-k).$$

You should put the larger variance on top so the F statistic is greater than unity. Notice that although this is an F test, it is not a test of linear restrictions on the regression parameters like the other F tests we have used. This is a test for a specific form of heteroskedasticity, tests for other types of heteroskedasticity are given below. If the variances are equal, the two equations can be estimated together using dummy variables

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

this will give the same estimates of the coefficients as running two separate regressions, but different estimators of the standard errors: this form imposes equality of variables, the separate regressions do not. For testing differences of individual coefficients, this can be rewritten

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ X_2 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 - \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

Then some can be allowed to differ and others kept the same between periods.

Packages like EViews also include a range of other ways to investigate structural stability of the parameters using recursive residuals such as the CUSUM and CUSUMSQ diagrams, which are particularly useful when one is uncertain about the breakpoint. These are presented as graphs of the statistics within two lines. If the graphs cross the lines it indicates structural instability. They also present recursive estimates, where the parameters are estimated on the first $k+1$ observations, the first $k+2$ and so on up to $T$. Breaks may show up in the estimates. They also give Andrews-Quandt tests which identify the most likely place for a break. Tests with an unknown break-point will have much less power than tests with a known break-point.

### 6.2. Serial Correlation

Suppose the data were generated by:

$$\begin{aligned} y_t &= \beta' x_t + v_t;\ v_t = \rho v_{t-1} + u_t \\ y_t &= \beta' x_t + \rho v_{t-1} + u_t \end{aligned}$$

where $u_t$ is a 'well-behaved' disturbance distributed $IN(0, \sigma^2)$; but we estimate

$$y_t = b' x_t + v_t$$

the estimated residuals

$$\widehat{v}_t = y_t - \widehat{b}' x_t = \beta' x_t + \rho v_{t-1} + u_t - \widehat{b}' x_t$$

$$\widehat{v}_t = (\beta - \widehat{b})' x_t + \rho v_{t-1} + u_t$$

we could test the hypothesis that $\rho = 0$, there is no serial correlation by running a regression of the estimated residuals on the regressors and the lagged residuals:

$$\widehat{v}_t = c' x_t + \rho \widehat{v}_{t-1} + u_t$$

and testing $\rho = 0$ with a t test. We replace the missing residuals (for period zero here) by their expected value zero. If we think there may be higher order correlations, we can add more lagged residuals and test the joint hypothesis that all the coefficients of the lagged residuals are zero, with an F test. For instance, if we have quarterly data, we would be interested in testing for up to fourth order serial correlation, i.e. all $\rho_i = 0, i = 1, 2, ..., 4$ in:

$$\widehat{v}_t = c' x_t + \rho_1 \widehat{v}_{t-1} + \rho_2 \widehat{v}_{t-2} + \rho_3 \widehat{v}_{t-3} + \rho_4 \widehat{v}_{t-4} + u_t$$

This is a different alternative hypothesis to that of no first order serial correlation, but the null hypothesis is the same.

### 6.3. Non-linearity.

#### 6.3.1. Linear in parameters and not in variables

Suppose we are explaining the logarithm of wages, $w_i$, of a sample of men $i = 1, 2, ..., N$ by age, A, and years of education, E. This is certainly not linear, at some age wages peak and then fall with age thereafter, similarly with education:

getting a PhD reduces expected earnings in the UK. In addition, the variables interact, wages peak later in life for more educated people. This suggests a model of the form:

$$w_i = a + bA_i + cA_i^2 + dE_i + eE_i^2 + fE_iA_i + u_i$$

This model is linear in parameters though it is non-linear in the variables and can be estimated by OLS on the transformed data. We expect, $b, d, f > 0$ and $c, e < 0$. The age at which earnings is maximised is given by the solution to:

$$\frac{\partial w}{\partial A} = b + 2cA + fE = 0$$

$$A^* = -\frac{b + fE}{2c}.$$

which if the estimated coefficients have the expected signs is positive (since $c < 0$) and peak earning age increases with education.

### 6.3.2. Testing for non-linearity

Here we have strong prior reasons to include squares and cross products. In other cases we do not, but just want to check whether there is a problem. Adding squares and cross-products can also use up degrees of freedom very fast. If there are k regressors, there are k(k+1)/2 squares and cross products, for k=5, 15 additional regressors. This is fine in large cross sections with thousands of observations, but in small samples it is a problem. Instead, we estimate a first stage linear regression:

$$y_t = \widehat{\beta}' x_t + \widehat{u}_t \tag{6.1}$$

with fitted values $\widehat{y}_t = \widehat{\beta}' x_t$; and run a second stage regression:

$$\widehat{u}_t = b' x_t + c\widehat{y}_t^2 + v_t$$

and test whether $c$ is signficantly different from zero.

Eviews does this Ramsey RESET test slightly differently. It runs

$$y_t = d' x_t + e\widehat{y}_t^2 + v_t$$

and tests whether $e$ is significantly different from zero. Noting that

$$y_t = \widehat{y}_t + \widehat{u}_t = \widehat{\beta}' x_t + \widehat{u}_t,$$

show that they give identifical test statistics.

Higher powers of $\widehat{y}_t$ can also be added. Notice that $\widehat{y}_t^2$ is being used as a measure of squares and cross-products of the $x_t$. For the simple model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

$$\begin{aligned} \widehat{y}_t^2 &= (\widehat{\beta}_1 + \widehat{\beta}_2 x_{2t} + \widehat{\beta}_3 x_{3t})^2 \\ &= \widehat{\beta}_1^2 + \widehat{\beta}_2^2 x_{2t}^2 + \widehat{\beta}_3^2 x_{3t}^2 + 2\widehat{\beta}_1 \widehat{\beta}_2 x_{2t} + ... \end{aligned}$$

Tests which use powers of the fitted values in this way are often known as RESET tests.

### 6.3.3. Non-linear in parameters

If our dependent variable is a proportion, $p_t$ taking values between zero and one, the logistic transformation is often used $\ln(p_t/(1-p_t)$. If this is made a function of time,

$$\ln\left(\frac{p_t}{1-p_t}\right) = a + bt + u_t$$

this gives an S shaped curve for $p_t$ over time, which often gives a good description of the spread of a new good (e.g. the proportion of the population that have a mobile phone) and can be estimated by least squares, since it is linear in the parameters. The form of the relationship is

$$p_t = \frac{1}{1 + \exp -(a + bt + u_t)}$$

Programs like EViews can handle inherently non-linear problems, so if we wanted to estimate a logistic with a saturation level so that $p_t = N_t/K$, where $N_t$ is the number of mobile phone owners and $K$ is the saturation level we could estimate

$$N_t = \frac{K}{1 + \exp -(a + bt)} + \varepsilon_t$$

directly by non-linear least squares. Notice the assumption about the errors is different. In the previous case the error was additive in the logit, here it is additive in the number. We would enter this in Eviews as

$$N = C(1)/(1 + \exp(C(2) + C(3) * @trend)).$$

@trend in EViews provides a trend. C(1) would be an estimate of $K$, C(2) of a and C(3) of b. In practice, unless the market is very close to saturation it is difficult to estimate $K$ precisely.

For non-linear models, the program uses an interative method to find the minimum of the sum of squared residuals or the maximum of the likelihood function. The function may not be well behaved and there may be multiple maxima or minima. See the applied exercise for an example.

### 6.4. Heteroskedasticity.

Suppose we estimate the first stage linear regression (6.1), then in heteroskedasticity tests we run second stage regressions using the squared residuals:

$$\widehat{u}_t^2 = \alpha + b'z_t + v_t$$

the null hypothesis is that the expected value of the squared residuals is a constant $\alpha$, so $b = 0$, and this can be tested with an F test. On the alternative hypothesis, the variance, squared residuals, change with $z_t$. There are lots of ways that the variance could change, thus lots of possible candidates for $z_t$. It could be $x_t$, the regressors; it could be the squares and cross-products of the regressors, often called the White test; it could be the squared fitted values, the RESET version; it could be lagged squared residuals, testing for ARCH (autoregessive conditional heteroskedasticity); etc.

### 6.5. Normality

If the residuals are normal then their coefficient of skewness (third moment) should be zero and ceoffiecient of kurtosis (fourth moment) three. This is tested by the Jarque-Bera test

$$T\left\{\frac{\mu_3^2}{6\mu_2^3} + \frac{1}{24}\left(\frac{\mu_4}{\mu_2^2} - 3\right)^2\right\} \sim \chi^2(2)$$

where $\mu_j = \sum_{t=1}^{T} \widehat{u}_t^j/T$:

## 7. Univariate Stochastic Processes

Suppose we have a series of observations on some economic variable, $y_t, t = 1, 2, ..., T$, which may already have been transformed, e.g. the logarithm of GDP.

It is useful to regard each $y_t$ as a random variable with a density function, $f_t(y_t)$ and we observe one realisation from the distribution for that period. A family of random variables indexed by time is called a stochastic process, an observed sample is called a realisation of the stochastic process. A stochastic process is said to be 'strongly stationary' if its distribution is constant through time, i.e. $f_t(y_t) = f(y_t)$. It is first order stationary if it has a constant mean. It is second order, or weakly or covariance stationary if also has constant variances and constant covariances between $y_t$ and $y_{t-i}$, i.e. the autocovariances (covariances with itself in previous periods) are only a function of $i$ (the distance apart of the observations) not $t$, the time they are observed. These autocovariances summarise the dependence between the observations and they are often represented by the autocorrelation function or correlogram, the vector (graph against i) of the autocorrelations $r_i = Cov(y_t, y_{t-i})/Var(y_t)$. If the series is stationary, the correlogram converges to zero quickly.

The order of integration is the number of times a series must be differenced to make it stationary (after perhaps removing deterministic elements like a linear trend). So a series, $y_t$, is said to be Integrated of order zero, I(0), if $y_t$ is stationary; integrated of order one, I(1), if $\Delta y_t = y_t - y_{t-1}$ is stationary; integrated of order two, I(2), if

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

is stationary. Notice that $\Delta^2 y_t \neq \Delta_2 y_t = y_t - y_{t-2}$.

In examining dynamics it will be useful to use the Lag Operator, L, sometimes known as the backward shift operator B.

$$\begin{aligned} Ly_t &= y_{t-1}; \quad L^2 y_t = y_{t-2}; \quad etc \\ \Delta y_t &= (1-L)y_t, \ \Delta^2 y_t = (1-L)^2 y_t. \end{aligned}$$

### 7.1. White noise processes

A stochastic process is said to be White Noise if

$$\begin{aligned} E(\varepsilon_t) &= 0; \\ E(\varepsilon_t^2) &= \sigma^2; \\ E(\varepsilon_t \varepsilon_{t-i}) &= 0, \ i \neq 0 \end{aligned}$$

We will use $\varepsilon_t$ below to denote white noise processes.

### 7.2. Autoregressive processes

A first order (one lag) autoregressive process (AR1) takes the form:

$$\begin{aligned} y_t &= \rho y_{t-1} + \varepsilon_t, \\ y_t(1-\rho L) &= \varepsilon_t, \end{aligned}$$

with $E(y_t) = 0$, and is stationary if $\mid \rho \mid < 1$. If it is stationary, then by repeated substitution, we get the sum of a geometric progression:

$$y_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3}.... \tag{7.1}$$

$$y_t = (1-\rho L)^{-1}\varepsilon_t, \tag{7.2}$$

the variance of $y_t$ is $E(y_t^2) = \sigma^2/(1-\rho^2)$ and the correlations between $y_t$ and $y_{t-i}$, $r_i = \rho^i$, so decline exponentially. A constant can be included $y_t = \alpha + \rho y_{t-1} + \varepsilon_t$, then $E(y_t) = \alpha/(1-\rho)$. If the process is stationary, the parameters of the AR model can be estimated consistently by Least Squares, though the estimates will not be unbiased ($y_{t-1}$ is uncorrelated with $\varepsilon_t$ but not independent since it is correlated with $\varepsilon_{t-1}$); the estimate of $\rho$ will be biased downwards.

A p th order autoregression (ARp) takes the form:

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + ... + \rho_p y_{t-p} + \varepsilon_t$$

$$y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} - ... - \rho_p y_{t-p} = \varepsilon_t$$

$$(1 - \rho_1 L - \rho_2 L^2 - ... - \rho_p L^p) y_t = \varepsilon_t.$$

The last expression is a p th order polynomial in the lag operator, which we can write $A^p(L)$. $y_t$ is stationary if all the roots (solutions), $z_i$, of $1 - \rho_1 z - \rho_2 z^2 - ... - \rho_p z^p = 0$ lie outside the unit circle (are greater than one in absolute value). If a root lies on the unit circle, some $z_i = 1$, the process is said to exhibit a unit root. The condition is sometimes expressed in terms of the inverse roots, which must lie inside the unit circle. Usually we just check that $\sum \rho_i < 1$ for stationarity.

Consider the case of an AR1 process

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

For stability, the solution to $(1 - \rho z) = 0$, must be greater than unity in absolute value, since this implies $z = 1/\rho$ this requires $-1 < \rho < 1$. For an $AR2$ the real parts of solution to the two solutions to the quadratic $(1 - \rho_1 z - \rho_2 z^2)$ must be greater than unity.

### 7.3. Random Walks

If $\rho = 1$, there is said to be a unit root and the AR1 becomes a random walk:

$$y_t = y_{t-1} + \varepsilon_t;$$

or $\Delta y_t = \varepsilon_t$. Substituting back

$$y_t = \varepsilon_t + \varepsilon_{t-1} + ...\varepsilon_1 + y_0;$$

so shocks have permanent effects. A random walk with drift is of the form: $\Delta y_t = \alpha + \varepsilon_t$.

In both these cases, $\Delta y_t$ is stationary, I(0), but $y_t$ is non-stationary, I(1). If there is no drift the expected value of $y_t$ will be constant at zero, if $y_0 = 0$, but the variance will increase with t. If there is a drift term the expected value of $y_t$, as well as the variance, will increase with t. Random walks appear very often in economics, e.g. the efficient market hypothesis implies that, to a first approximation, asset prices should be random walks.

### 7.4. Moving Average processes

A first order moving average process (MA1) takes the form

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

a q th order moving average:

$$y_t = \varepsilon_t + \mu_1\varepsilon_{t-1} + \mu_2\varepsilon_{t-2} + ... + \mu_q\varepsilon_{t-q};$$

$$y_t = (1 + \mu_1 L + \mu_2 L^2 + ... + \mu_q L^q)\varepsilon_t = B^q(L)\varepsilon_t.$$

$Cov(y_t, y_{t-i}) = 0$, for $i > q$. A finite order moving average process is always stationary. Any stationary process can be represented by a (possibly infinite) moving average process. Notice that the AR1 is written as an infinite MA process in (7.1). The parameters of the MA model cannot be estimated by OLS, but maximum likelihood estimators are available. If a MA process is invertible we can write it as an AR, i.e. $B^q(L)^{-1}y_t = \varepsilon_t$. Notice that if we take a white noise process $y_t = \varepsilon_t$ and difference it we get

$$\Delta y_t = \varepsilon_t - \varepsilon_{t-1}$$

a moving average process with a unit coefficient.

### 7.5. Autoregressive Integrated Moving Average (ARIMA) processes

Combining the AR and MA processes, gives the ARMA process. The first order ARMA(1,1) with intercept is

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t + \mu\varepsilon_{t-1}$$

In practice, the data are differenced enough times, say d, to make them stationary and then modelled as an ARMA process of order p and q. This gives the Autoregressive Integrated Moving Average, ARIMA(p,d,q) process, which can be written using the lag polynomials above as:

$$A^p(L)\Delta^d y_t = \alpha + B^q(L)\varepsilon_t.$$

For instance, the ARIMA(1,1,1) process is

$$\Delta y_t = \alpha + \rho\Delta y_{t-1} + \varepsilon_t + \mu\varepsilon_{t-1}$$

ARIMA models often describe the univariate dynamics of a single economic time-series quite well and are widely used for forecasting.

### 7.6. Trend and difference stationary processes

Most economic time-series, e.g. log GDP, are non-stationary, trended. The trend can be generated in two ways. First, the traditional assumption was that the series could be regarded as stationary once a deterministic trend was removed. For instance:

$$y_t = \alpha + \rho y_{t-1} + \gamma t + \varepsilon_t \tag{7.3}$$

with $\mid \rho \mid < 1$ is a trend stationary process. The effects of the shocks $\varepsilon_t$ are transitory and die away through time, since $\varepsilon_{t-i}$ is multiplied by $\rho^i$ when you substitute back, see (7.1) above. If the variables are in logs, the long run growth rate is $g = \gamma/(1-\rho)$. Second the series could be regarded as a random walk with drift, difference stationary:

$$\begin{aligned} \Delta y_t &= \alpha + \varepsilon_t \\ y_t &= \alpha + y_{t-1} + \varepsilon_t \end{aligned}$$

The long run growth rate is $\alpha$.

We want to test the null of a difference stationary process (one with a unit root) against the alternative of a trend stationary process. Substiture $\gamma = g(1-\rho)$

then substract $y_{t-1}$ from both sides, so we can write the trend stationary process as:

$$\begin{aligned} \Delta y_t &= \alpha + (\rho - 1)(y_{t-1} - gt) + \varepsilon_t \quad (7.4) \\ \Delta y_t &= \alpha + \beta(y_{t-1} - gt) + \varepsilon_t \quad (7.5) \end{aligned}$$

where $\beta = \rho - 1$. If $\rho = 1$ or equivalently $\beta = 0$, we get the random walk with drift:with growth rate $\alpha$. Substituting back we get

$$\begin{aligned} y_t &= \alpha + (\alpha + y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ y_t &= y_{t-2} + 2\alpha + \varepsilon_t + \varepsilon_{t-1} \end{aligned}$$

continuing the process to period zero, we get:

$$y_t = y_0 + \alpha t + \sum_{i=0}^{t-1} \varepsilon_{t-i}.$$

In this case, the difference stationary process, the efects of the shocks are permanent or persistent, they last for ever, and the series is determined by an initial value, $y_0$, a deterministic trend $\alpha t$, and a 'stochastic trend', $\sum_{i=0}^{t-1} \varepsilon_{t-i}$, the sum of past errors.

If we had not restricted (7.4) so that the trend term dropped out when $\beta = 0$, there would be a quadratic trend in $y_t$. Show this by substituting back in

$$\begin{aligned} y_t &= \alpha + y_{t-1} + \gamma t + \varepsilon_t \\ y_t &= \alpha + (\alpha + y_{t-2} + \gamma(t-1) + \varepsilon_{t-1}) + \gamma t + \varepsilon_t \quad (7.6) \end{aligned}$$

etc.

### 7.7. Testing for unit roots

Choosing between the trend stationary and difference stationary model is a matter of determining whether $\beta = 0$ or equivalently $\rho = 1$; whether there is a 'unit root' in $y_t$. To do this we can estimate (7.4) by running a regression of $\Delta y_t$ on a constant, $y_{t-1}$ and a linear trend; estimate $\widehat{\beta}$ the coefficient on the lagged level of $y_{t-1}$; construct the 't statistic' $\tau_\beta = \widehat{\beta}/SE(\widehat{\beta})$ to test $H_0 : \beta = 0$; against $H_1 : \beta < 0$. If we do not reject the null we conclude that there is a unit root in

$y_t$, it is integrated of order one, I(1), stationary after being differenced once. If we reject the null we conclude that $y_t$ is trend stationary I(0). This is a one-sided test and if $\widehat{\beta} > 0$, we do not reject the null of a unit root. The test statistic $\tau_\beta$ does not have a standard t distribution, but a Dickey Fuller distribution and the critical value is -3.5 at the 5% significance level, when a trend is included. This is because under $H_0$ the regressor, $y_{t-1}$ is non-stationary. If there is no trend included in the regression the 5% critical value is -2.9. Most programs will give you these critical values or the relevant p values.

To get good estimates of (7.4) we require that $\varepsilon_t$ is white noise. Often this will not be the case and the error will be serially correlated. To remove this serial correlation, lags of the dependent variable are added to give the 'Augmented Dickey Fuller' (ADF) regression:

$$\Delta y_t = \alpha + \beta y_{t-1} + \gamma t + \sum_{i=1}^{p} \delta_i \Delta y_{t_-i} + \varepsilon_t$$

where p is chosen to try to make the residual white noise. Show that this is a reparameterisation of a AR(p+1) with trend. Again the procedure is to use the t ratio on $\beta$ with the non standard critical values to test the null hypothesis $\beta = 0$ against the alternative $\beta < 0$.

To test for $I(2)$ versus $I(1)$ you just take a further difference:

$$\Delta^2 y_t = \alpha + \beta \Delta y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta^2 y_{t_-i} + \varepsilon_t$$

if it was thought that there might be a trend in the change (not common for economic series) it could be included also. Again $H_0 : \beta = 0$; against $H_1 : \beta < 0$.

There are a range of other procedures for determining whether there is a unit root. They differ, for instance, in how they correct for serial correlation (in a parametric way like the ADF where you allow for lags or in a non-parametric way like Phillips Peron where you allow for window size); whether they include other variables; whether they use the null of a unit root like the ADF or the null of stationarity, like KPSS; whether they use GLS detrending; and whether they use both forward and backward regressions. EViews gives you a lot of choices.

Most of these tests have low power, it is very difficult to distinguish $\rho = 1$ from a stationary alternative in which $\rho$ is close to unity. The power of the tests depends on the span of the data not the number of observations. For instance UK unemployment rates 1945-1985 appear I(1), UK unemployment rates 1855-1985

appear I(0). The tests are also sensitive to step changes, an I(0) process with a single change in level will appear I(1), as it should since the shock (change in level) is permanent. The order of integration is a univariate statistical summary of how the time series moves over the sample, it is not an inherent structural property of the series. Whether you treat a variable as $I(0)$ or $I(1)$ depends on the purpose of the exercise, for estimation it is often safer to treat it as $I(1)$.

## 8. Dynamic Linear Regression, ARDL models

When we look at the relationship between variables, there are usually lags between the change in one variable and the effect on another. The distributed lag of order q, DL(q) regression model takes the form:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + ... + \beta_q x_{t-q} + u_t$$

notice that it is similar to a moving average, except that here the shocks are observed, $x_t$, rather than being unobserved. We can combine the distributed lag with an autoregressive component to give the ARDL(p,q) process:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + ... + \beta_q x_{t-q} + u_t \tag{8.1}$$

where $u_t$ is usually a white noise error, though it could be moving average. If the error is white noise, the parameters can be estimated consistently by OLS, though the estimates are not unbiased. $y_t$ is stationary, conditional on $x_t$, the process is stable, if all the roots (solutions), $z_i$, of the characteristic equation

$$1 - \alpha_1 z - \alpha_2 z^2 - ... - \alpha_p z^p = 0 \tag{8.2}$$

lie outside the unit circle (are greater than one in absolute value). We usually check that $\sum \alpha_i < 1$. In this case, if $x_t$ is constant, say at $x$, then $y_t$ will converge to a constant, say $y$, and the long run relation between them will be:

$$y = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i} + \frac{\sum_{i=0}^{q} \beta_i}{1 - \sum_{i=1}^{p} \alpha_i} x = \theta_0 + \theta_x x.$$

This can be obtained by setting $y_{t-i} = y$ and $x_{t-i} = x$ for all i. This long-run solution is usually interpreted as the long-run equilibrium or target value for $y_t$ and can be calculated from the estimated regression coefficients. Standard errors for the long-run coefficients can be calculated by the delta method, which is available in most programmes.

This procedure is appropriate in quite a wide variety of circumstances including if all the variables are I(0) and $x_t$ is exogenous; or if all the variables are I(1), there is a single cointegrating relationship and $x_t$ is exogenous; or if there are mixtures of I(0) and cointegrating I(1) variables such that $u_t$ is white noise. See below for more details.

### 8.1. ARDL(1,1)

ARDL models or dynamic linear regressions are widely used to examine the relationship between economic variables. We will use the ARDL(1,1) for illustration, this is:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t. \tag{8.3}$$

It is stable if $-1 < \alpha_1 < 1$, and then has a long run solution:

$$y_t^* = \frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0 + \beta_1}{1-\alpha_1} x_t = \theta_0 + \theta_x x_t.$$

Where $y_t^*$ is the target or long run equilibrium value for $y_t$ to which it would tend in the absence of further shocks to $x_t$ and $u_t$. There are a number of other useful ways of rewriting (reparameterizing) (8.3).

#### 8.1.1. Bewley transform

A trick to get the standard errors is to write (8.3) as

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

and divide through by $(1-\alpha_1)$ to give

$$y_t = \frac{\alpha_0}{(1-\alpha_1)} - \frac{\alpha_1}{(1-\alpha_1)}\Delta y_t + \frac{(\beta_0+\beta_1)}{(1-\alpha_1)}x_t - \frac{\beta_1}{(1-\alpha_1)}\Delta x_t + \frac{u_t}{(1-\alpha_1)}$$

while $\Delta y_t$ is clearly correlated with $u_t$ the equation can be estimated by the method of instrumental variables (see below) using one, $y_{t-1}$, $x_t$ and $x_{t-1}$ as instruments. The coefficient of $x_t$ will be the long-run coefficient and its standard error will be the same as got by the delta method. :

### 8.1.2. Error Correction Models

Write (8.3) as

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

$$\Delta y_t = a_0 + b_0\Delta x_t + a_1 y_{t-1} + b_1 x_{t-1} + u_t \tag{8.4}$$

where $a_0 = \alpha_0; b_0 = \beta_0; a_1 = (\alpha_1 - 1); b_1 = \beta_0 + \beta_1$; or in terms of adjustment to a long-run target:

$$\Delta y_t = \lambda_1 \Delta y_t^* + \lambda_2(y_{t-1}^* - y_{t-1}) + u_t$$

where the long-run target or equilibrium (as calculated above) is

$$y_t^* = \theta_0 + \theta_x x_t,$$

and the $\lambda_i$ are adjustment coeficients which measure how $y$ adjusts to changes in the target and deviations from the target. Notice $a_0 = \lambda_2\theta_0; a_1 = -\lambda_2; b_0 = \lambda_1\theta_x; b_1 = \lambda_2\theta_x$. This form is usually known as an 'Error (or equilibrium) Correction Model' ECM. The dependent variable changes in response to changes in the target and to the error, the deviation of the actual from the equilibrium in the previous period: $(y_{t-1}^* - y_{t-1})$.

An alternative parameterization, which unlike the ECM nests the partial adjustment model is:

$$\Delta y_t = \alpha_0 + (\alpha_1 - 1)y_{t-1} + (\beta_0 + \beta_1)x_t - \beta_1\Delta x_t + u_t.$$

When you **reparameterize** a model, as we did above, you estimate exactly the same number of parameters (4 in this case), just written in different ways.You will get identical estimates of say, the long-run coefficient, whether you estimate it as an ARDL, ECM or by a non-linear procedure. The statistical properties of the model do not change, the estimated residuals, standard error of the regression and the maximised log-likelihood are identical between the different versions. $R^2$ will change, because the proportion of variation explained is measured in terms of a different dependent variable, $\Delta y_t$ in the ECM rather than $y_t$ in the ARDL. Any RESET tests that use fitted values of the dependent variable will also change. Use the misspecification tests which use the fitted values of $\Delta y_t$.

### 8.2. Restrictions

When you **restrict** a model, you reduce the number of parameters estimated and such restrictions are testable. The ARDL(1,1) nests a number of interesting restricted special cases, including:

(a) Static: $\alpha_1 = 0; \beta_1 = 0$.
(b) First difference: $\alpha_1 = 1; \beta_1 = -\beta_0$
(c) Partial Adjustment Model:$\beta_1 = 0$
(d) First order disturbance serial correlation: $\beta_1 = -\beta_0\alpha_1$
(e) Unit long-run coefficient: $\beta_1 + \beta_0 + \alpha_1 = 1$
(f) Random Walk with drift: $\alpha_1 = 1; \beta_1 = \beta_0 = 0$.

A useful procedure in many circumstances is to start with a general model, e.g. the ARDL(1,1) and test down to specific restricted cases. This general to specific procedure has the advantage that any tests on the general model are valid. Whereas if you start from the restricted model, the tests will not be valid if the model is misspecified.

Case (d) is got by assuming that the model is:

$$y_t = \alpha + \beta x_t + v_t; \; v_t = \rho v_{t-1} + \varepsilon_t$$

where $\varepsilon_t$ is white noise, this can be written:

$$y_t = \alpha + \beta x_t + \rho v_{t-1} + \varepsilon_t$$

noting that

$$v_t = y_t - \alpha - \beta x_t; \text{ and } \; v_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}$$
$$y_t = \alpha + \beta x_t + \rho(y_{t-1} - \alpha - \beta x_{t-1}) + \varepsilon_t$$
$$y_t = \alpha(1-\rho) + \beta x_t + \rho y_{t-1} - \beta\rho x_{t-1} + \varepsilon_t$$

which is of the same form as (8.3) with the restriction that the coefficient of $x_{t-1}$ equals the negative of the product of the coefficients of $x_t$ and $y_{t-1}$, i.e. $\beta_1 = -\beta_0\alpha_1$ in terms of the parameters of the unrestricted model. This is sometimes called the common factor model, since it can be written $(1-\rho L)y_t = (1-\rho L)(\alpha + \beta x_t) + \varepsilon_t$, both sides of the static model are multiplied by the common factor $(1-\rho L)$. The restricted model (with AR1 errors) is not linear in the parameters and is estimated by Generalised Least Squares or Maximum Likelihood.

In case (e) the restricted model can be written:

$$\Delta y_t = a_0 + b_0\Delta x_t + a_1(y_{t-1} - x_{t-1}) + e_t.$$

and the restriction is equivalent to assuming $b_1 = -a_1$ in (8.4).

### 8.3. ARDL(1,1,1)

The structure generalises to more explanatory variables, e.g. the ARDL(1,1,1) model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \gamma_0 z_t + \gamma_1 z_{t-1} + u_t. \tag{8.5}$$

has a long run solution:

$$y = \frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0+\beta_1}{1-\alpha_1}x + \frac{\gamma_0+\gamma_1}{1-\alpha_1}z = \theta_0 + \theta_x x + \theta_z z.$$

Notice that our error correction adjustment process

$$\Delta y_t = \lambda_1 \Delta y_t^* + \lambda_2(y_{t-1}^* - y_{t-1}) + u_t$$
$$y_t^* = \theta_0 + \theta_x x_t + \theta_z z_t,$$

now imposes restrictions. In the case of one exogenous variable, there were four ARDL parameters $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ and four theoretical parameters $(\lambda_1, \lambda_2, \theta_0, \theta_x)$ so no restrictions. In the case of two exogenous variables there are six ARDL parameters, $(\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma_0, \gamma_1)$ but only five theoretical parameters $(\lambda_1, \lambda_2, \theta_0, \theta_x, \theta_z)$. What is the restriction?

### 8.4. Adaptive Expectations

Define the expected value of $x_{t+1}$ conditional on information available at time t, as:

$$E(x_{t+1} \mid I_t) = x_t^e;$$

agents determine their actions according to:

$$y_t = \beta x_t^e + u_t \tag{8.6}$$

and determine their expectations according to:

$$x_t^e - x_{t-1}^e = \phi(x_t - x_{t-1}^e)$$

they adjust their forecast proportional to the forecast error they made in the previous period (note $x_{t-1}^e$ is the forecast of $x_t$ made in the previous period). This can be written:

$$\begin{aligned} x_t^e &= \phi x_t + (1-\phi)x_{t-1}^e \\ (1-(1-\phi)L)x_t^e &= \phi x_t \\ x_t^e &= \frac{\phi x_t}{(1-(1-\phi)L)} = \phi \sum_{i=0}^{\infty} (1-\phi)^i x_{t-i} \end{aligned}$$

substituting this exponentially weighted moving average of past $x_t$ in (8.6) gives:

$$y_t = \beta \left( \frac{\phi x_t}{(1-(1-\phi)L)} \right) + u_t \tag{8.7}$$

premultiply by $(1-(1-\phi)L)$ to give

$$(1-(1-\phi)L)y_t = \beta\phi x_t + (1-(1-\phi)L)u_t$$

$$y_t = \beta\phi x_t + (1-\phi)y_{t-1} + u_t - (1-\phi)u_{t-1}$$

an ARDL(1,0) with a MA1 error, with a restriction that the AR and MA coefficients should be equal and of opposite sign.

This type of transformation (known as the Koyck transform) can be used to get rid of a variety of exponentially weighted infinite distributed lags.

# 9. Cointegration

## 9.1. Introduction

Suppose $y_t$ and $x_t$ are I(1) then in general any linear combination of them will also be I(1). If there is a linear combination that is I(0), they are said to cointegrate. If they cointegrate, they have a common stochastic trend which is cancelled out by the linear combination; and this linear combination is called the cointegrating vector, which is often interpreted as an equilibrium relationship.

Suppose we have data on $s_t$, $p_t$, $p_t^*$, the logarithms of the spot exchange rate (domestic currency per unit foreign), domestic and foreign price indexes and that each of these are I(1). Purchasing Power Parity says that the real exchange rate $e_t = s_t - p_t + p_t^*$ should be stationary, i.e. $e_t = e + u_t$ where $e$ is the equilibrium real exchange rate and $u_t$ is a stationary (not necessarily white noise) error. The cointegrating vector is then $(1, -1, 1)$. It is quite common in economics to get ratios of non-stationary variables being approximately stationary. These 'great ratios' include the real exchange rate, the savings ratio, the velocity of circulation of money, the capital-output ratio, the share of wages in output, the profit rate, etc. In each case a linear combination of the logarithm of the variables with cointegrating vectors of plus and minus ones should be stationary and this can be tested using the unit root tests described above.

The coefficient does not need to be unity. If

$$y_t = \alpha + \beta x_t + u_t \tag{9.1}$$

and $u_t$ is stationary, the cointegrating vector is $(1, -\beta)$ since $(y_t - \beta x_t = \alpha + u_t)$ is I(0).

If $y_t$ and $x_t$ are I(1) and do not cointegrate, say they are independent unrelated random walks, the error in (9.1) will be I(1) and this will be a 'spurious' regression. As $T \to \infty$, the $R^2$ of this regression will go to unity and the t ratio for $\widehat{\beta}$ will go to a non-zero random variable. Thus even if there is no relationship, the regression would indicate a close relationship. Therefore it is important to test for cointegration. A similar issue arises in the ARDL(1,1). Write it in the ECM form:

$$\Delta y_t = a_0 + b_0 \Delta x_t + a_1 y_{t-1} + b_1 x_{t-1} + u_t.$$

This equation does not seem to balance, the left hand side $\Delta y_t$ is I(0) and there are two I(1) terms $y_{t-1}$ and $x_{t-1}$ on the right hand side. It only balances if a linear combination of the I(1) terms is I(0), that is if $y_t$ and $x_t$ cointegrate so that $y_t - \theta_x x_t$ is I(0) with cointegrating vector $(1, -\theta_x)$, in:

$$\Delta y_t = a_0 + b_0 \Delta x_t + \lambda(y_{t-1} - \theta_x x_{t-1}) + u_t \tag{9.2}$$

Notice that if they cointegrate $\lambda$ must be non-zero and negative (this is the feedback that keeps $y_t$ and $x_t$ from diverging. We can test for this, though the critical values are non standard, see below. Notice we are free to normalise the cointegrating vector, since $a_1 y_{t-1} + b_1 x_{t-1}$ is I(0), we could also have called the cointegrating vector $(a_1, b_1) = (\lambda, -\lambda\theta_x)$.

With only two I(1) variables there can only be a single cointegrating vector, but with more than two variables there can be more than one cointegrating vector and any linear combination of these cointegrating vectors will also be a cointegrating vector. Suppose that we have data on domestic and foreign interest rates and inflation $(r_t, r_t^*, \Delta p_t, \Delta p_t^*)$ and all are I(1) (this implies that $p_t$ is I(2)). If real interest rates ($r_t - \Delta p_t$ and $r_t^* - \Delta p_t^*$) are I(0) with cointegrating vectors $(1, 0, -1, 0)$ and $(0, 1, 0, -1)$; then the real interest rate differential $(r_t - \Delta p_t) - (r_t^* - \Delta p_t^*)$ would also be I(0), with cointegrating vector $(1, -1, -1, 1)$.

## 9.2. Ways to test for cointegration.

### 9.2.1. Known cointegrating vector

If the cointegrating vector is known a priori (as with the real exchange rate or real interest rate examples above) we can form the hypothesised I(0) linear combination (the log of the real exchange rate or the real interest rates) and use an ADF test to determine whether it is in fact I(0).

### 9.2.2. Single unknown cointegrating vector

There are three procedures here.

(a) Those that can be used for multiple unknown cointegrating vectors discussed below.

(b) Estimating an ARDL model and testing for the existence of a long-run relationship, i.e. test the null hypothesis that the levels $x_{t-1}$ and $y_{t-1}$ should not appear in the equation or equivalently that $\lambda = 0$ in (9.2) above, using the appropriate (non-standard) critical values, which are given in Pesaran, Shin and R.J. Smith, Journal of Applied Econometrics, 2001, p289-326, Bounds Testing Approaches to the Analysis of Level Relationships..

(c) Running the levels equation (9.1) above and testing whether the residuals are I(1), using an ADF test and the appropriate critical values, which are different from those for an ADF on an ordinary variable. This is the original Engle-Granger procedure. Although the estimates of (9.1) are 'superconsistent' (converge to their true values at rate $T$ rather than $\sqrt{T}$), (9.1) is clearly misspecified because it omits the dynamics, thus the estimates can be badly biased in small samples. In addition doing a unit root test on the residuals, imposes very strong restrictions on the short-run dynamics, which may not be appropriate. Thus the original Engle-Granger procedure is not recommended in most cases. If you know that one variable is exogenous use (b), if you do not know which is the exogenous variable start with (a) and test for exogeneity.

### 9.2.3. Multiple unknown cointegrating vectors

Again there are a variety of procedures, but the most commonly used is the Johansen procedure discussed below. This procedure operates in the context of a VAR, which we consider first.

## 9.3. Vector Autoregressions and cointegration

### 9.3.1. VARs

The generalisation of an AR2 to a vector is the VAR2:

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t$$

where $y_t$ is now a $m \times 1$ vector, $A_0$ a $m \times 1$ vector, $A_1$ and $A_2$ are $m \times m$ matrices and $\varepsilon_t \sim N(0, \Sigma)$, where $\Sigma$ is a $m \times m$ matrix with elements $\sigma_{ij}$.

For $m = 2, ...y_t = (y_{1t}, y_{2t})'$ the VAR is:

$$\begin{aligned} y_{1t} &= a_1^0 + a_{11}^1 y_{1t-1} + a_{12}^1 y_{2t-1} + a_{11}^2 y_{1t-2} + a_{12}^2 y_{2t-2} + \varepsilon_{1t}, \\ y_{2t} &= a_2^0 + a_{21}^1 y_{1t-1} + a_{22}^1 y_{2t-1} + a_{21}^2 y_{1t-2} + a_{22}^2 y_{2t-2} + \varepsilon_{2t}. \end{aligned}$$

Each equation of the VAR can be estimated consistently by OLS and the covariance matrix $\Sigma$ can be estimated from the OLS residuals,

$$\widehat{\sigma}_{ij} = \frac{1}{T} \sum_{t=1}^{T} \widehat{\varepsilon}_{it} \widehat{\varepsilon}_{jt}$$

where $\widehat{\sigma}_{11}$ is the estimated variance of $\varepsilon_{1t}$, $\widehat{\sigma}_{12}$ the estimated covariance of $\varepsilon_{1t}$ and $\varepsilon_{2t}$.

A variable $y_{2t}$ is said to Granger cause $y_{1t}$ if knowing current values of $y_{2t}$ helps you to predict future values of $y_{1t}$ equivalently, current $y_{1t}$ is explained by past $y_{2t}$. In this case, $y_{2t}$ is Granger causal with respect to $y_{1t}$ if either $a_{12}^1$ or $a_{12}^2$ are non zero. You can test that they are both zero with a standard F test of linear restrictions. The restricted model just excludes $y_{2,t-1}$ and $y_{2,t-2}$ from the equation for $y_{1t}$. Granger causality is rarely the same as economic causality, particularly because expectations cause consequences to precede their cause: weather forecasts Granger Cause the weather.

More lags can be included and you can decide the appropriate lag length by Likelihood Ratio tests or model selection criteria like the AIC or SBC. Make sure that you use the same sample for the restricted and unrestricted model; i.e. do not use the extra observation that becomes available when you shorten the lag length. If the lag length is $p$, each equation of the VAR with intercept has $1 + mp$ parameters. This can get large, 4 lags in a 4 variable VAR gives 17 parameters in each equation. Be careful about degrees of freedom.

A pth order VAR

$$y_t = A_0 + \sum_{i=1}^{p} A_i y_{t-i} + \varepsilon_t$$

is stationary if all the roots of the determinantal equation $\mid I - A_1 z - A_2 z^2 - ... - A_p z^p \mid = 0$ lie outside the unit circle. When you estimate a VAR, EViews will give you a graph of the inverse roots, which should lie inside the unit circle for the variables to all be stationary.

We can reparameterise the VAR2:

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t$$

as:

$$\begin{aligned} y_t - y_{t-1} &= A_0 - (I - A_1 - A_2) y_{t-1} - A_2 (y_{t-1} - y_{t-2}) + \varepsilon_t \\ \Delta y_t &= A_0 - \Pi y_{t-1} + \Gamma \Delta y_{t-1} + \varepsilon_t \end{aligned}$$

and the VARp as:

$$\Delta y_t = A_0 - \Pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t.$$

Notice that this is the vector equivalent of the Augmented Dickey Fuller regression that we used above for testing for unit roots. Express the $\Gamma_i$ in terms of the $A_i$.

### 9.3.2. Cointegration in VARs

If all the variables, the m elements of $y_t$, are I(0), $\Pi$ is a full rank matrix. If all the variables are I(1) and not cointegrated, $\Pi = 0$, and a VAR in first differences is appropriate. If the variables are I(1) and cointegrated, with $r$ cointegrating vectors, then there are r cointegrating relations, combinations of $y_t$ that are I(0),

$$z_t = \beta' y_t$$

where $z_t$ is a $r \times 1$ vector and $\beta'$ is a $r \times m$ matrix. Then we can write the model as:

$$\Delta y_t = A_0 - \alpha z_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t,$$

in which the I(0) dependent variable is only explained by I(0) variables and $\alpha$ is a $m \times r$ matrix of 'adjustment coefficients' which measure how the deviations from equilibrium (the $r$ I(0) variables $z_{t-1}$) feed back on the changes. This can also be written:

$$\Delta y_t = A_0 - \alpha \beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t,$$

so $\Pi = \alpha\beta'$ has rank $r < m$ if there are r cointegrating vectors. If there are $r < m$ cointegrating vectors, then $y_t$ will also be determined by $m - r$ stochastic trends,

and will have $m - r$ roots on the unit circle and $m$ roots outside the unit circle. If there is cointegration, some of the $\alpha$ must be non-zero, there must be some feedback on the $y_t$ to keep them from diverging, i.e. there must be some Granger causality in the system.

If there are $r$ cointegrating vectors and $\Pi$ has rank $r$, it will have $r$ non-zero eigenvalues and Johansen provided a way of estimating the eigenvalues and two tests for determining how many of the eigenvalues are different from zero. These allow us to determine $r$, though the two tests may give different answers. The Johansen estimates of the cointegrating vectors $\beta$ are the associated eigenvectors.

There is an 'identification' problem, since the $\alpha$ and $\beta$ are not uniquely determined. We can always choose a non-singular $r \times r$ matrix P such that $(\alpha P)(P^{-1}\beta) = \Pi$ and the new estimates $\alpha^* = (\alpha P)$ and $\beta^* = (P^{-1}\beta)$ would be equivalent, though they might have very different economic interpretations. Put differently, if $z_{t-1} = \beta' y_{t-1}$ are I(0) so are $z^*_{t-1} = P^{-1}\beta' y_{t-1}$, since any linear combination of I(0) variables is I(0). We need to choose the appropriate P matrix to allow us to interpret the estimates. This requires $r^2$ restrictions, $r$ on each cointegrating vector. One of these is provided by normalisation, we set the coefficient of the 'dependent variable' to unity, so if $r = 1$ this is straightforward (though it requires the coefficient set to unity to be non-zero). If there is more than one cointegrating vector it requires prior economic assumptions. The Johansen identification assumption, that the $\beta$ are eigenvectors with unit length and orthogonal, do not allow an economic interpretation. Programs like EViews or Microfit allow you to specify the $r^2$ just identifying restrictions and test any extra 'over-identifying' restrictions.

As we saw above with the Dickey Fuller regression, there is also a problem with the treatment of the deterministic elements. If we have a linear trend in the VAR, and do not restrict the trends, the variables will be determined by $m - r$ quadratic trends. To avoid this (economic variables tend to show linear not quadratic trends), we enter the trends in the cointegrating vectors,

$$\Delta y_t = A_0 - \alpha(\beta' y_{t-1} + ct) + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t,$$

so if an element of $\alpha$ is zero the trend drops out. Most programs give you a choice of how you enter trends and intercepts; unrestricted intercepts and restricted trends, option 4 in Eviews, is a good choice for trended economic data.

### 9.3.3. Example: money demand

Consider a VAR1 in the logarithms of real money and income, which are both I(1) with a linear trend:

$$\begin{aligned} m_t &= a_{10} + a_{11}m_{t-1} + a_{12}y_{t-1} + \gamma_1 t + \varepsilon_{1t} \\ y_t &= a_{20} + a_{21}m_{t-1} + a_{22}y_{t-1} + \gamma_2 t + \varepsilon_{2t} \end{aligned}$$

and $z_t = m_t - \beta y_t$ is I(0). The cointegrating vector is $(1, -\beta)$ and we have normalised the equation by setting the coefficient of $m_t$ to unity. This just identifies the cointegrating vector for r=1. The VECM is:

$$\begin{aligned} \Delta m_t &= a_{10} + (a_{11} - 1)m_{t-1} + a_{12}y_{t-1} + \gamma_1 t + \varepsilon_{1t} \\ \Delta y_t &= a_{20} + a_{21}m_{t-1} + (a_{22} - 1)y_{t-1} + \gamma_2 t + \varepsilon_{2t}, \end{aligned}$$

or

$$\begin{aligned} \Delta m_t &= a_{10} + \pi_{11}m_{t-1} + \pi_{12}y_{t-1} + \gamma_1 t + \varepsilon_{1t} \\ \Delta y_t &= a_{20} + \pi_{21}m_{t-1} + \pi_{22}y_{t-1} + \gamma_2 t + \varepsilon_{2t}. \end{aligned}$$

We can write this

$$\begin{aligned} \Delta m_t &= a_{10} + \pi_{11}(m_{t-1} + \frac{\pi_{12}}{\pi_{11}}y_{t-1}) + \gamma_1 t + \varepsilon_{1t} \\ \Delta y_t &= a_{20} + \pi_{21}(m_{t-1} + \frac{\pi_{22}}{\pi_{21}}y_{t-1}) + \gamma_2 t + \varepsilon_{2t}. \end{aligned}$$

Imposing the cointegration restriction, that the long-run coefficients are the same in both equations,

$$\frac{\pi_{12}}{\pi_{11}} = \frac{\pi_{22}}{\pi_{21}}$$

it becomes:

$$\begin{aligned} \Delta m_t &= a_{10} - \alpha_1(m_{t-1} - \beta y_{t-1}) + \gamma_1 t + u_{1t} \\ \Delta y_t &= a_{20} - \alpha_2(m_{t-1} - \beta y_{t-1}) + \gamma_2 t + u_{2t} \end{aligned}$$

where $-\alpha_1 = \pi_{11}$ etc. Thus

$$\Pi = \begin{bmatrix} -\alpha_1 & +\alpha_1\beta \\ -\alpha_2 & +\alpha_2\beta \end{bmatrix}$$

which is clearly of rank 1, since a multiple of the first column equals the second column. A natural over-identifying restriction to test in this context would be that $\beta = 1$. To restrict the trend we could put it in the cointegrating vector, saving one further parameter:

$$\begin{aligned} \Delta y_t &= a_{10} - \alpha_1(m_{t-1} - \beta y_{t-1} + \gamma t) + u_{1t} \\ \Delta m_t &= a_{10} - \alpha_2(m_{t-1} - \beta y_{t-1} + \gamma t) + u_{2t} \end{aligned}$$

If $y_t$ is weakly exogenous the $\alpha_1 = 0$, which can be tested.

# 10. Endogenous regressors and IV Estimation

## 10.1. Exogeneity

Exogeneity is a difficult concept, Hendry's text Dynamic Econometrics is probably the best available treatment. There are a number of different definitions, which fall into two classes of approach.

The first approach starts with the joint distribution of the random variables, $y_t, x_t$, which can be written as the product of the distribution of $y_t$ conditional on $x_t$ and the marginal distribution of $x_t$ :

$$D_j(y_t, x_t; \theta_j) = D_c(y_t \mid x_t; \theta_c) D_m(x_t; \theta_m) \tag{10.1}$$

$\theta_j$ is a vector of parameters of the joint distribution, $\theta_c$ of the conditional distribution, $\theta_m$ of the marginal. The distribution that we will be interested in is the distribution of $y_t$ conditional on $x_t$ and the parameters that we will be interested in are the parameters of the conditional distribution $\theta_c$ which we will usually denote by $\theta$. **Weak exogeneity** requires that the parameters of interest should be functions only of the parameters of the conditional distribution, $\theta_c = \theta = (\beta, \sigma^2)$, and that the parameters of the conditional and marginal distributions should be 'variation free': there are no restrictions linking them. Essentially this says that we can ignore the information in the marginal distribution of $x$ for the purpose of estimating particular parameters. Notice that exogeneity is not an inherent property of $x$, it is only defined relative to the parameters you want to estimate. $x$ may be exogenous for some parameters and not for others. This is the assumption that we need for efficient inference about the parameters of interest. The main reasons for it failing in economics are simultaneity, where the regressors are jointly determined with the dependent variable (prices and quantities are simultaneously

determined by demand and supply), and measurement errors in the regressors. In both cases we need information about the processes generating the regressors to consistently estimate the parameters of interest. **Strong exogeneity** is weak exogeneity plus Granger Non-causality of $y_t$ with respect to $x_t$, we need this assumption for forecasting. **Super exogeneity** requires that the parameters of the conditional distribution, $\theta_c$, should be invariant to changes in the parameters of the marginal distribution of $x_t$. In this case even if the process generating $x_t$ changes, the parameters of our regression do not change. We need this for policy analysis which usually involves changing right hand side policy variables and essentially this assumption precludes the Lucas Critique. Notice that these three definitions are presented in terms of the distributions of the observables, $y_t$ and $x_t$.

The second approach, very common in the text books, presents the assumptions in terms of the unobservable error or disturbance $u_t$. Notice that our assumption, in terms of the conditional distribution of $y$, $D_c(y \mid X; \theta) \sim N(X\beta, \sigma^2 I)$, is equivalent to an assumption in terms of the unconditional distribution of the disturbance $u \sim N(0, \sigma^2 I)$. In this framework, there are three types of exogeneity assumptions that are made about $X$. Firstly, it may be a set of fixed numbers, **non-stochastic** or deterministic. These phrases are all equivalent ways to describe the fact that $X$ is not a random variable. Apart from trends and seasonals non-stochastic variables are rare in economics. Secondly, it may be **strictly exogenous,** a set of random variables which are distributed independently of the disturbance term. Thirdly, it may be **predetermined** a set of random variables which are uncorrelated with the disturbance term. If $X$ is strictly exogenous, $x_t$ is uncorrelated with the whole sequence of $u_t$, $t = 1, 2, ..., T$. If it is predetermined, it is only uncorrelated with the current value of $u_t$. Typically predetermined variables are lagged (past) values of $y_t$ which are included in the $x_t$.

### 10.2. The Simultaneous Equations Model.

Consider the simple demand and supply model for an agricultural product in structural form as

$$\begin{aligned} q_t^d &= \gamma_{10} + \beta_{12} p_t + \gamma_{11} y_t + u_{1t} \\ q_t^s &= \gamma_{20} + \beta_{22} p_t + \gamma_{22} w_t + u_{2t} \end{aligned}$$

demand is determined by price and income, supply is determined by price and the weather and price adjusts so that demand equals supply $q_t^d = q_t^s = q_t$. This

system simultaneously determines price and quantity in terms of the exogenous variables income and the weather and the errors:

$$\begin{aligned} p_t &= [\beta_{12} - \beta_{22}]^{-1} \{(\gamma_{20} - \gamma_{10}) - \gamma_{11} y_t + \gamma_{22} w_t + (u_{2t} - u_{1t})\} \\ q_t &= [\beta_{12} - \beta_{22}]^{-1} \{(\beta_{12}\gamma_{20} - \beta_{22}\gamma_{10}\} - \beta_{22}\gamma_{11} y_t + \beta_{12}\gamma_{22} w_t + (\beta_{12} u_{2t} - \beta_{22} u_{1t})\} \end{aligned}$$

$$\begin{aligned} p_t &= \pi_{10} + \pi_{11} y_t + \pi_{12} w_t + v_{1t} \\ q_t &= \pi_{20} + \pi_{21} y_t + \pi_{22} w_t + v_{2t} \end{aligned}$$

this is called the reduced form. The standard demand-supply system in economics is normalised in an unusual form, making quantity the dependent variable in both equations. Usually systems are normalised so that each endogenous variable is the dependent variable in each equation. Normalisation, specifies that a coefficient of a dependent variable equals unity.

We can write the system in matrix notation as

$$\begin{bmatrix} 1 & -\beta_{12} \\ 1 & -\beta_{22} \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \gamma_{10} & \gamma_{11} & 0 \\ \gamma_{20} & 0 & \gamma_{22} \end{bmatrix} \begin{bmatrix} 1 \\ y_t \\ w_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{B}\boldsymbol{y}_t &= \boldsymbol{\Gamma}\boldsymbol{x}_t + \boldsymbol{u}_t; \quad \mathbf{E}(\mathbf{u}_t \mathbf{u}_t') = \boldsymbol{\Omega} \\ \boldsymbol{y}_t &= \boldsymbol{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{x}_t + \boldsymbol{B}^{-1}\mathbf{u}_t \\ \boldsymbol{y}_t &= \boldsymbol{\Pi}\boldsymbol{x}_t + \mathbf{v}_t \\ \mathbf{E}(\mathbf{v}_t \mathbf{v}_t') = \boldsymbol{\Sigma} &= \boldsymbol{B}^{-1}\boldsymbol{\Omega}\boldsymbol{B}^{-1\prime} \end{aligned}$$

Where $B$ is a $m \times m$ matrix, $\Gamma$ is a $m \times k$ matrix.

In the demand-supply case $m = 2$, $k = 3$. Notice that if $\mathbf{x}_t = \mathbf{y}_{t-1}$ the reduced form of the system is a VAR. In general OLS estimates of the structural form will be inconsistent since in the demand equation $u_{1t}$ will be correlated with $p_t$ (which is a function of $u_{1t}$ as the reduced form equations show). We can estimate the reduced form by OLS. The identification problem is whether we can recover the $m \times m + m \times k$ $\mathbf{B}$ and $\mathbf{\Gamma}$ coefficients from the $m \times k$ estimates in $\mathbf{\Pi}$. We are obviously $m^2$ coefficients short and the information has to come from somewhere else, e.g. economic theory. For each equation we need $d \geqslant m$ extra pieces of information, one of these will come from normalisation, we set the coefficient of the dependent variable to unity. This is the order condition, a necessary but not

sufficient condition for identification. This condition is written in lots of different but equivalent ways in the literature. One way of expressing it for a particular equation is that the number of excluded exogenous variables (not appearing in that equation) must be greater or equal to the number of included right hand side endogenous variables. The necessary condition is the rank condition. When $d < m$, the equation is said to be underidentified or not identified; when $d = m$ it is said to be exactly identified or just identified; when $d \geqslant m$ it is said to be overidentified. You can have a system with some equations identified and others not identified.

In the demand and supply example, both equations are exactly identified because we have two restrictions in each case, $d = 2, m = 2$. In the demand equation we have $\beta_{11} = 1; \gamma_{12} = 0$. In the supply equation $\beta_{21} = 1; \gamma_{21} = 0$. Estimation can then be done by Two stage Least Squares. First estimate the reduced form and obtain the predicted values for $p_t$ and $q_t$ :

$$\begin{aligned} \widehat{p}_t &= \widehat{\pi}_{10} + \widehat{\pi}_{11} y_t + \widehat{\pi}_{12} w_t \\ \widehat{q}_t &= \widehat{\pi}_{20} + \widehat{\pi}_{21} y_t + \widehat{\pi}_{22} w_t \end{aligned}$$

these are just functions of the exogenous variables and so are not correlated with $u_{1t}$ and $u_{2t}$ and can be used in two second stage regressions estimating the structural equations

$$\begin{aligned} q_t &= \gamma_{10} + \beta_{12} \widehat{p}_t + \gamma_{11} y_t + e_{1t} \\ q_t &= \gamma_{20} + \beta_{22} \widehat{p}_t + \gamma_{22} w_t + e_{2t} \end{aligned}$$

where $e_{1t} = u_{1t} + \beta_{i2} \widehat{v}_{1t}$ neither of which are correlated with $\widehat{p}_t$. Two Stage Least Squares, 2SLS, is an example of Instrumental Variables, IV, discussed below.

Consider another example, the simple Keynesian model of identity and consumption function:

$$\begin{aligned} Y_t &= C_t + I_t, \\ C_t &= \alpha + \beta Y_t + u_t. \end{aligned}$$

Notice that identification is not an issue for the identity, since there are no coefficients to estimate. The (restricted) reduced form is

$$\begin{aligned} Y_t &= \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{u_t}{1-\beta} \\ C_t &= \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} I_t + \frac{u_t}{1-\beta}. \end{aligned}$$

Notice the coefficient of investment in the income equation is the standard Keynesian multiplier. The (unrestricted) reduced form which we can estimate is

$$\begin{aligned} Y_t &= \pi_{10} + \pi_{11} I_t + v_{1t} \\ C_t &= \pi_{20} + \pi_{21} I_t + v_{2t}, \end{aligned}$$

where $v_{1t} = v_{2t}$, etc. Clearly $Y_t$ is correlated with $u_t$ in the consumption function, since as the reduced form shows $u_t$ determines $Y_t$ through consumption. What is the covariance between $Y_t$ and $u_t$? However we could estimate $\beta$ by "indirect least squares", ILS, as the ratio of the two reduced form coefficients of investment, where the lower case letters indicate deviations from the mean:

$$\widehat{\beta}^{ILS} = \frac{\widehat{\pi}_{21}}{\widehat{\pi}_{11}} = \frac{\sum c_t i_t / \sum i_t^2}{\sum y_t i_t / \sum i_t^2} = \frac{\sum c_t i_t}{\sum y_t i_t}.$$

This can only be done in exactly identified cases, like this, where all the various estimators (2SLS, IV, ILS and others) give the same estimates. Note that the fourth term in the equation is the IV estimator that appears below for the exactly identified case.

Above we considered identification just by restrictions on the coefficient matrices, $\mathbf{B}$ and $\mathbf{\Gamma}$. But we can also get identification partly through restrictions on the covariance matrix $\mathbf{\Omega}$. If we assume that $\mathbf{\Omega}$ is diagonal, this gives us $m(m-1)/2$ restrictions, that all the off diagonal elements are zero. If we also assume that $\mathbf{B}$ is triangular, all the elements above the diagonal are zero, this gives us another $m(m-1)/2$ restrictions. Together with the $m$ normalisation restrictions, this totals $m^2$ and the system is identified. Such a system is called recursive and can be estimated by OLS on each equation. An example is:

$$\begin{aligned} y_{1t} &= \gamma_1 x_t + \varepsilon_{1t} \\ y_{2t} &= \beta_{21} y_{1t} + \gamma_2 x_t + \varepsilon_{2t} \\ y_{3t} &= \beta_{31} y_{1t} + \beta_{32} y_{2t} + \gamma_3 x_t + \varepsilon_{3t} \end{aligned}$$

with $E(\varepsilon_{it}\varepsilon_{jt}) = 0$. $\varepsilon_{2t}$ is not correlated with $y_{1t}$ because there is no direct link, $y_{2t}$ does not influence $y_{1t}$, and no indirect link, $\varepsilon_{2t}$ is not correlated with $\varepsilon_{1t}$.So OLS is consistent. The system is

$$\begin{bmatrix} 1 & 0 & 0 \\ -\beta_{21} & 1 & 0 \\ -\beta_{31} & -\beta_{32} & 1 \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} x_t + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}.$$

This identification assumption is the basis of the Orthogonalised Impulse Response functions, obtained by Choleski decomposition, used to examine the effects of shocks to the VARs. EViews will impose these restrictions using the order you list the variables in. It also provides Generalised Impulse Response Functions, which do not require identifying assumptions and are invariant to the order. But these cannot be given a structural interpretation, since they do not identify the structural errors $\varepsilon_{it}$.

Notice that the VAR is the reduced form of a structural system in which instead of exogenous variables there appears the predetermined lagged values:

$$\begin{aligned} \boldsymbol{B}\boldsymbol{y}_t &= \boldsymbol{\Gamma}\boldsymbol{y}_{t-1}+\boldsymbol{u}_t;\ \ \mathbf{E}(\mathbf{u}_t\mathbf{u}_t')=\boldsymbol{\Omega} \\ \boldsymbol{y}_t &= \boldsymbol{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{y}_{t-1}+\boldsymbol{B}^{-1}\,\mathbf{u}_t \\ \boldsymbol{y}_t &= \boldsymbol{\Pi}\boldsymbol{y}_{t-1}+\mathbf{v}_t \\ \mathbf{E}(\mathbf{v}_t\mathbf{v}_t')=\boldsymbol{\Sigma} &= \boldsymbol{B}^{-1}\boldsymbol{\Omega}\boldsymbol{B}^{-1\prime} \end{aligned}$$

### 10.3. Instrumental Variables

Let us return to the LRM

$$y = X\beta + u$$

where $X$ is a $T \times k$ matrix, but the X are not exogenous, so $E(X'u) \neq 0$. This may happen because of simultaneity (some of the X are jointly determined with the $y$) or because some of the X are measured with error. In either case the OLS estimates will be biased and inconsistent. Suppose that there exists a $T \times i$, matrix of 'Instruments', $W$, where $i \geq k$, which are correlated with $X$ so that $E(W'X) \neq 0$ but are not correlated with the disturbances so that $E(W'u) = 0$. $W$ will include the elements of $X$ that are exogenous (including the column of ones for the constant), but we need at least one instrument for each endogenous $X$. If $i = k$, the model is said to be just-identified, if $i > k$ it is said to be over-identified. The condition $i \geq k$ is the same order condition, we encountered in simultaneous systems. There is also the rank condition from $E(W'X) \neq 0$ to ensure that $(W'X)$ is of full rank and $(W'X)^{-1}$ exists.

If the model is just or exactly identified, the consistent instrumental variable estimator is

$$\beta^{IV} = (W'X)^{-1}W'y$$

with variance-covariance matrix $\sigma^2(W'X)^{-1}W'W(W'X)^{-1}$. The efficiency of the estimator will increase (the size of the standard errors reduce) with the correlation between $W$ and $X$. Notice this estimator chooses the $\beta$ that imposes the

orthogonality condition:

$$\begin{aligned} W'\widetilde{u} &= 0 \\ W'(y - X\widetilde{\beta}) &= 0 \\ W'y &= W'X\widetilde{\beta} \\ (W'X)^{-1}W'y &= \widetilde{\beta}. \end{aligned}$$

Notice that in the case of a single right hand side endogenous variable, like the Keynesian consumption function above (where $y$ corresponds to $C_t$, $X$ to $Y_t$ and $W$ to $I_t$) the IV estimator is the ratio of the coefficient of the regression of $y_t$ on $w_t$ to the coefficient of the regression of $x_t$ on $w_t$.

If the model is over identified, the Generalised Instrumental Variable Estimator (GIVE), which is the same as the Two Stage Least Squares Estimator (2SLS) is obtained by first regressing each of the $X$ on the $W$;

$$X = WB + V$$

to give the $i \times k$ matrix of coefficients $\widehat{B} = (W'W)^{-1}W'X$, then calculating the predicted values of $X$ as: $\widehat{X} = W\widehat{B} = W(W'W)^{-1}W'X$. Substituting $X = \widehat{X} + \widehat{V}$ into the original regression we get:

$$y = (\widehat{X} + \widehat{V})\beta + u = \widehat{X}\beta + (\widehat{V}\beta + u).$$

Now $\widehat{X}$ is uncorrelated with $u$ since it is only a function of the $W$ which are uncorrelated with $u$, and is uncorrelated with $\widehat{V}$ by construction. Therefore it satisfies our exogeneity conditions. The GIVE estimator is

$$\begin{aligned} \beta^{GIV} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y \\ &= (X'W(W'W)^{-1}W'X)^{-1}X'W(W'W)^{-1}W'y \\ &= (X'P_wX)^{-1}X'P_wy \end{aligned}$$

with $P_w = W(W'W)^{-1}W'$ being a projection matrix. Its variance covariance matrix is $\sigma^2(X'P_wX)^{-1}$ and we estimate the residuals using the actual $X$ not their fitted values:

$$s^2_{IV} = (y - X\beta^{GIV})'(y - X\beta^{GIV})/(T - k)$$

This estimator chooses $\beta$ to make $X'P_wu = 0$. It minimises the estimate of $u'P_wu$, the IV minimand, rather than $u'u$ as OLS does. Many programs report the IV

minimand, which will be zero when the model is just identified. Show this by mutiplying out

$$(y - X\beta^{IV})'W(W'W)^{-1}W'(y - X\beta^{IV})$$

for the just identified case where $\beta^{IV} = (W'X)^{-1}W'y$.

### 10.4. Example: Endogenous variables

Suppose $x_{it}$ denote potentially endogenous variables, $w_{it}$ exogenous variables and the structural model is

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 w_{1t} + u_t$$

with $w_{2t}, w_{3t}, w_{4t}$ as potential instruments. Note that $X = [1\ x_{1t}\ x_{2t}\ w_{1t}]$, and $W = [1\ w_{1t}\ w_{2t}\ w_{3t}\ w_{4t}]$ so $k = 4$, $i = 5$ and the degree of overidentification is one. To get the fitted values, you run the two 'reduced form' regressions:

$$\begin{aligned} x_{1t} &= b_{10} + b_{11}w_{1t} + b_{12}w_{2t} + b_{13}w_{3t} + b_{14}w_{4t} + v_{1t} \\ x_{2t} &= b_{20} + b_{21}w_{1t} + b_{22}w_{2t} + b_{23}w_{3t} + b_{24}w_{4t} + v_{2t} \end{aligned}$$

to give you estimates of $\widehat{x}_{1t}, \widehat{x}_{2t}, \widehat{v}_{1t}, \widehat{v}_{2t}$; use the fitted values in the regression:

$$y_t = \beta_0 + \beta_1\widehat{x}_{1t} + \beta_2\widehat{x}_{2t} + \beta_3 w_{1t} + e_t \tag{10.2}$$

where $e_t = u_t + \beta_1\widehat{v}_{1t} + \beta_2\widehat{v}_{2t}$. The OLS estimates from this regression give the GIVE estimates of $\beta_i$ and the residuals are estimated as:

$$\widetilde{u}_t = y_t - (\beta_0^{GIV} + \beta_1^{GIV}x_{1t} + \beta_2^{GIV}x_{2t} + \beta_3^{GIV}w_{1t})$$

i.e. not using the fitted values.

You do not have to do GIVE/2SLS estimation in two stages in practice, since it is programmed into most packages. You just choose the option and list the instruments in addition to the model. Do not forget to include constant and right hand side exogenous variables among the instruments. However, it is usually a good idea to look at the F statistic on the reduced form regressions. A rule of thumb is that this should be greater than about 10. If the instruments are weak, do not explain $x_{it}$ very well, then the GIVE estimates will be badly biased and have large variance even in large samples.

If the instruments (or more precisely the over-identifying restrictions which exclude $w_{2t}, w_{3t}, w_{4t}$ from the structural model) are valid, these GIVE or 2SLS

residuals should be uncorrelated with the instruments. This can be tested by a Sargan (Bassman) test which involves regressing the GIVE residuals on all the instruments:

$$\widetilde{u}_t = c_0 + c_1 w_{1t} + c_2 w_{2t} + c_3 w_{3t} + c_4 w_{4t} + \varepsilon_t$$

and testing the hypothesis $c_1 = c_2 = c_3 = c_4 = 0$, this will be distributed $\chi^2(i-k)$, i.e. with degrees of freedom equal to the number of overidentifying restrictions. This can also be expressed as the ratio of the IV minimand (see above) to the GIVE variance. When the model is just identified, the IV minimand is zero, so the test is not defined.

To test whether the $x_{it}$ are in fact exogenous you can use the Wu-Hausman test. To do this you save the residuals from the reduced form regressions, $\widehat{v}_{1t}, \widehat{v}_{2t}$, and include them in the original regression, i.e. run by OLS:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 w_{1t} + \delta_1 \widehat{v}_{1t} + \delta_2 \widehat{v}_{2t} + u_t$$

then test the null that they are exogenous $H_0 : \delta_1 = \delta_2 = 0$. Rejection of the null (significant reduced form residuals) indicates that one or both of them are endogenous and GIVE should be used. This tests whether there is a significant difference between the OLS and GIVE estimates.

### 10.5. Example: measurement error

One cause of correlation between errors and regressors is measurement error. Suppose the model is

$$y_t = \beta x_t^* + \varepsilon_t; \tag{10.3}$$

where the variables are measured as deviations from their means and the true value $x_t^*$ is not observed, but we observe

$$x_t = x_t^* + v_t \tag{10.4}$$

where

$$\begin{aligned} E(\varepsilon_t) &= E(v_t) = 0 \\ E(\varepsilon_t) &= \sigma_\varepsilon^2;\ E(v_t) = \sigma_v^2 \end{aligned}$$

and $\varepsilon_t$ and $v_t$ are independent of each other and $x_t^*$. In some cases, e.g. where $x_t^*$ was the expected value of $x_t$ we may have suitable instruments and can apply instrumental variables, but suppose we do not.

Now

$$\begin{aligned} y_t &= \beta x_t^* + \varepsilon_t \\ &= \beta(x_t - v_t) + \varepsilon_t \\ &= \beta x_t + (\varepsilon_t - \beta v_t) \\ &= b x_t + u_t \end{aligned}$$

Clearly $x_t$ and $u_t$ are correlated $E(x_t, u_t) = E((x_t^* + v_t)(\varepsilon_t - \beta v_t) = -\beta\sigma_v^2$, hence $b$ will be an inconsistent estimator for $\beta$. $x_t$ is not weakly exogeous for $\beta$, because we need to know information about the marginal distribution of $x_t$, i.e. $\sigma_v^2$. We can observe the variances for $y_t$ and $x_t$ and their covariance:

$$S_{xx} = \frac{1}{T}\sum x_t^2;\ S_{yy} = \frac{1}{T}\sum y_t^2;\ S_{xy} = \frac{1}{T}\sum x_t y_t.$$

The variables are defined as deviations from their means. Assuming large samples, we can match these up with their theoretical values; defining the variance of $x_t^*$ as $\sigma_*^2$

$$\begin{aligned} S_{xx} &= \sigma_*^2 + \sigma_v^2 \\ S_{yy} &= \beta^2\sigma_*^2 + \sigma_\varepsilon^2 \\ S_{xy} &= \beta\sigma_*^2 \end{aligned}$$

The first line is got by squaring (10.4), and using the fact that the covariance of $v_t$ and $x_t^*$ is zero; the second line is got by squaring (10.3) ; the third line is got by multiplying (10.3) by (10.4). The OLS estimator from a regression of $y_t$ on $x_t$ is

$$\widehat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\beta\sigma_*^2}{\sigma_*^2 + \sigma_v^2} < \beta.$$

So unless $\sigma_v^2 = 0$, the direct least squares estimator is biased downwards. Consider taking the inverse of the coefficient of the reverse regression of $x_t$ on $y_t$ which is

$$\widehat{d} = \frac{S_{yy}}{S_{xy}} = \frac{\beta^2\sigma_*^2 + \sigma_\varepsilon^2}{\beta\sigma_*^2} > \beta$$

so unless $\sigma_\varepsilon^2 = 0$, this reverse least squares estimator is biased upwards. This gives us a bound in large samples

$$\widehat{b} < \beta < \widehat{d}$$

This may be useful in seeing the size of the effect of the possible measurement error. Unfortunately this does not generalise to more than two variables in any simple way; but with more variables there may be other ways to deal with measurement error.

Up to now we have considered point identification, a parameter is either identified or not identified. In this case the parameter is identified as being within a bound, $\beta$ is between $\widehat{b}$ and $\widehat{d}$. There are other cases of identification within bounds. The model is not point identified because we have three pieces of information $S_{ij}$ and four theoretical parameters. One extra piece of information would identify it. If we knew that the errors in measurement were the same size as the errors in equation $\sigma_\varepsilon^2 = \sigma_v^2$ (or any other known ratio) this would identify it. In the case where $\sigma_\varepsilon^2 = \sigma_v^2 = \sigma^2$ then

$$\begin{aligned} S_{xx} &= \sigma_*^2 + \sigma^2 \\ S_{yy} &= \beta^2\sigma_*^2 + \sigma^2 \\ S_{xy} &= \beta\sigma_*^2 \end{aligned}$$

From the third equation, $\sigma_*^2 = S_{xy}/\beta$; from the first equation

$$\sigma^2 = S_{xx} - \sigma_*^2 = S_{xx} - S_{xy}/\beta$$

substituting these in the second equation gives

$$S_{yy} = \beta^2 \left(S_{xy}/\beta\right) + S_{xx} - S_{xy}/\beta.$$

Rearranging this shows $\beta$ is a solution to the quadratic equation

$$\beta^2 S_{xy} + \beta(S_{xx} - S_{yy}) - S_{xy} = 0.$$

Which by the usual formula

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

gives

$$\frac{-(S_{xx} - S_{yy}) \pm \sqrt{(S_{xx} - S_{yy}) + 4S_{xy}^2}}{2S_{xy}}.$$

### 10.5.1. Principal Components

If you have a number of indicators of an unobserved variable or factor, then Principal Components, PC, can provide an estimate. These are used in factor augmented VARs and other applications. The PC of a $T \times N$ data matrix $\mathbf{X}$ (which is usually standardised by subtracting the mean of the variable and dividing by the standard deviation) are the linear combination which explains as much as possible of the variance of all the $\mathbf{X}$. The first principal component is $z_1 = \mathbf{X}a_1$ where the variance of $z$ i.e. $z_1'z_1 = \sum z_{1t}^2 = a_1'\mathbf{X}'\mathbf{X}a_1$ is maximised. Notice that if the data are standardised, $\mathbf{X}'\mathbf{X}$ is the correlation matrix of the data, otherwise it is the covariance matrix.This $z_1'z_1$ can be made as large as you like depending on the units of $a_1$ so we need to choose a normalisation that determines scale, it is usual to use $a_1'a_1 = 1$. Set this up as a Lagrangian,

$$\begin{aligned} \pounds &= a_1'\mathbf{X}'\mathbf{X}a_1 - \lambda_1(a_1'a_1 - 1) \\ \frac{\partial \pounds}{\partial a_1} &= \mathbf{X}'\mathbf{X}a_1 - \lambda_1 a_1 = 0. \end{aligned}$$

Thus $\lambda_1$ is the largest eigenvalue and $a_1$ the corresponding eigenvector. If the data are standardised $\lambda_1$ tells you the proportion of the variation in $\mathbf{X}$ explained by the first PC. One can get the other PCs in a similar way and they will be orthogonal (uncorrelated). This gives you $N$ new variables which are linear combinations of the $\mathbf{X}$. One uses a subset of these corresponding to the $r$ largest eigenvalues. There are various ways to choose $r$, one is to use any PCs where the eigenvalues from standardised data are greater than one. In EViews to get PCs define a group, the variables in $\mathbf{X}$, open the group; choose View and one of the options will be to calculate the PCs for the group. These are known as static PCs or factors, dynamic factors take the PCs or the long-run covariance matrix (spectral density matrix) of $\mathbf{X}$.