A Modern Approach to Teaching Econometrics

David F. Hendry and Bent Nielsen*
Economics Department, Oxford University.

September 24, 2009

Abstract

We explain the computer-based approach to the teaching of econometrics used in Oxford from the elementary to the advanced. The aims are to enable students to critically evaluate published applied studies and to undertake empirical research in economics. The unified theoretical framework is that of likelihood, using likelihood-ratio tests for inference and evaluation, and focusing on developing well-specified empirical models of interesting economic issues. A sequence of increasingly realistic models is developed from independent, identically distributed binary data through to selecting cointegrated equations in the face of structural breaks–in a one-year course.

Preface by David Hendry

Although my first interchanges with Clive Granger involved disagreements over modeling non-stationary economic time series, that stimulus led to his famous formulation of the concept of cointegration, and associated developments in econometric modeling (see ?, ??, ?, ??, and http://nobelprize.org/nobel_prizes/economics/laureates/2003/granger-lecture.pdf, based on ?, ??, and ?, ??). Clive was already well known both for his ideas on causality (see ?, ??, appraised in ?, ??, and distinguished from exogeneity in ?, ??), and for re-emphasizing the dangers in applying static regression models to integrated data (in ?, ??, following pioneering research by ?, ??). From my first visit to the University of California at San Diego in 1975, where Clive had moved in 1974, our friendship blossomed, built around a common desire to improve the quality of econometric model building, especially by a better match to the empirical evidence: Clive's contributions to doing so have been one of the most successful research programmes in econometrics, and are a lasting contribution (for formal Obituaries, see ?, ??, and ?, ??). My own approach focused on implementing modeling methods, and led to our only joint publication (?, ??), discussing automatic modeling. Clive also kept the theory of economic forecasting under the spotlight when it was not in fashion (see ?, ?, ???), another interest we had in common (including our accidentally adopting the same title in ?, ??). In addition to his astonishing creativity and innovative ideas, Clive was a master of written and presentational clarity, so we also shared a desire to communicate both with students (Clive supervised a large number of successful doctoral students) and colleagues on a world-wide basis. The paper which follows develops modeling ideas in the teaching domain, where major changes in how we explain and teach econometrics to the next generation could further enhance the quality with which econometrics is applied to the many pressing problems facing the world.

*The background research has been financed by the United Kingdom Economic and Social Research Council through the funding of RES-062-23-0061 and RES-000-27-0179. We are indebted to Jennifer L. Castle, Jurgen A. Doornik, Vivien L. Hendry for many helpful comments on a previous draft, to Jennie and Jurgen for many invaluable contributions to the research and software, and in memory of Clive Granger's long-term creative stimulus.

1 Introduction

There are six reasons why now is a good time to review the teaching of econometrics. Over the last quarter century, there have been:
(1) massive changes in the coverage, approaches, and methods of econometrics;
(2) huge improvements in computer hardware and computational methods;[1]
(3) improvements to software, data, and graphics capabilities, which have been at least as impressive;
(4) considerable advances in teaching methods, from mathematical derivations written on blackboards, through overheads to live computer projection;
(5) few discussions of computer-based teaching of econometrics since ? (?)? proposed an approach based on PcGive (?, ??, describe its history);
(6) new approaches to teaching econometrics (see e.g., ?, ??), emphasizing empirical modeling.

The last is the focus of this paper, where every student is taught while having continuous computer access to automatic modeling software (in Oxford, based on ?, ??, within OxMetrics: see ?, ??). Computer-based teaching of econometrics is feasible at all levels, from elementary, through intermediate, to advanced. We cannot cover all those aspects, and will only briefly describe how we teach the first steps in econometrics, noting some tricks that help retain student interest, before moving on to model selection in non-stationary data. Such ideas can even be communicated to less mathematically oriented undergraduates, enabling them to progress in a year from introducing independent, identically distributed (IID) binary data to selecting cointegrated equations in the face of structural breaks.

At the outset, we assume no knowledge of even elementary statistical theory, so first explain the basic concepts of probability; relate those to distributions, focusing on location, spread and shape; turn to elementary notions of randomness; then apply these ideas to distributions of statistics based on data. There are six central themes:

- likelihood;
- testing assumptions;
- economically relevant empirical applications;
- mastery of software to implement all the procedures;
- an emphasis on graphical analysis and interpretation;
- rigorous evaluation of all 'findings'.

Derivations are only used where necessary to clarify the properties of methods, concepts, formulations, empirical findings, and interpretations, but also simultaneously upskill students' mathematical competence. By adopting the common framework of likelihood, once that fundamental idea is fully understood in the simplest IID binary setting, generalizations to more complicated models and data generation processes (DGPs) follow easily. A similar remark applies to evaluation procedures, based on likelihood ratio tests: the concept is the same, even though the distribution may alter with more complicated and realistic DGPs with time dependence, non-stationary features etc. (which include changes in means and variances as well as stochastic trends). The theory and empirical sections of the course proceed in tandem. Once a given statistical or econometric idea has been introduced, explained and illustrated, it is applied in a computer class where every student has their own workstation on line to the database and software.

Sections 2 and 3 describe the first steps in theory and practice respectively, then §4 discusses regression analysis graphically and as a generic 'line-fitting' tool. Those set the scene for introducing simple

[1] These have, of course, been ongoing from hand calculations at its start in the 1930s, through punched card/tape-fed mainframe computers, workstations, to powerful PCs and laptops.

models and estimation in §5, leading to more general models in §6, and model selection in §7. Section 8 notes some implications for forecasting–Clive's other great interest–and §9 concludes.

2 Theory first steps

To introduce elementary statistical theory, we consider binary events in a Bernoulli model with independent draws, using sex of a child at birth as the example. This allows us to explain sample versus population distributions, and hence codify these notions in distribution functions and densities.

Next, we consider inference in the Bernoulli model, discussing expectations and variances, then introduce elementary asymptotic theory (the simplest law of large numbers and central limit theorem for IID processes) and inference.

It is then a small generalization to consider continuous variables, where we use wages (w_i) in a cross section as the example. The model thereof is the simplest case, merely $w_i = \beta + u_i$, where β is the mean wage and u_i characterizes the distribution around the mean. Having established that case, regression is treated as precisely the same model, so is already familiar. Thus, building on the simple estimation of means leads to regression, and hence to logit regression, and then on to bivariate regression models.

2.1 The way ahead

In the lectures, regression is applied to an autoregressive analysis of the Fulton fish-market data from ? (?)?. The natural next step is to model price and quantity jointly as a system, leading to simultaneity and identification, resolved using as instruments dummies for 'structural breaks' induced by stormy weather at sea. In the system, over-identified instrumental variables regression is simply reduced-rank regression. This makes it easy to move on to unit roots and a system analysis of cointegration, picking up model selection issues en route, and illustrating the theory by Monte Carlo simulation experiments. Thus, each topic segues smoothly into the next.

3 Empirical first steps

Simultaneously, we teach *OxMetrics* and *PcGive* so that students can conduct their own empirical work. Here, we focus on the computer-class material, which moves in parallel, but analyzes different data each year. We have collected a large databank of long historical time series from 1875–2000 (available at http://press.princeton.edu/titles/8352.html), and offer the students the choice of modeling any one of the key series, such as the unemployment rate (denoted U_r), gross domestic product (g, where a lower-case letter denotes the log of the corresponding capital), price inflation (Δp, where P is the implicit deflator of G), or real wages ($w - p$). We assume students chose U_r, leading to a classroom-wide discussion of possible economic theories of unemployment based around supply and demand for a factor of production–but replete with ideas about 'too high real wages', 'lack of demand for labour', 'technological change', 'population growth', '(im)migration', 'trade union power', 'overly high unemployment benefits' etc.–as well as the relevant institutions, including companies, trade unions and governments. An advantage of such a long-run data series is that many vague claims are easily rebutted as the sole explanation, albeit that one can never refute that they may be part of the story.

The next step is to discuss the measurement and sources of data for unemployment and working population, then get everyone to graph the level of U_r (here by *OxMetrics* as in fig. 1). It is essential to carefully explain the interpretation of graphs in considerable detail, covering the meaning of axes, their units, and any data transforms, especially the roles and properties of logs. Then one can discuss the

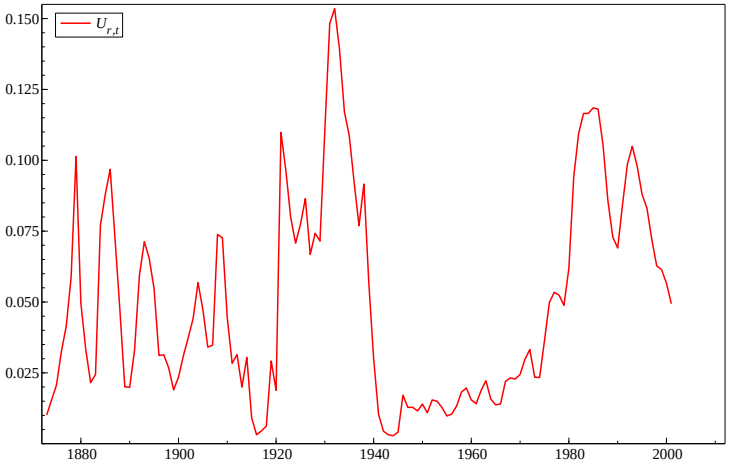

Figure 1 Graph of historical data on UK unemployment rate.

salient features, such as major events, cycles, trends and breaks, leading to a first notion of the concept of non-stationarity, an aspect that Clive would have liked. It is also easy to relate unemployment to the student's own life prospects, and those of their fellow citizens. Everyone in the class is made to make a new comment in turn–however simple–about some aspect of the graph, every time. Nevertheless, it is usually several sessions before students learn to mention the axes' units first.

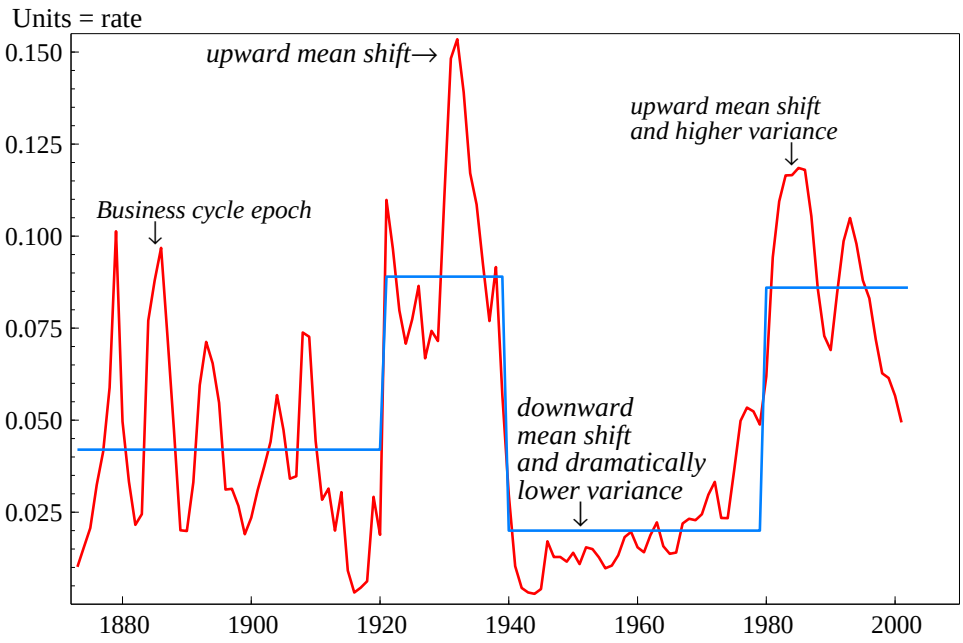

Figure 2 General comments about unemployment rate.

We aim to produce students who will be able to critically interpret published empirical findings, and sensibly conduct their own empirical research, so one cannot finesse difficulties such as non-stationarity and model selection. The non-stationarity visible in whatever series is selected is manifest(e.g., fig. 1), including shifts in means and variances, any 'epochs' of markedly different behaviour, and changes in persistence. Figure 2 shows possible general comments, whereas fig. 3 adds specific details, most of which might arise in discussion. Thus, a detailed knowledge of the historical context is imparted as a key aspect of modeling any time series. This also serves to reveal that most of the major shifts are due to non-economic forces, especially wars and their aftermaths.

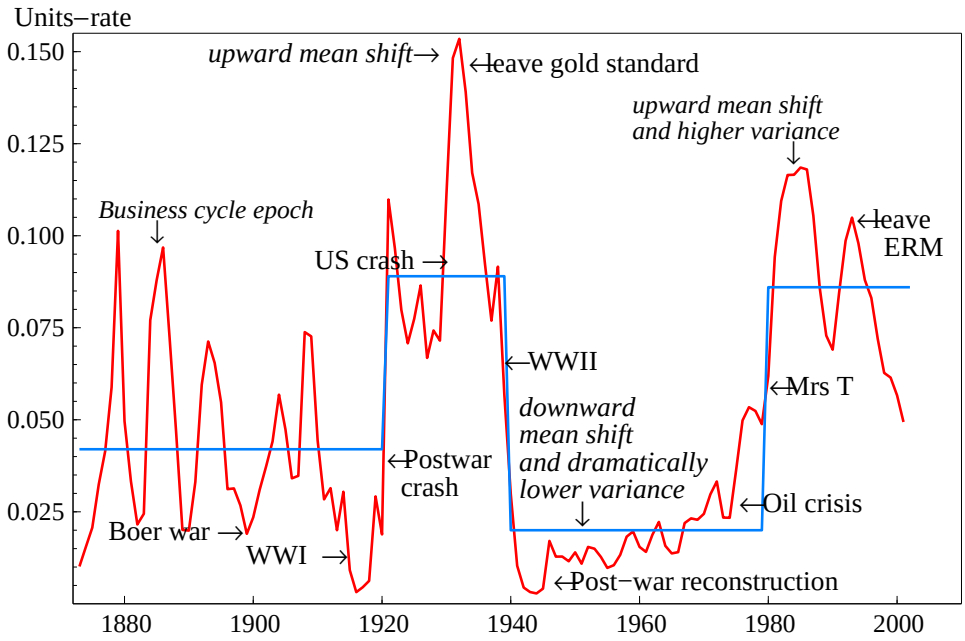

Figure 3 All comments about unemployment rate.

Next, the students are required to calculate the differences of $U_{r,t}$, namely $\Delta U_{r,t} = U_{r,t} - U_{r,t-1}$ and plot the resulting series as in fig. 4. Again, everyone is asked to discuss the hugely different 'look' of the graph, especially its low persistence, constant mean of zero, and how it relates to fig. 1, including the possibility that changes in the reliability and coverage of the data sources may partly explain the apparent variance shift, as well as better economic policy. $\Delta U_{r,t}$ is clearly not stationary, despite differencing, another key lesson.

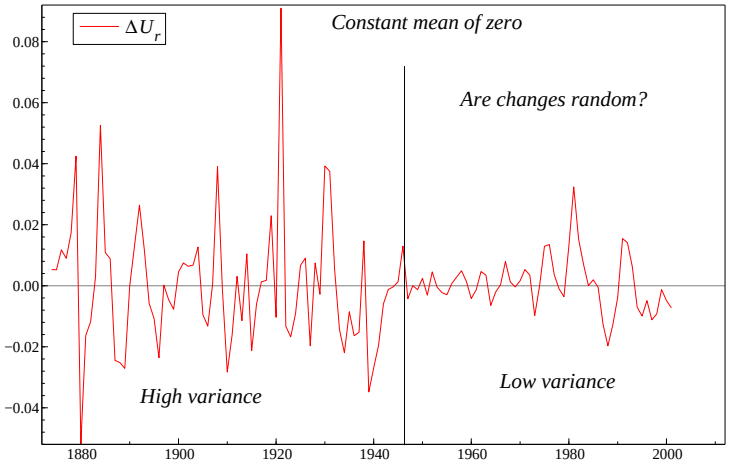

Figure 4 Changes in the unemployment rate.

3.1 Theory and evidence

A further feature of our approach is to relate distributional assumptions to model formulation. The basic example is conditioning in a bivariate normal distribution, which is one model for linear regression. In turn that leads naturally to the interpretation of linear models and their assumptions, and hence to model design, modeling, and how to judge a model. Here, the key concepts are a well-specified

model (matching the sample distribution), congruence (also matching the substantive context), exogeneity (valid conditioning), and encompassing (explaining the gestalt of results obtained by other models of the same phenomena). These are deliberately introduced early on when only a few features need to be matched, as they will be important when models become larger and more complicated, usually requiring computer-based automatic selection.

In fact, it is difficult to formulate simple theoretical models of unemployment with any ability to fit the long non-stationary sample of data in fig. 1. Consequently, to start down the road of linking theory and evidence in a univariate model, we postulate a 'golden-growth' explanation for deviations of unemployment from its historical mean, so assume that $U_{r,t}$ is non-integrated yet non-stationary. The measure of the steady-state equilibrium determinant is given by:

$$d_t = R_{L,t} - \Delta p_t - \Delta g_t \tag{1}$$

where $R_{L,t}$ is the long bond rate: precise definitions of the data are provided in ? (?)?, who also develops a model based on (1). When the real cost of capital $(R_{L,t} - \Delta p_t)$ exceeds the real growth rate Δg_t, then $d_t > 0$ so the economy will slow and $U_{r,t}$ will rise; and conversely when $d_t < 0$, so $U_{r,t}$ will converge to its historical mean for $d_t = 0$. Of course all the variables in (1) are jointly determined with $U_{r,t}$ in any equilibrium, so the resulting 'explanation' is conditional, but allows a simple yet viable model to be developed, with the possibility of later extension to a multivariate setting.

As ever, we first look at the graphs of each series as in fig. 5a, and discuss their properties, then consider d_t in relation to $U_{r,t}$ in panel b (where the series are adjusted for means and ranges to maximize the visual correlation).

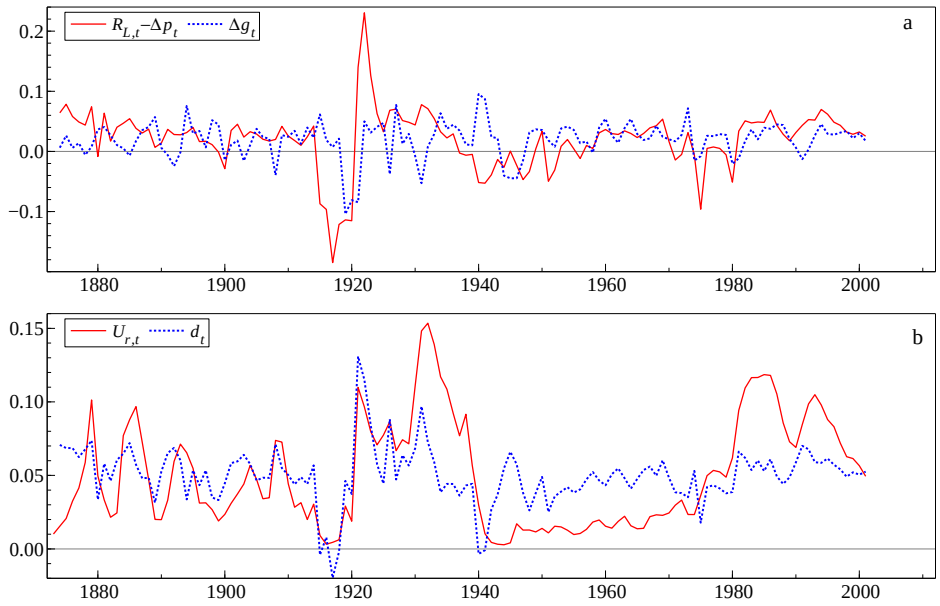

Figure 5 Graphs of $R_{L,t} - \Delta p_t$ and Δg_t; $U_{r,t}$ and d_t.

4 Regression concepts

Once regression theory has been explained, it is easily illustrated by a scatter plot, as in fig. 6a. Adding a single line allows one to explain the slope and intercept graphically. By further adding the projections

from the data to the line, least squares can be understood visually as the line that minimizes the squared deviations: see fig. 6b. Such graphs now reveal that the $U_{r,t}$ on d_t relation is fine in the tails, but 'erratic' in the middle.

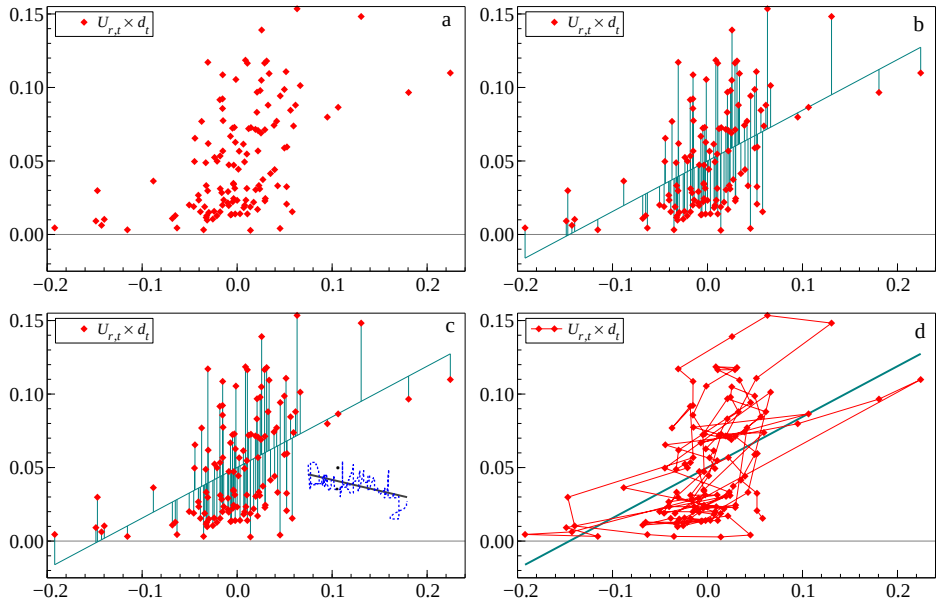

Figure 6 Scatter plots and regressions of $U_{r,t}$ on d_t.

Having established the basics of regression, a more penetrating analysis moves on to the five key concepts that underpin linear regression interpreted as conditioning in a bivariate normal distribution:

- exogeneity of the regressor;
- IID errors;
- normality;
- linear functional form;
- parameter constancy.

Failure on any of these induces model mis-specification, and likelihood ratio, or equivalent, tests of the corresponding assumptions can be introduced. At each stage, we relate the maths derivations as needed for understanding the graphs–but always using the same basic principles: data graphs suggest a putative DGP and hence a model of its distribution function, leading to the likelihood function. Maximize that likelihood as a function of the postulated parameters, obtaining an appropriate statistic from the score equation, and derive its distribution. Finally, apply the resulting method to the data, interpret the results and evaluate the findings to check how well the evidence matches the assumed DGP. The same approach is just applied *seriatim* to ever more complicated cases.

4.1 Regression as 'non-parametric'

To 'demystify' regression analysis as just line fitting, use the 'pen' in *OxMetrics* to have each students write his/her name on a graph, then run a regression through it: see fig. 6c. The pixels are mapped to world coordinates (which can be explained using the graphics editor), so become 'data' in the ($U_{r,t}$, d_t) space, and hence one can estimate a regression for that subset. Consequently, projections can even be added to the signature regression. Most students are intrigued by this capability, and many gain important insights, as well as being amused.

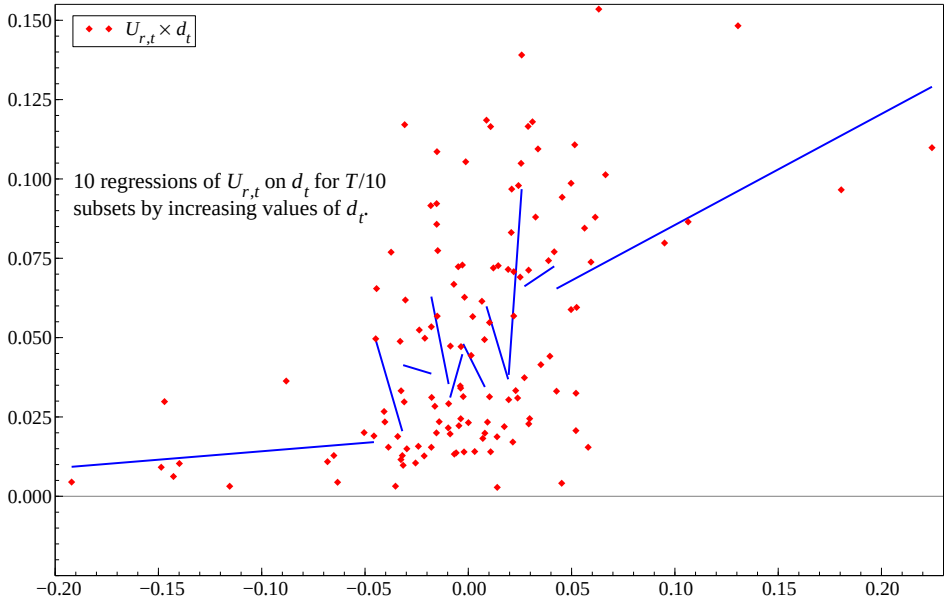

Figure 7 Regressions for $U_{r,t}$ on d_t for each tenth of the data.

Next, show how to join up the original data points to create "Phillips' loops" (see ?, ??), tracing out the dynamics as in fig. 6d. This serves to highlight the absence of the time dimension from the analysis so far, and is a convenient lead into taking account of serial dependence in data.

Many routes are possible at this point–one could further clarify the underlying concepts, or models, or methods of evaluation. We illustrate several sequential regressions graphically as that leads to recursive methods for investigating parameter constancy: see fig. 7. Alternatively, *OxMetrics* graphs provide an opportunity to introduce the basics of LaTex by naming variables, as shown in the graphs here, or by writing formulae on the figures. Even minimal LaTex skills will prove invaluable later as estimated models can be output that way, and pasted directly into papers and reports (as used below).

4.2 Distributions

Graphics also ease the visualization of distributions. Plotting the histograms of $U_{r,t}$ and $\Delta U_{r,t}$ with their interpolated densities yields fig. 8. Such figures can be used to explain non-parametric/kernel approaches to density estimation, or simply described as a 'smoothed' histogram.

More importantly, one can emphasize the very different features of the density graphs for the level of $U_{r,t}$ and its change. The former is like a uniform distribution–many values are roughly equally likely. The distribution for $\Delta U_{r,t}$ is closer to a normal with some outliers. Thus, differencing alters distributional shapes. This can be explained as the *unconditional* distribution of $U_{r,t}$ versus its distribution *conditional on the previous value*, so panel b is plotting the distribution of the deviation of $U_{r,t}$ from $U_{r,t-1}$.

4.3 Time series and randomness

That last step also serves to introduce the key concept of non-randomness. Regression subsumes correlation, which has by now been formally described, so can be used to explain correlograms as correlations between successively longer lagged values: fig. 9 illustrates.

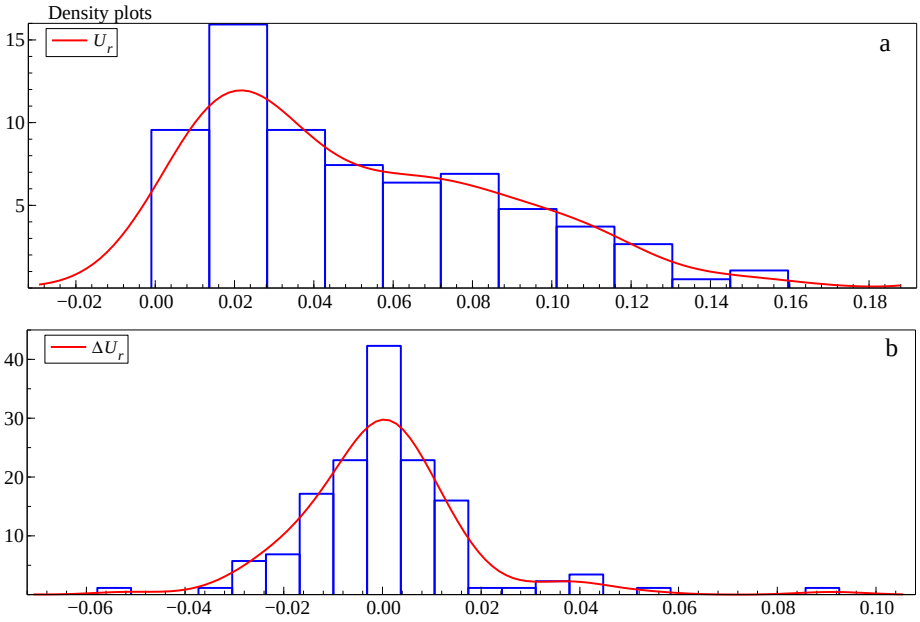

Figure 8 Distribution of the unemployment rate $U_{r,t}$ and its change $\Delta U_{r,t}$.

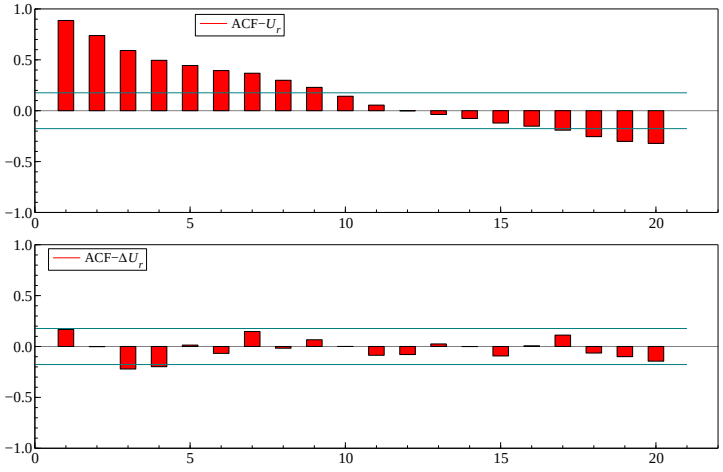

Figure 9 Correlograms for the unemployment rate $U_{r,t}$ and its change $\Delta U_{r,t}$.

The plots reveal that $U_{r,t}$ has many high autocorrelations–indeed the successive autocorrelations almost lie on a downward linear trend–whereas $\Delta U_{r,t}$ has almost no autocorrelation at any lag. Thus, changes in $U_{r,t}$ are 'surprise-like': again this comparison highlights the huge difference between unconditional and conditional behaviour. In turn, we can exploit the different behaviour of $U_{r,t}$ and $\Delta U_{r,t}$ to introduce dynamics, by plotting $U_{r,t}$ against its own lag $U_{r,t-1}$ then graphically adding the regression, as in fig. 10a (and panel b for $\Delta U_{r,t}$).

4.4 Well-specified models

Now one can explain well-specified models as needing all the properties of the variables in a model to match simultaneously–in terms of dynamics, breaks, distributions, linear relations, etc.– otherwise there will be systematic departures from any claimed properties. Tests of each null hypothesis are then discussed, albeit using Lagrange multiplier approximate F-tests rather than likelihood ratio, namely:
F_{ar} for k^{th}-order serial correlation as in ? (?)? and ? (?)?;

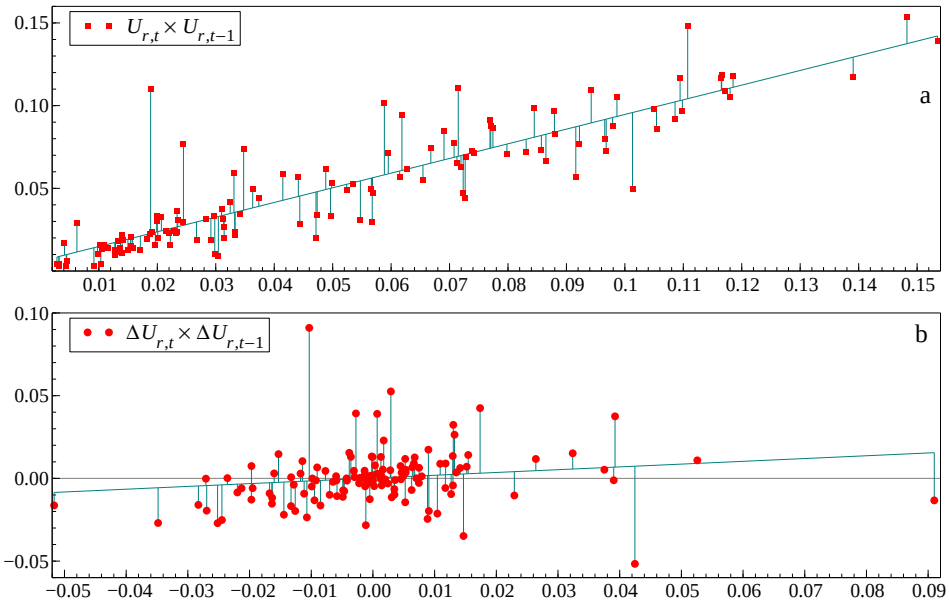

Figure 10 Unemployment rate $U_{r,t}$ and its change regressed on their own first lag.

$\mathsf{F}_{\mathsf{het}}$ for heteroskedasticity as in ? (?)?;
$\mathsf{F}_{\mathsf{reset}}$ for functional form following ? (?)?);
$\mathsf{F}_{\mathsf{arch}}$ for k^{th}-order autoregressive conditional heteroskedasticity, from ? (?)?;
$\mathsf{F}_{\mathsf{Chow}}$ for parameter constancy over k periods as in ? (?)?; and
$\chi^2_{\mathsf{nd}}(2)$ for normality (a chi-square test: see ?, ??).
Below * and ** denote significant at 5% and 1% respectively.

Having established the basics, the scene is set for formal estimation of a regression.

5 Model estimation

Estimating the static or long-run (later interpreted as 'cointegrated') relation $U_{r,t} = \beta_0 + \beta_1 d_t + e_t$ yields:

$$\begin{aligned} \widehat{U}_{r,t} &= \underset{(0.003)}{0.050} + \underset{(0.052)}{0.345}\, d_t \qquad (2) \\ \mathsf{R}^2 &= 0.26 \;\; \widehat{\sigma} = 0.0315 \;\; \mathsf{F}_{\mathsf{GUM}}(1,126) = 44.64^{**} \end{aligned}$$

Here, R^2 is the squared multiple correlation, $\widehat{\sigma}$ is the residual standard deviation, and coefficient standard errors are shown in parentheses. The test $\mathsf{F}_{\mathsf{GUM}}$ is for significance of the general unrestricted model, that is the joint significance of all regressors (d_t) apart from the intercept. The estimates suggest that unemployment rises/falls as the real long-run interest rate is above/below the real growth rate (i.e., $d_t \lesseqgtr 0$). All the assumptions are easily tested, yielding $\mathsf{F}_{\mathsf{ar}}(2,124) = 180.4^{**}$; $\mathsf{F}_{\mathsf{arch}} = 229.9^{**}$; $\mathsf{F}_{\mathsf{reset}}(1,125) = 0.33$; $\mathsf{F}_{\mathsf{het}}(2,123) = 2.62$; $\chi^2_{nd}(2) = 15.0^{**}$. These tests show that the model is poorly specified.

Figure 11 records the fitted and actual values, their cross-plot, the residuals scaled by $\widehat{\sigma}$, and their histogram and density with N[0,1] for comparison, visually confirming the formal tests. Once again, it is clear that the model is badly mis-specified, but it is not clear which assumptions are invalid. However, we have now successfully applied the key concepts to *residuals*.

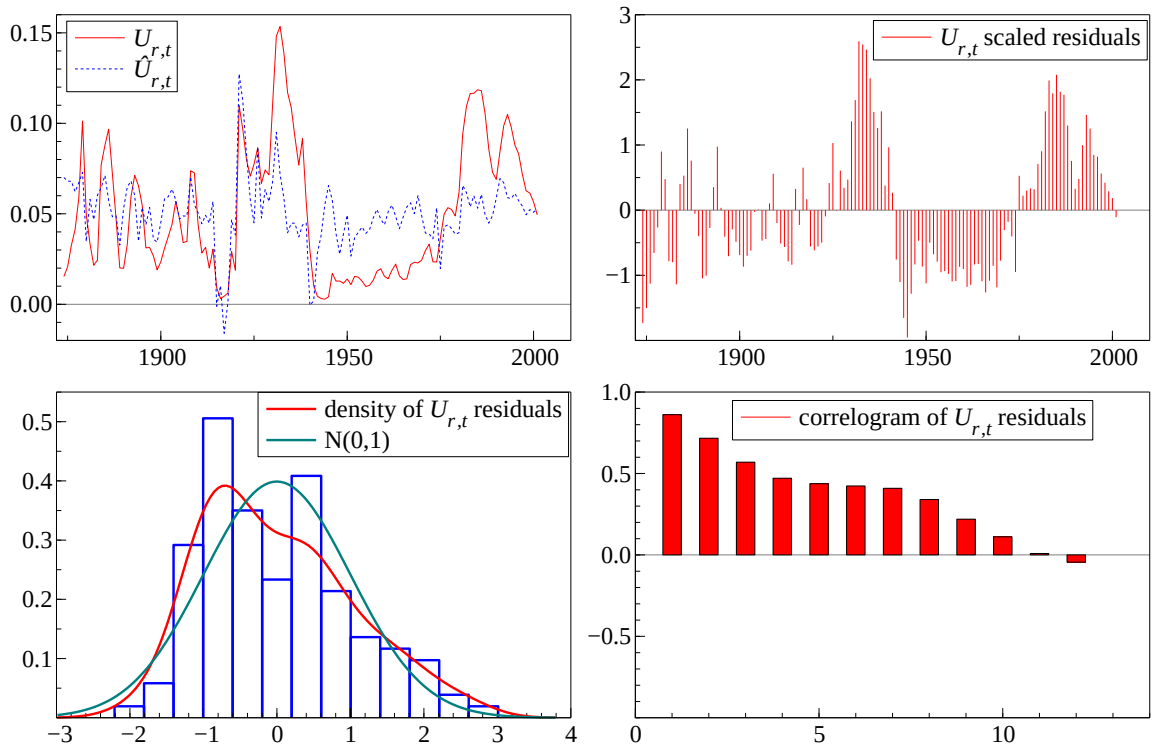

Figure 11 Graphical output from U_r on d.

5.1 Simple dynamic models

Another univariate model worth illustrating is that of $U_{r,t}$ on $U_{r,t-1}$, namely $U_{r,t} = \gamma_0 + \gamma_1 U_{r,t-1} + \epsilon_t$. This form was implicit in fig. 10–and can also be related to the earlier graphs for $\Delta U_{r,t}$:

$$\begin{aligned} \widehat{U}_{r,t} &= \underset{(0.003)}{0.006} + \underset{(0.040)}{0.887}\, U_{r,t-1} \qquad (3) \\ \mathsf{R}^2 &= 0.79 \;\; \widehat{\sigma} = 0.017 \;\; \mathsf{F}_{\mathsf{GUM}}(1,126) = 485.7^{**} \\ \chi^2_{nd}(2) &= 33.0^{**} \;\; \mathsf{F}_{\mathsf{ar}}(2,124) = 3.8^{*} \;\; \mathsf{F}_{\mathsf{arch}}(2,124) = 0.55 \\ \mathsf{F}_{\mathsf{het}}(2,123) &= 0.42 \;\; \mathsf{F}_{\mathsf{reset}}(1,125) = 0.01 \end{aligned}$$

Most of the mis-specification tests are considerably improved, but the model in (3) is still mis-specified, with an obvious outlier in 1920, as fig. 12(b) shows. The long-run solution in (3) is $0.006/(1-0.887)$ or 5.3% unemployment–which is close to the intercept in (2)–and although one cannot in fact reject the hypothesis of a unit root, that provides an opportunity to explain the rudiments of stochastic trends, possibly illustrated by Monte Carlo simulation of the null distribution.

However, this is crunch time: having postulated our models of the DGP, we find strong rejection on several tests of specification, so something has gone wrong. Multiple testing concepts must be clarified: each test is derived under its separate null, but assuming all other aspects are well specified. Consequently, any other mis-specification rejections contradict the assumptions behind such derivations: once any test rejects, none of the others is trustworthy as the assumptions underlying their calculation are also invalidated. Moreover, simply 'correcting' any one problem–such as serial correlation–need not help, as the source may be something else altogether, such as parameter non-constancy over time. A more viable approach is clearly needed–leading to general-to-simple...

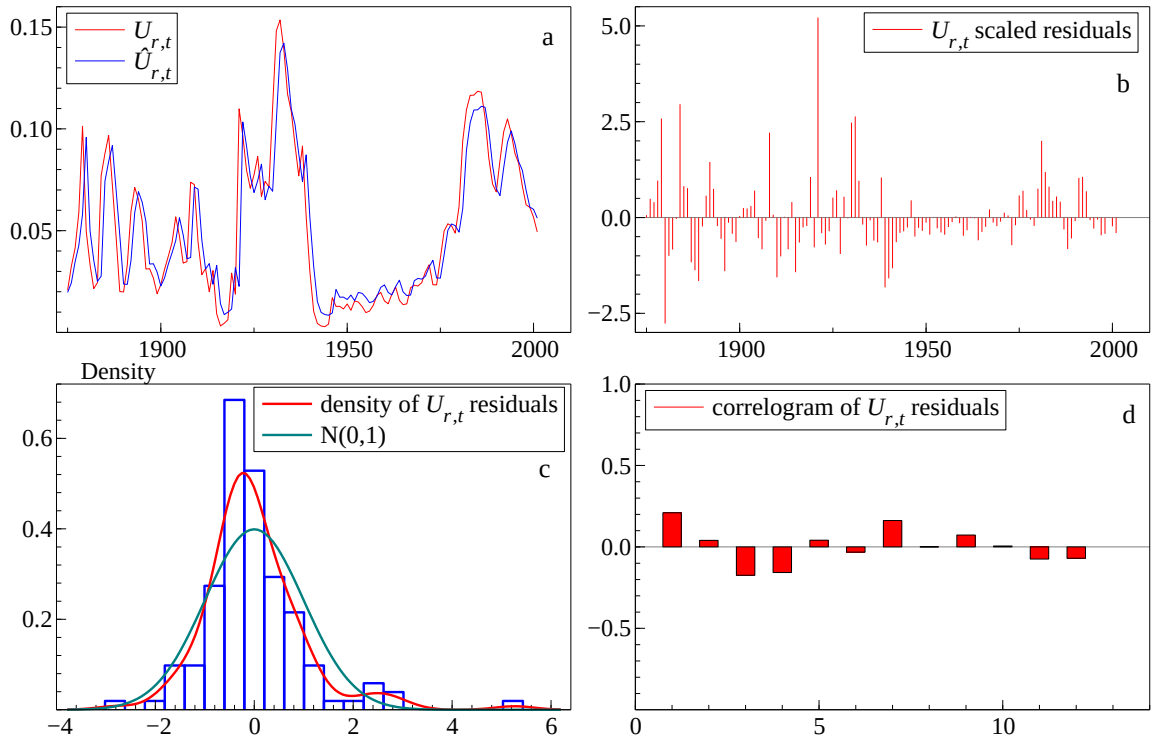

Figure 12 $U_{r,t}$ on $U_{r,t-1}$ graphical output.

6 More general models

It is time to introduce a dynamic model which also has regressors, nesting both (3) and (2), namely $U_{r,t} = \beta_0 + \beta_1 d_t + \beta_2 U_{r,t-1} + \beta_3 d_{t-1} + \upsilon_t$. Estimation delivers:

$$\begin{aligned} \widehat{U}_{r,t} &= \underset{(0.002)}{0.007} + \underset{(0.03)}{0.24}\, d_t + \underset{(0.04)}{0.86}\, U_{r,t-1} - \underset{(0.03)}{0.10}\, d_{t-1} \qquad (4) \\ \mathsf{R}^2 &= 0.88 \;\; \widehat{\sigma} = 0.013 \;\; \mathsf{F}_{\mathsf{GUM}}(3,123) = 308.2^{**} \;\; \mathsf{F}_{\mathsf{ar}}(2,121) = 2.5 \\ \chi^2_{nd}(2) &= 7.2^{*} \;\; \mathsf{F}_{\mathsf{arch}}(1,121) = 3.1 \;\; \mathsf{F}_{\mathsf{het}}(6,116) = 4.2^{**} \;\; \mathsf{F}_{\mathsf{reset}}(1,122) = 4.2^{*} \end{aligned}$$

Although (4) is not completely well-specified, it is again much better, and certainly dominates both earlier models, as F-tests based on the 'progress' option in *OxMetrics* reveal. While illustrating progressive research, the exercise also reveals the inefficiency of commencing with overly simple models, as nothing precluded commencing from (4). Assuming cointegration has been explained, one can show that a unit root can be rejected in (4) ($\mathsf{t}_{ur} = -3.9^{**}$ on the PcGive unit-root test: see ?, ??, and ?, ??), so U_r and d are 'cointegrated' (or co-breaking, as in ?, ??). Next, the long-run solution can be derived by taking the expected value of the error as zero, and setting the levels to constants such that:

$$(1 - \beta_2)\, U^*_r = \beta_0 + (\beta_1 + \beta_3)\, d^*,$$

so:

$$U^*_r = \frac{\beta_0}{1 - \beta_2} + \frac{\beta_1 + \beta_3}{1 - \beta_2} d^*,$$

which yields $U^*_r = 0.052 + 1.02 d^*$ for the estimates in (4). The coefficient of d at unity is much larger than that of 0.35 in (2), and suggests a one-for-one reaction in the long run.

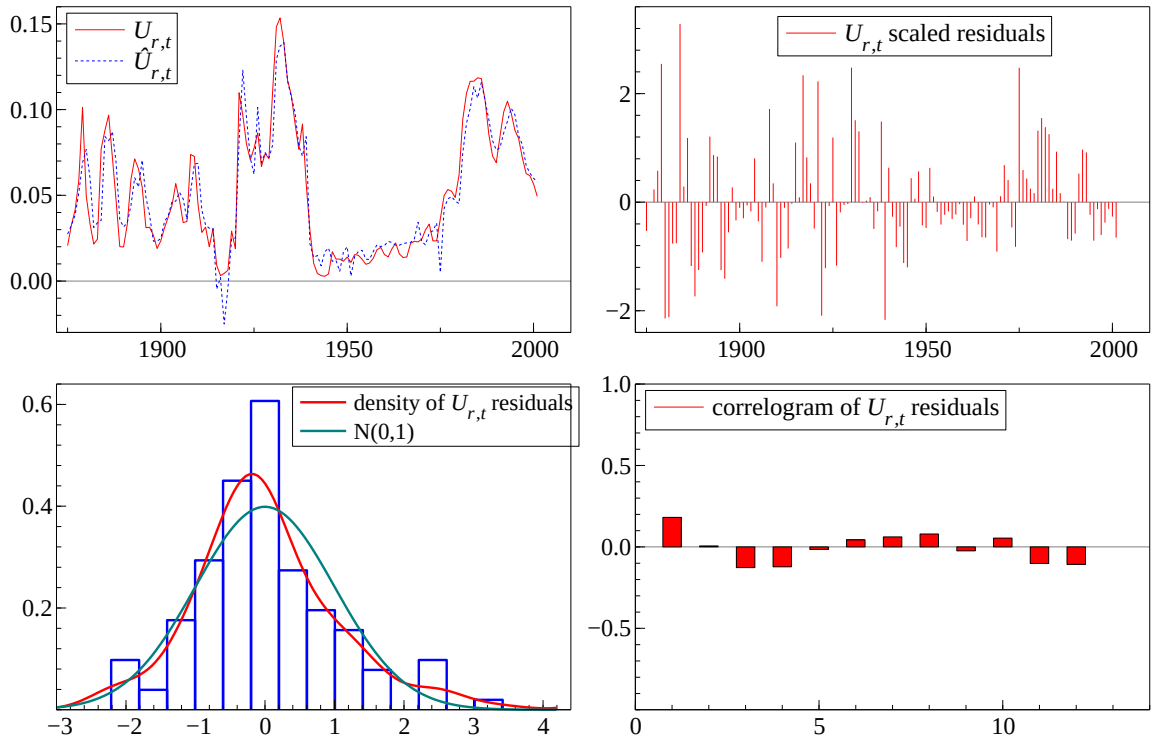

Figure 13 General $U_{r,t}$ model graphical output.

6.1 Extensions

There is as much to discuss as ones desires at this stage. For example, there are few outliers, but there are some negative fitted values (suggesting a logit formulation, which may also attenuate the residual heteroscedasticity). Challenge students to formulate alternative explanations, and test their proposals against the evidence–and see if they can encompass (4), by explaining its performance from their model. One can also check model constancy by formal recursive methods, building on the earlier graphical approach. Figure 14 records the outcome for (4): despite the apparently 'wandering' estimates, the constancy tests–which are scaled by their 1% critical values–do not reject. It is surprising that such a simple representation as (4) can describe the four distinct epochs of unemployment about equally accurately.

7 Model selection

Having shown the dangers of simple approaches, general-to-specific model selection needs to be explained. In a general dynamic model, one cannot know in advance which variables will matter: some will, but some will not, so selection is required. Indeed, any test followed by a decision entails selection, so in empirical research, selection is ubiquitous, however unwilling practitioners are to admit its existence. 'Model uncertainty' is pandemic–every aspect of an empirical model is uncertain, from the existence of any such relation in reality, the viability of any 'corroborating' theory, and the measurements of the variables, as well as the choice of the specification and every assumption needed in the formulation, such as exogeneity, constancy, linearity, independence etc. One must confront such issues openly if graduating students are to be competent practitioners.

It is feasible to sketch the theory of model selection in the simplest case. We use the idea of choosing between two decisions, namely keeping or eliminating a variable, where there are two states of nature,

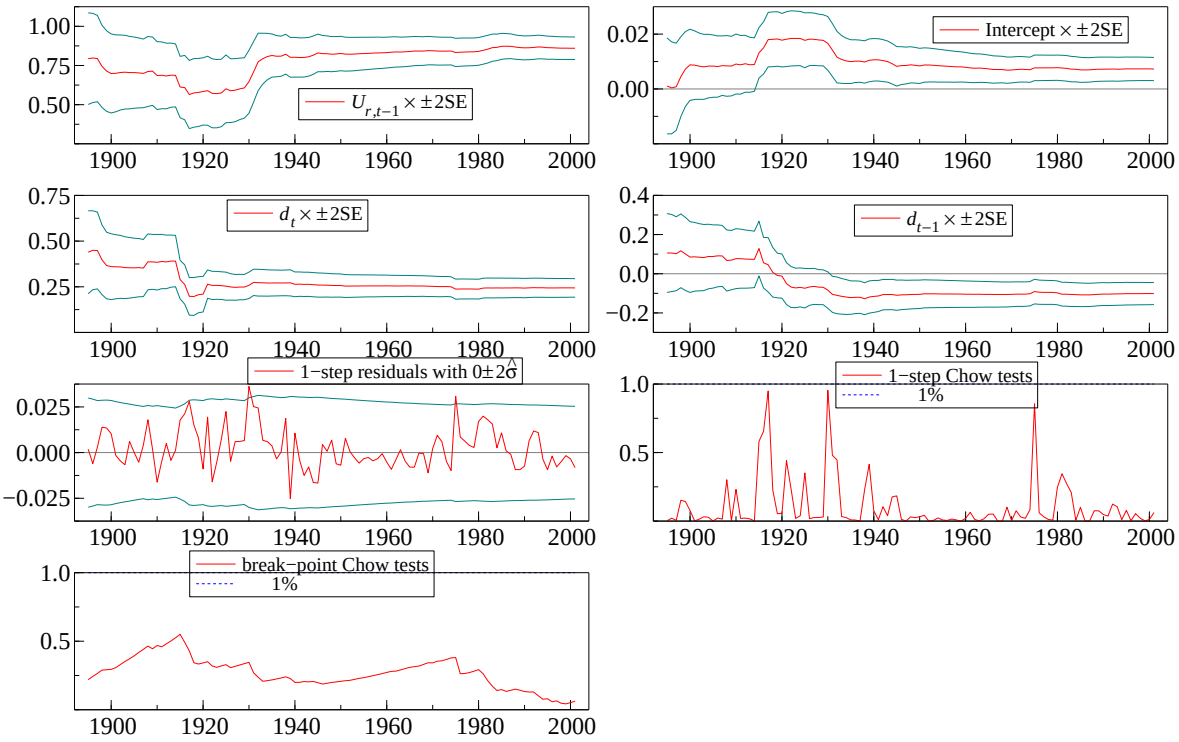

Figure 14 $U_{r,t}$ model recursive output.

namely the variable is in fact relevant or irrelevant in that setting. The mistakes are 'retain an irrelevant variable' and 'exclude a relevant variable', akin to probabilities of type I and II errors. Consider the perfectly orthogonal, correctly specified regression model:

$$y_t = \beta_1 z_{1,t} + \beta_2 z_{2,t} + \epsilon_t \tag{5}$$

where all variables have zero means, $\mathsf{E}[z_{1,t} z_{2,t}] = 0$, the β_i are constant, and $\epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]$. Denote the t^2-statistics testing H_0: $\beta_j = 0$ by t_j^2, and let c_α be the desired critical value for retaining a variable in (5) when $\mathsf{t}_j^2 \geq c_\alpha^2$. When either (or both) $\beta_j = 0$ in (5), the probability of falsely rejecting the null is determined by the choice of c_α–conventionally set from $\alpha = 0.05$. There is a 5% chance of incorrectly retaining one of the variables on t^2, but a negligible probability (0.0025) of retaining both. When one (or both) $\beta_j \neq 0$, the power of the t^2-statistic to reject the null depends on the non-centrality $\beta_j^2 / \mathsf{V}\left[\widehat{\beta}_j\right] \simeq T \beta_j^2 \sigma_{z_j}^2 / \sigma_\epsilon^2$ where $\mathsf{E}[z_{j,t}^2] = \sigma_{z_j}^2$: this can be evaluated by simulation. Thus, all the factors affecting the outcome of selection are now in place.

7.1 Understanding model selection

The interesting case, however, is generalizing to:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \tag{6}$$

where N is large (say 40). Order the N sample t^2-statistics as $\mathsf{t}_{(N)}^2 \geq \mathsf{t}_{(N+1)}^2 \geq \cdots \geq \mathsf{t}_{(1)}^2$, then the cut-off between included and excluded variables is given by $\mathsf{t}_{(n)}^2 \geq c_\alpha^2 > \mathsf{t}_{(n-1)}^2$, so n are retained and $N - n$ eliminated. Thus, variables with larger t^2 values are retained on average, and all others are eliminated. Importantly, only one decision is needed to select the model even for $N = 1000$ when

there are $2^{1000} = 10^{301}$ possible models. Consequently, 'repeated testing' does not occur, although path searches during model reduction may give the impression of 'repeated testing'. Moreover, when N is large, one can set the average false retention rate at one irrelevant variable by setting $\alpha = 1/N$, so $\alpha N = 1$, at a possible cost in lower correct retention.

Of course, there is sampling uncertainty, as the t_j^2 are statistics with distributions, and on any draw, those close to c_α^2 could randomly lie on either side–for both relevant and irrelevant variables. It is important to explain the key role of such marginal decisions: empirical t^2-values close to the critical value c_α^2 are the danger zone, as some are likely to arise by chance for irrelevant variables, even when α is as small as 0.001. Fortunately, it is relatively easy to explain how to bias correct the resulting estimates for sample selection, and why doing so drives estimates where t_j^2 just exceeds c_α^2 close to zero (see e.g. ?, ??).

Students have now covered the basic theory of *Autometrics* (see ?, ??), and, despite their inexperience, can start to handle realistically complicated models using automatic methods, which has led to a marked improvement in the quality of their empirical work. Nevertheless, a final advance merits discussion, namely handling more variables than observations in the canonical case of impulse-indicator saturation.

7.2 Impulse-indicator saturation

The basic idea is to 'saturate' a regression by adding T indicator variables to the candidate regressor set. Adding all T indicators simultaneously to any equation would generate a perfect fit, from which nothing is learned. Instead, exploiting their orthogonality, add half the indicators, record the significant ones, and remove them: this step is just 'dummying out' $T/2$ observations as in ? (?)?. Now add the other half, and select again, and finally combine the results from the two models and select as usual. A feasible algorithm is discussed in ? (?)? for a simple location-scale model where $x_i \sim \mathsf{IID}\left[\mu, \sigma_x^2\right]$, and is extended to dynamic processes by ? (?)?. Their theorem shows that after saturation, $\widetilde{\mu}$ is unbiased, and αT indicators are retained by chance on average, so for $\alpha = 0.01$ and $T = 100$, then 1 indicator will be retained by chance under the null even though there are more variables than observations. Thus, the procedure is highly efficient under the null that there are no breaks, outliers or data contamination.

Autometrics uses a more sophisticated algorithm than just split halves, (see ?, ??), but the selection process is easily illustrated live in the classroom. Here we start with 2 lags of both $U_{r,t}$ and d_t and set $\alpha = 0.0025$ as $T = 126$. Selection locates 8 significant outliers (1879, 1880, 1884, 1908, 1921, 1922, 1930, and 1939: the 19^{th} century indicators may be due to data errors), and yields (not reporting the indicators):[2]

$$\begin{aligned}
\widehat{U}_{r,t} &= \underset{(0.0015)}{0.004} + \underset{(0.02)}{0.15}\, d_t + \underset{(0.06)}{1.29}\, U_{r,t-1} - \underset{(0.02)}{0.09}\, d_{t-1} - \underset{(0.06)}{0.39}\, U_{r,t-2} \qquad (7)\\
\mathsf{R}^2 &= 0.95 \;\; \widehat{\sigma} = 0.008 \;\; \mathsf{F}_{\mathsf{ar}}(2,111) = 1.59\\
\chi^2_{nd}(2) &= 7.98^{*} \;\; \mathsf{F}_{\mathsf{arch}}(1,1249) = 0.09 \;\; \mathsf{F}_{\mathsf{het}}(14,103) = 1.03 \;\; \mathsf{F}_{\mathsf{reset}}(2,111) = 1.41
\end{aligned}$$

The long-run solution from (7) is $U_r^* = 0.05 + 0.62 d^*$ so has a coefficient of d that is smaller than in (4). However, no diagnostic test is significant other than normality, and the model is congruent, other than the excess of zero residuals visible in the residual density.

[2]Three more indicators for 1887, 1910 and 1938 are retained at $\alpha = 0.01$, with a similar equation.

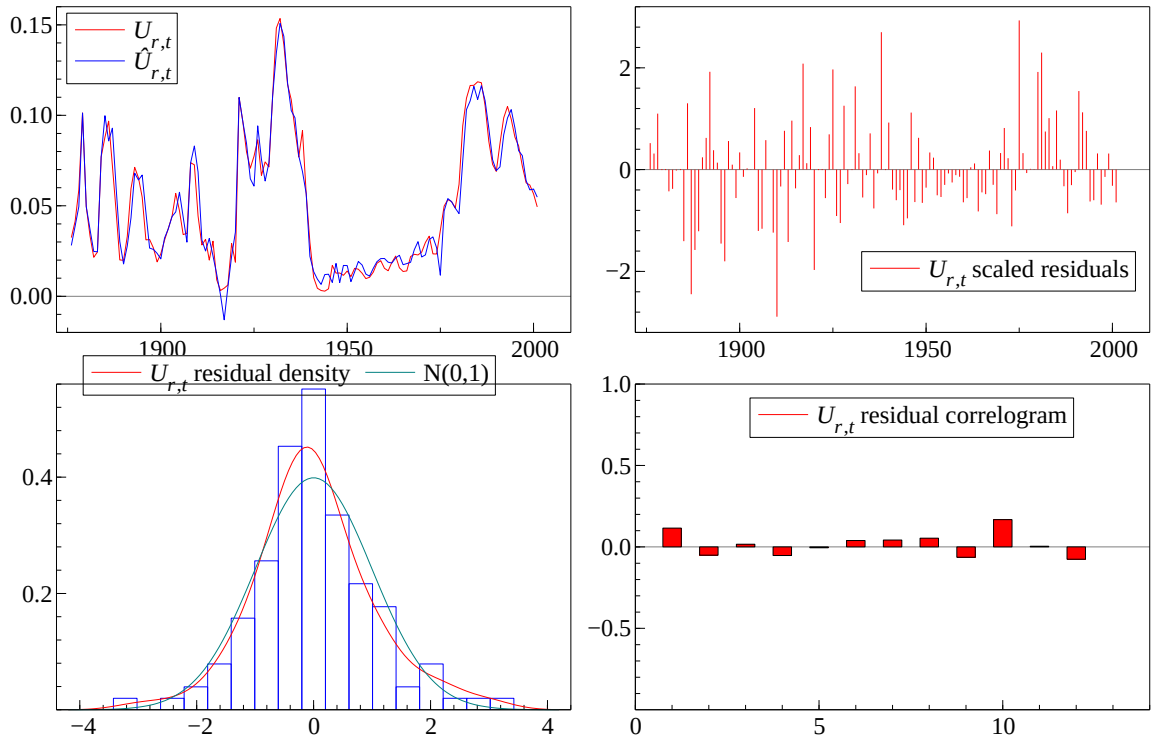

Figure 15 Graphical description of final model of $U_{r,t}$.

7.3 Monte Carlo of model selection

How to evaluate how well such general-to-specific model selection works? Everyone in the class generates a different artificial sample from the same DGP, which they design as a group, then they all apply *Autometrics* to their own sample. Pool the class results and relate the outcomes to the above theoretical 'delete/keep' calculations–then repeat at looser/tighter selection criteria, with and without impulse-indicator saturation to see that the theory matches the practice, and works.

7.4 Evaluating the selection

At an advanced level, exogeneity issues can be explored, based on impulse-indicator saturation applied to the marginal model for the supposedly exogenous variable (see ?, ??). Here, that would be d_t, so develop a model of it using only lagged variables and indicators: *Autometrics* takes under a minute from formulation to completion at (say) $\alpha = 0.0025$, as for (7).

$$\begin{aligned} d_t &= \underset{(0.05)}{0.55}\, d_{t-1} - \underset{(0.03)}{0.16}\, I_{1915} - \underset{(0.03)}{0.13}\, I_{1917} + \underset{(0.03)}{0.24}\, I_{1921} \\ &\quad + \underset{(0.03)}{0.12}\, I_{1926} + \underset{(0.03)}{0.10}\, I_{1931} - \underset{(0.03)}{0.14}\, I_{1940} - \underset{(0.03)}{0.09}\, I_{1975} \qquad (8) \\ \widehat{\sigma} &= 0.028 \;\; \mathsf{F}_{\mathsf{ar}}(2,116) = 1.32 \;\; \mathsf{F}_{\mathsf{arch}}(1,124) = 0.19 \\ \chi^2_{nd}(2) &= 2.20 \;\; \mathsf{F}_{\mathsf{het}}(2,116) = 4.72^* \;\; \mathsf{F}_{\mathsf{reset}}(2,116) = 4.52^* \end{aligned}$$

Only the indicator for I_{1921} is in common, and the remainder are not, so must all co-break with (7). All the dates correspond to recognizable historical events, albeit that other important dates are nor found. Even that for 1921 (one of most eventful years for the UK) is 0.056 in (7) as against 0.24 in (8), so does not suggest that a break in the latter is communicated to the former, which would violate

exogeneity. Adding the indicators from (8) to (7), however, delivers $\mathsf{F}(6,107) = 4.28^{**}$ so strongly rejects exogeneity, even beyond the 0.0025 level used in selection. The main 'culprit' is 1975 (which was omitted from (7) by *Autometrics* as it induced failure in several diagnostic tests), but interestingly, the long-run solution is now $U_r^* = 0.042 + 1.02d^*$ so is back to the original. None of the indicators from (7) is significant if added to (8).

8 Forecasting

First, one must establish how to forecast. Given the unemployment equation above, for 1-step forecasts, $U_{r,T},\ d_T$ are known, the past indicators are now zero, but d_{T+1} needs to be forecast if (7) is to be used for $\widehat{U}_{r,T+1}$. Thus, a system is needed, and is easily understood in the 2-stages of forecasting d_{t+1} from (8):

$$\widehat{d}_{T+1} = \underset{(0.05)}{0.55}\ d_T \tag{9}$$

and use that in:

$$\widehat{U}_{r,T+1} = \underset{(0.0015)}{0.004} + \underset{(0.02)}{0.15}\ \widehat{d}_{T+1} + \underset{(0.06)}{1.29}\ U_{r,t} - \underset{(0.02)}{0.09}\ d_T - \underset{(0.06)}{0.39}\ U_{r,T-1} \tag{10}$$

Brighter students rapidly notice that the net effect of d on the forecast outcome is essentially zero as substituting (9) into (10) yields $0.55 \times 0.15 - 0.09 = -0.0075$. Thus, the forecast model is no better than an autoregression in $U_{r,t}$. Indeed simply selecting that autoregression delivers (indicators not reported):

$$\widehat{U}_{r,t} = \underset{(0.06)}{1.29}\ U_{r,t-1} - \underset{(0.06)}{0.34}\ U_{r,t-2}$$

with $\widehat{\sigma} = 0.0096$. This is the first signpost that 'forecasting is different'.

That 'vanishing trick' would have been harder to spot when the model was expressed in equilibrium-correction form to embody the long-run relation $e = U_r - 0.05 - d$ as a variable:

$$\Delta\widehat{U}_{r,t} = \underset{(0.05)}{0.37}\ \Delta U_{r,t-1} + \underset{(0.02)}{0.17}\ \Delta d_t - \underset{(0.02)}{0.07}\ e_{t-1}$$

Since multiple breaks have already been encountered, it is easy to explain the real problem confronting economic forecasting, namely breaks. Simply extrapolating an in-sample estimated model (or a small group of models pooled in some way) into the future is a risky strategy in processes where location shifts occur. Here, the key shift would be in the equilibrium mean of 5% unemployment, and that has not apparently occurred over the sample, despite the many 'local mean shifts' visible in figure 2. To make the exercise interesting, we go back to 1979 and the election of Mrs. Thatcher, and dynamically forecast U_r and d over the remainder of the sample as shown in figure 16 (top row) with $\pm 2\widehat{\sigma}$ error fans.

The forecast failure over the first few years in U_r is clear, associated with the failure to forecast the jump in d, following her major policy changes. Going forward two years and repeating the exercise (bottom row) now yields respectable forecasts.

9 Conclusion

Computer-based teaching of econometrics enhances the students' skills, so they can progress from binary events in a Bernoulli model with independent draws to model selection in non-stationary data in a

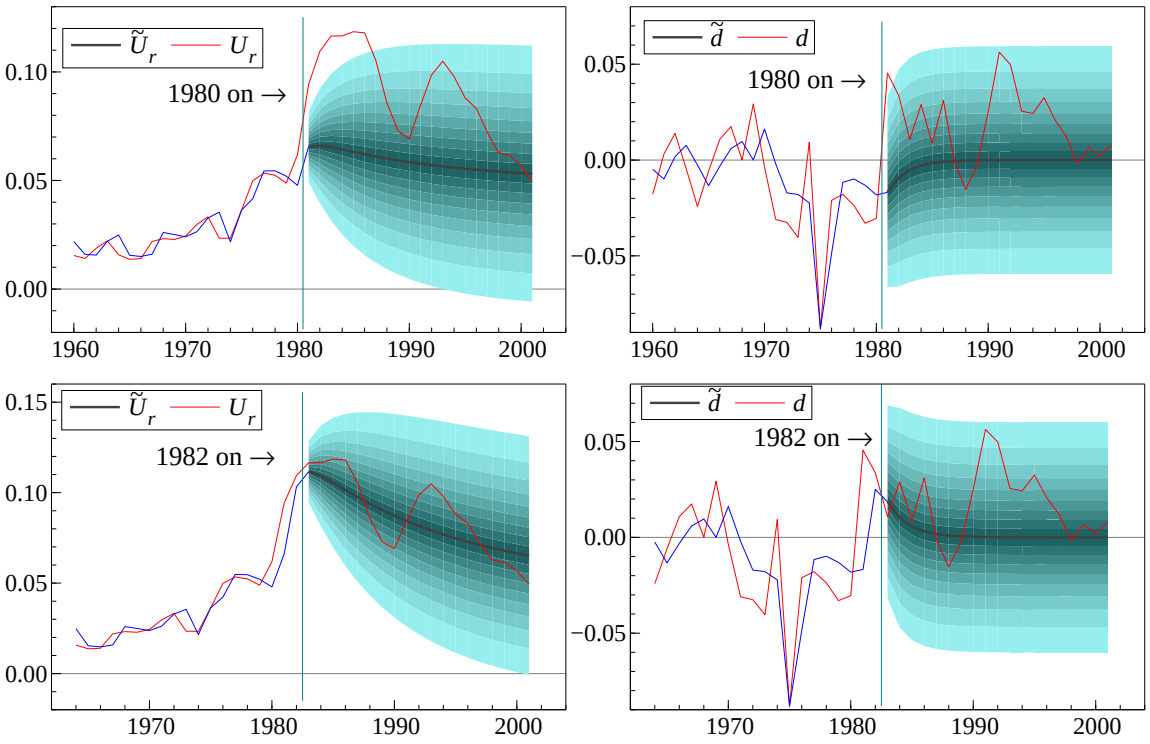

Figure 16 Dynamic forecasts of $U_{r,t}$ and d_t over 1980–2001.

year-long course which closely integrates theory and empirical modeling. Even in that short time, they can learn to build sensible empirical models of non-stationary data, aided by automatic modeling. We believe Clive would have approved.