Empirical Model Discovery

David F. Hendry

Department of Economics, Oxford University

OxMetrics Conference, Washington, March, 2010

Research jointly with Jennifer Castle, Jurgen Doornik, Bent Nielsen and Søren Johansen

'Any sufficiently advanced technology is indistinguishable from magic.' Arthur C. Clarke, *Profiles of The Future*, 1961

Introduction

Many features of models not derivable from theory
Need empirical evidence on:
which variables are actually relevant,
any lagged responses (dynamic reactions),
functional forms of connections (non-linearity),
structural breaks and unit roots (non-stationarities),
simultaneity (or exogeneity), expectations, etc.

Almost always must be data-based on available sample:
need to discover what matters empirically.

Four key steps:
(1) define the framework–the target for modelling;
(2) embed that target in a general formulation;
(3) search for the simplest acceptable representation;
(4) evaluate the findings.

Automatic methods can outperform

[A] formulation: many candidate variables, long lag lengths, non-linearities, outliers, and parameter shifts

[B] selection: handle more variables than observations, yet deliver high success rates by multi-path search

[C] estimation: near unbiased estimates despite selection

[D] evaluation: automatically conduct a range of pertinent tests of specification and mis-specification

Now routinely create models with $N > 200$ variables: and select successfully even if only $T = 150$ observations.

This talk aims to explain how we do so: it only *appears* to be magic!

Approach embodied in *Autometrics*: see Doornik (2009)
So let's perform some magic.

Basis of approach

Data generation process (DGP):
joint density of all variables in economy

Impossible to accurately theorize about or model precisely
Too high dimensional and too non-stationary

Need to reduce to manageable size in 'local DGP' (LDGP):
the DGP in space of r variables $\{\mathbf{x}_t\}$ being modelled

Theory of reduction explains derivation of LDGP:
joint density $\mathsf{D}_{\mathbf{x}}(\mathbf{x}_1 \dots \mathbf{x}_T|\boldsymbol{\theta})$**.**
Acts as DGP, but 'parameter' $\boldsymbol{\theta}$ **may be time varying**

Knowing LDGP, can generate 'look alikes' for relevant $\{\mathbf{x}_t\}$
which only deviate from actual data by unpredictable noise

Cannot do better than know $\mathsf{D}_{\mathbf{x}}(\cdot)$**–**
so the LDGP $\mathsf{D}_{\mathbf{x}}(\cdot)$ **is the target for model selection**

Formulating a 'good' LDGP

Choice of r variables, $\{\mathbf{x}_t\}$, to analyze determines modelling target LDGP, $\mathsf{D}_{\mathbf{x}}(\cdot)$, and its properties.

Prior reasoning, theoretical analysis, previous evidence, historical and institutional knowledge all important.

Should be $90 + \%$ of effort in an empirical analysis.

Aim to avoid complex, non-linear and non-constant LDGPs. Crucial not to omit substantively important variables: small set $\{\mathbf{x}_t\}$ more likely to do so

Given $\{\mathbf{x}_t\}$, have defined the target $\mathsf{D}_{\mathbf{x}}(\cdot)$ in (1).

Now embed that target in a general model formulation.

Extensions for discovering a 'good' model

Second of four key steps: **extensions of $\{\mathrm{x}_t\}$ determine how well LDGP is approximated**.

Four main groups of automatic extensions:
additional candidate variables that 'might' be relevant;
lag formulation, implementing a **sequential factorization**;
functional form transformations for non-linearity;
impulse-indicator saturation (IIS) for parameter non-constancy and data contamination.

Must also handle mapping to non-integrated data, **conditional factorizations**, and simultaneity.

'Good choices' facilitate locating a congruent parsimonious-encompassing model of LDGP.
Congruence: matches the evidence on desired criteria;
parsimonious: as small a model as viable;
encompassing: explains the results of all other models.

Selecting and evaluating the model

Extensions create the general unrestricted model (GUM). The GUM should nest the LDGP, making it a special case; **reductions commence from GUM to locate specific model**.

Selection is step (3):
search for the simplest acceptable representation.
Much of this talk concerns how that selection is done, and checking how well it works.

Finally, step (4):
evaluate the findings–and the selection process.
Includes tests of new aspects, such as
super exogeneity (essentially causality) for policy, and
parameter invariance (constancy across regimes).

Route map

(1) **Discovery**

(2) Automatic model extension

(3) 1-cut model selection

(4) Automatic estimation

(5) Automatic model selection

(6) Model evaluation

(7) Excess numbers of variables $N > T$

(8) An empirical example: food expenditure

Conclusions

Discovery

Discovery: learning something previously unknown.
Cannot know how to discover what is not known–
unlikely to be a 'best' way of doing so.

Many discoveries have element of chance:
luck: Fleming–penicillin from a dirty petrie dish
serendipity: Becquerel–discovery of radioactivity
'natural experiment': Dicke–role of gluten in celiac disease
trial and error: Edison–incandescent lamp
brilliant intuition: Faraday–dynamo from electric shock
false theories: Kepler–regular solids for planetary laws
valid theories: Pasteur–germs not spontaneous generation
systematic exploration: Lavoisier–oxygen not phlogiston
careful observation: Harvey–circulation of blood
new instruments: Galileo–moons around Jupiter
self testing: Marshall–ulcers caused by *Helicobacter pylori*.

Common aspects of discovery

Theoretical reasoning also frequent: Newton, Einstein.
Science is both inductive and deductive.

Must distinguish between:
context of discovery–where 'anything goes', and
context of evaluation–rigorous attempts to refute.

Accumulation and consolidation of evidence crucial:
data reduction a key attribute of science (think $E = mc^2$).

Six aspects in common to above examples of discovery.

First, going ***outside*** existing state.
Second, ***search*** for something.
Third, ***recognition*** of significance of what is found.
Fourth, ***quantification*** of what is found.
Fifth, ***evaluating*** discovery to ascertain its 'reality'.
Sixth, ***parsimoniously summarize*** vast information set.

Discovery in economics

Discoveries in economics mainly from theory.

An economic analysis suggests:

$$y = f(\mathbf{x}) \tag{1}$$

where y depends on n 'explanatory' variables $\mathbf{x}$,
for a sample of T observations, $\{y_t, \mathbf{x}_t\}$.
Form of $f(\cdot)$ in (1) depends on:
utility or loss functions of agents,
constraints they face, & information they possess,
aggregation across heterogeneous individuals,
specification of a unit of time.
In addition, observations may be contaminated,
underlying processes integrated,
abrupt shifts may induce various forms of breaks.
Previous econometrics covertly concerned discovery.

Classical econometrics: covert discovery

Postulated:

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \epsilon_t, \quad t = 1, \ldots, T \tag{2}$$

Aim to obtain 'best' estimate of the constant parameters $\boldsymbol{\beta}$, given the n correct variables, $\mathbf{x}$, 'independent' of $\{\epsilon_t\}$ and uncontaminated observations, $\mathcal{T}$, with $\epsilon_t \sim$IID$\left[0, \sigma_\epsilon^2\right]$.

Many tests to 'discover' departures from assumptions of (2), followed by recipes for 'fixing' them–
covert and unstructured empirical model discovery.

Model selection: discovering the 'best' model

Start from (2):

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \epsilon_t, \quad t = 1, \ldots, T$$

assuming N 'correct' initial $\mathbf{x}$ & $\mathcal{T}$, and valid conditioning where $\mathbf{x}$ includes many candidate regressors.

Aim to discover the subset of relevant variables, $\mathbf{x}_t^*$, then estimate the associated parameters, $\boldsymbol{\beta}^*$.

Departures from assumptions often simply ignored: 'information criteria' usually do not even check.

May select 'best model' in set, yet be a poor approximation to LDGP.

Robust statistics: discovering the best sample

Same start (2), but aim to find a **'robust' estimate** of β by selecting over $\mathcal{T}$.

Worry about data contamination and outliers, so select sample, $\mathcal{T}^*$, where outliers least in evidence, given correct set of relevant variables $\mathbf{x}$.

Other difficulties still need separate tests, and must be fixed if found.

$\mathbf{x}$ rarely selected jointly with $\mathcal{T}^*$, so assumes $\mathbf{x} = \mathbf{x}^*$.

Similarly for non-parametric methods:
aim to discover 'best' functional form & distribution assuming correct $\mathbf{x}$, no data contamination, constant β, etc., all rarely checked.

Automatic empirical model discovery

Re-frame empirical modelling as discovery process: part of a progressive research strategy.

Starting from T observations on N variables $\mathbf{x}$, aim to find $\boldsymbol{\beta}^*$ for s lagged functions $\mathbf{g}(\mathbf{x}_t^*) \dots \mathbf{g}(\mathbf{x}_{t-s}^*)$ of a subset of n variables $\mathbf{x}^*$, jointly with $\mathcal{T}^*$ and $\{d_t\}$– indicators for breaks, outliers etc.

Embeds initial $y = f(\mathbf{x})$, but much more general.

Globally, learning must be simple to general; but locally, need not be.

Implications for automatic methods

Same six stages as for discovery in general.

First, going **outside** current view by **automatic creation of a general model** from variables specified by investigator.

Second, **search** by **automatic selection** to find viable representations: too large for manual labor.

Third, criteria to **recognize** when search is completed: **congruent parsimonious-encompassing model**.

Fourth, **quantification** of the outcome: translated into **unbiasedly estimating the resulting model**.

Fifth, **evaluate** discovery to check its 'reality: **new data, new tests or new procedures**.
Can also evaluate the selection process itself.

Sixth, **summarize** vast information set in **parsimonious but undominated model**.

Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) 1-cut model selection

(4) Automatic estimation

(5) Automatic model selection

(6) Model evaluation

(7) Excess numbers of variables $N > T$

(8) An empirical example: food expenditure

Conclusions

Extensions outside standard information

Extensions determine how well LDGP is approximated

Create three extensions automatically:

(i) lag formulation to implement **sequential factorization**;

(ii) functional form transformations for **non-linearity**;

(iii) impulse-indicator saturation (IIS) for **parameter non-constancy and data contamination**.

(i) Create s lags $\mathbf{x}_t \dots \mathbf{x}_{t-s}$ to formulate general linear model:

$$y_t = \beta_0 + \sum_{i=1}^{s} \lambda_i y_{t-i} + \sum_{i=1}^{r} \sum_{j=0}^{s} \beta_{i,j} x_{i,t-j} + \epsilon_t \tag{3}$$

$\mathbf{x}_t$ could also be modelled as a system:

$$\mathbf{x}_t = \boldsymbol{\gamma} + \sum_{j=1}^{s} \boldsymbol{\Gamma}_j \mathbf{x}_{t-j} + \boldsymbol{\epsilon}_t \tag{4}$$

Automatic non-linear extensions

Test for non-linearity in general linear model by low-dimensional portmanteau test in Castle and Hendry (2010) (cubics of **principal components** of the $\mathbf{x}_t$).

(ii) If reject, create $\mathbf{g}(\mathbf{x}_t)$, otherwise $\mathbf{g}(\mathbf{x}_t) = \mathbf{x}_t$: presently, implemented general cubics with exponential functions.

Number of potential regressors for cubic polynomials is:

$$M_K = K\,(K+1)\,(K+5)\,/6.$$

Explosion in number of terms as $K = r \times s$ increases:

K	1	2	3	4	5	10	15	20	30	40
M_K	3	9	19	30	55	285	679	1539	5455	12300

Quickly reach $M_K > T$: search must handle that case.
Most easily explained for IIS.
(Investigating **squashing functions**, to better approximate non-linearity in economics, suggested by Hal White)

Impulse-indicator saturation

(iii) To tackle multiple breaks & data contamination (outliers), add T impulse indicators to candidates for T observations.

Consider $x_i \sim$ IID $\left[\mu, \sigma_\epsilon^2\right]$ for $i = 1, \ldots, T$

μ is parameter of interest

Uncertain of outliers, so add T indicators $I_k = \mathbf{1}_{\{k=k_i\}}$ to set of candidate regressors.

First, include half of indicators, record significant:
just 'dummying out' $T/2$ observations for estimating μ
Then omit, include other half, record again.
Combine sub-sample indicators, & select significant.

αT indicators selected on average at significance level α

Feasible 'split-sample' impulse-indicator saturation (IIS) algorithm: see Hendry, Johansen and Santos (2008)

Dynamic generalizations

Johansen and Nielsen (2009) extend IIS to both stationary and unit-root autoregressions

When distribution is symmetric, adding T impulse-indicators to a regression with r variables, coefficient $\boldsymbol{\beta}$ (not selected) and second moment $\boldsymbol{\Sigma}$:

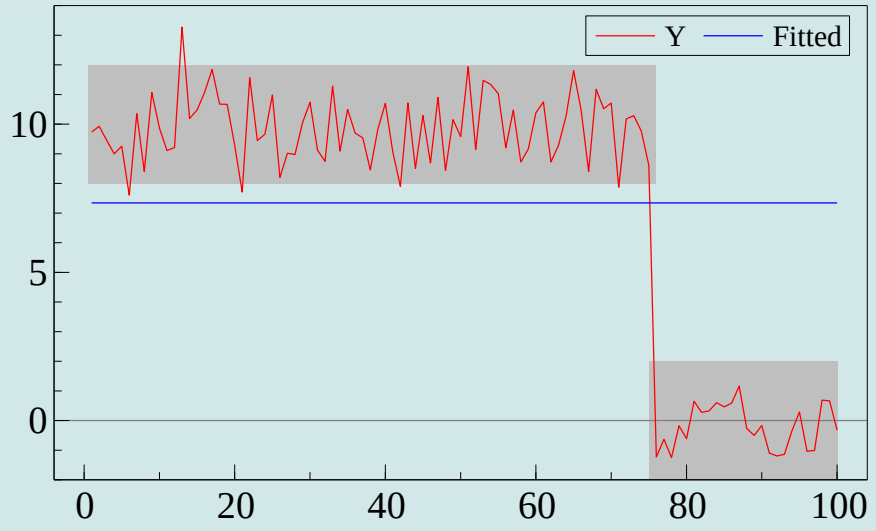
$$T^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_r\left[\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\Omega}_\beta\right]$$

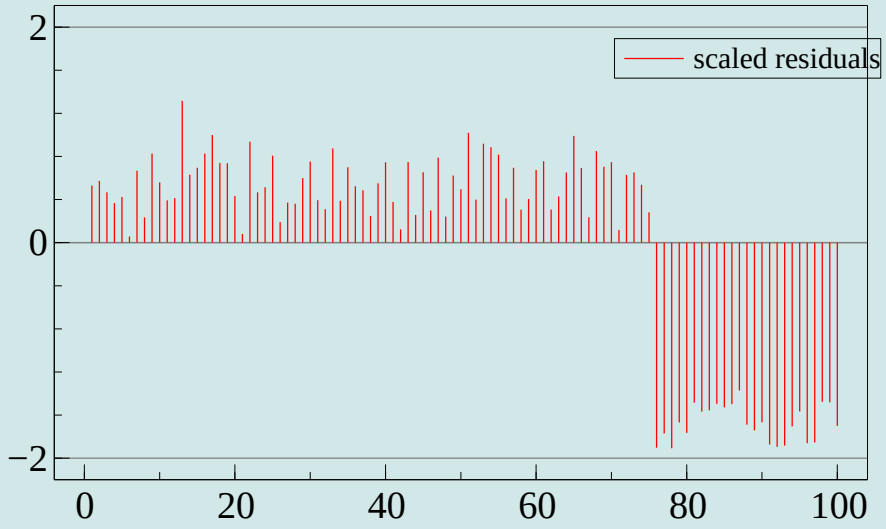
Efficiency of IIS estimator $\widetilde{\boldsymbol{\beta}}$ with respect to OLS $\widehat{\boldsymbol{\beta}}$ measured by $\boldsymbol{\Omega}_\beta$ depends on c_α and distribution

Must lose efficiency under null: but small loss αT– only 1% at $\alpha = 1/T$ if $T = 100$, despite T extra candidates.

Potential for major gain under alternatives of breaks and/or data contamination: variant of robust estimation

Structural break example

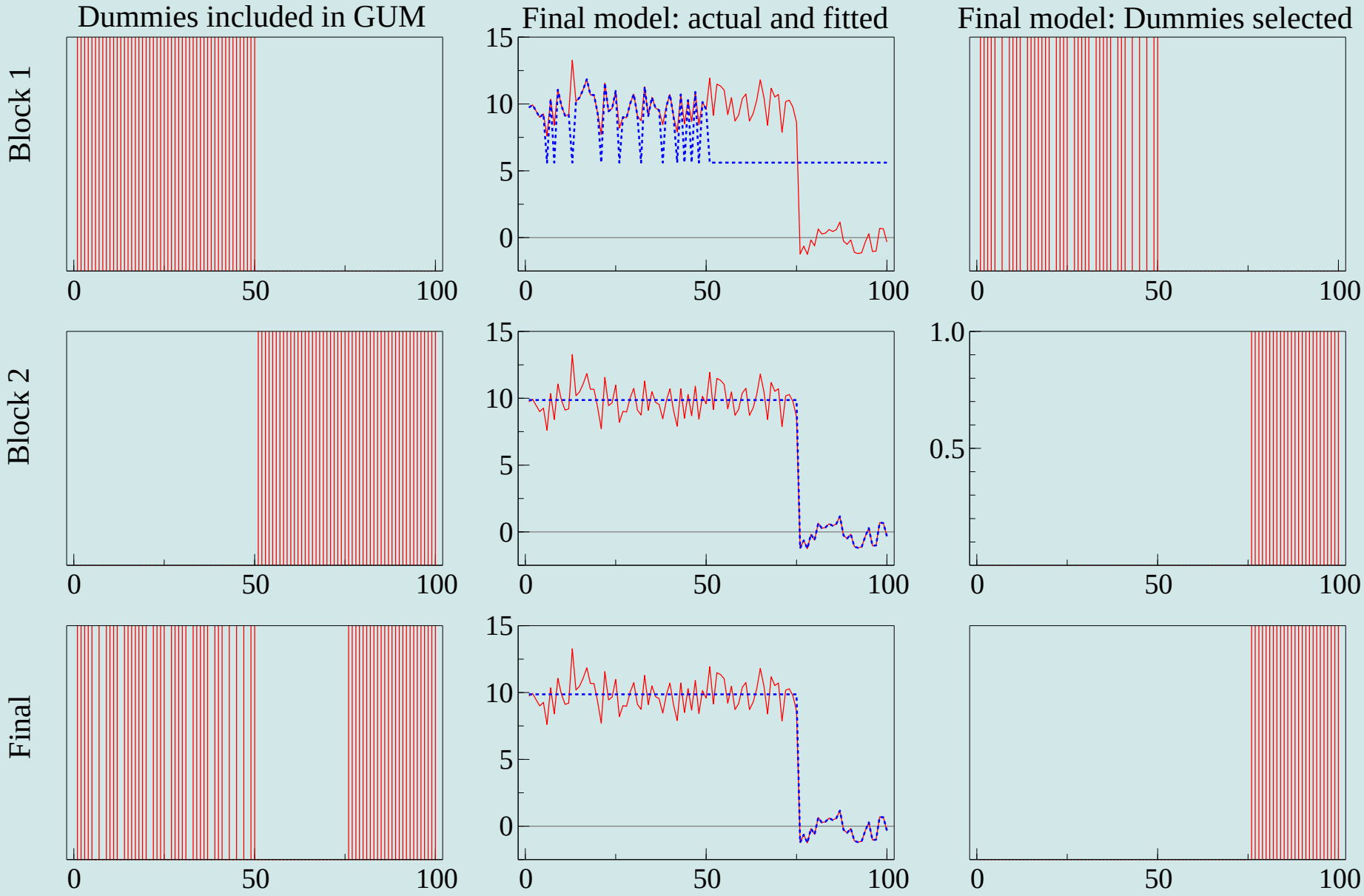




- Size of the break is **10 standard errors** at $0.75T$
- There are **no outliers** in this mis-specified model as all residuals $\in [-2, 2]$ SDs:
  outliers $\neq$ structural breaks
- step-wise regression has **zero power**

Let's see what Autometrics reports

‘Split-sample’ search in IIS



Specification of GUM

Most major formulation decisions now made:
which variables;
their lag lengths;
functional forms;
structural breaks.

Leads to general unrestricted model (GUM):

$$y_t = \sum_{i=1}^{K} \beta_i z_{i,t} + \sum_{i=1}^{K} \theta_i z_{i,t}^2 + \sum_{i=1}^{K} \gamma_i z_{i,t}^3 + \sum_{i=1}^{K} \sum_{j>i}^{K} \lambda_{i,j} z_{i,t} z_{j,t}$$

$$+ \sum_{i=1}^{K} \sum_{j>i}^{K} \sum_{k>j}^{K} \psi_{i,j,k} z_{i,t} z_{j,t} z_{k,t} + \sum_{i=1}^{T} \delta_i 1_{\{i=t\}} + \epsilon_t$$

$K = r \times s$ potential regressors, $\mathbf{z}_t$, after lags of $\mathbf{x}_t$, plus T indicators $1_{\{i=t\}}$

Bound to have $N > T$: consider exogeneity later.

Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) **1-cut model selection**

(4) Automatic estimation

(5) Automatic model selection

(6) Model evaluation

(7) Excess numbers of variables $N > T$

(8) An empirical example: food expenditure

Conclusions

Four main stages, then evaluation

1] **'1-cut' selection** for **orthogonal** designs with $N << T$.

2] Selection matters, so consider effects of **bias correction** on distributions of estimates

3] Compare '1-cut' with ***Autometrics***, which works in non-orthogonal models, still with $N << T$.

4] **More variables N than observations T**

Having resolved selection, next consider evaluation:

5] **Multiple breaks**, using IIS

6] Impact of **diagnostic testing**

7] Role of **encompassing** in automatic selection

8] Testing **exogeneity** and **invariance**

Understanding model selection

Consider a perfectly orthogonal regression model:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \tag{5}$$

$\mathsf{E}[z_{i,t} z_{j,t}] = \lambda_{i,i}$ for $i = j$ & $0 \; \forall i \neq j$, $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$ and $T >> N$.

Order the N sample t^2-statistics testing H_0: $\beta_j = 0$:

$$\mathsf{t}^2_{(N)} \geq \mathsf{t}^2_{(N-1)} \geq \cdots \geq \mathsf{t}^2_{(1)}$$

Cut-off m between included and excluded variables is:

$$\mathsf{t}^2_{(m)} \geq c^2_\alpha > \mathsf{t}^2_{(m-1)}$$

Larger values retained: all others eliminated.

Only one decision needed even for $N \geq 1000$:
'repeated testing' does not occur, and
'goodness of fit' is never considered.

Maintain average false null retention at **one variable** by $\alpha \leq 1/N$, with α declining as $T \to \infty$

Interpretation

Path search gives impression of 'repeated testing'.
Confused with selecting from 2^N possible **models**
(here $2^{1000} = 10^{301}$, an impossible task).
We are selecting **variables**, not models, & only N variables.

But selection matters, as only retain 'significant' outcomes.
Sampling variation also entails retain irrelevant, or miss relevant, by chance near selection margin.

Conditional on selecting, estimates biased away from origin: **but can bias correct as know** c_α.

Small efficiency cost under null for examining many candidate regressors, even $N >> T$**.**

Almost as good as commencing from LDGP at same c_α.

Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) **1-cut model selection**

(4) **Automatic estimation**

(5) Automatic model selection

(6) Model evaluation

(7) Excess numbers of variables $N > T$

(8) An empirical example: food expenditure

Conclusions

Simulation MSEs

Table 1 shows impact of bias corrections on retained irrelevant and relevant variables.

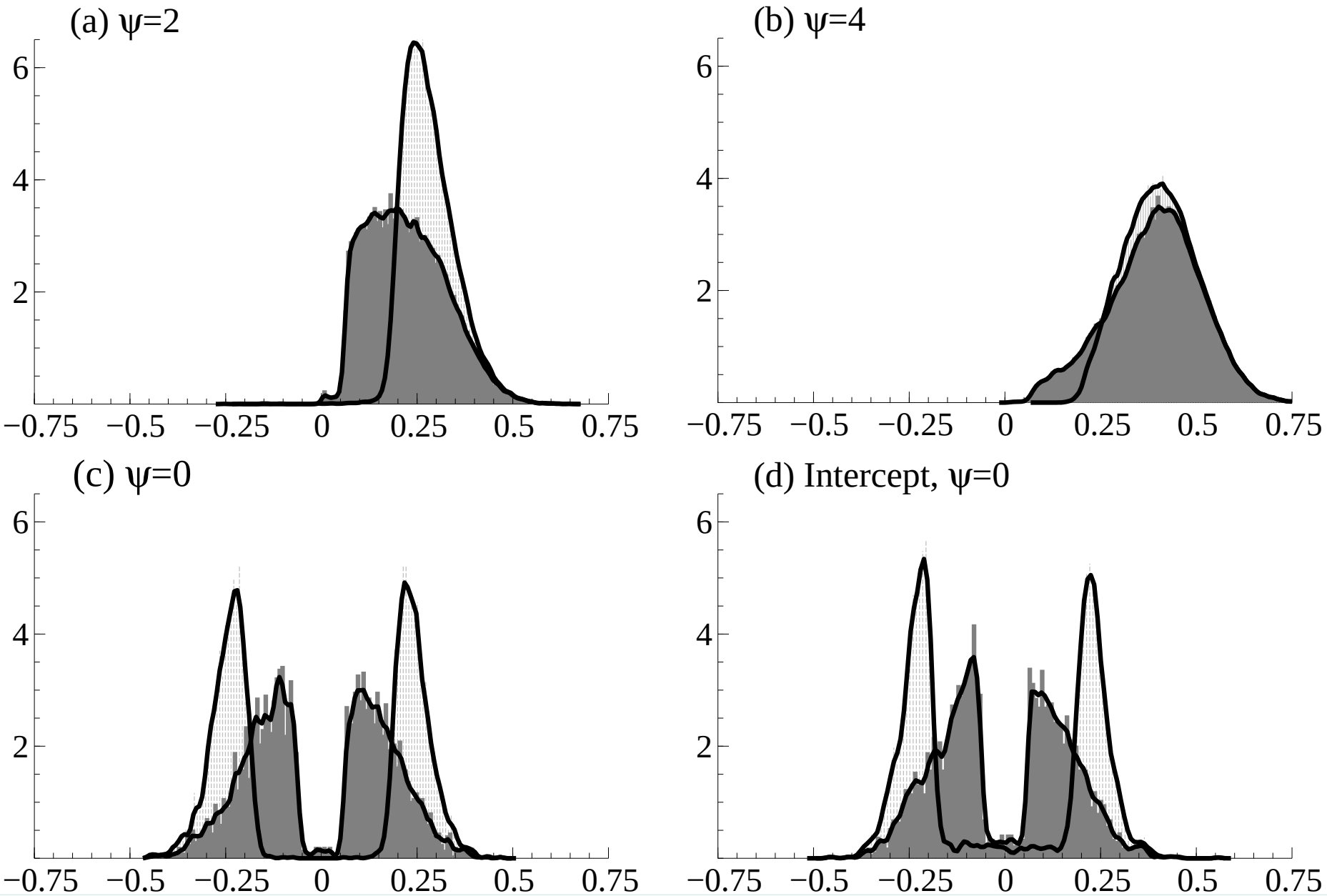
α	1%	0.1%	1%	0.1%
	average CMSE over 990 irrelevant variables		average CMSE over 10 relevant variables	
uncorrected $\widetilde{\beta}$	0.84	1.23	1.0	1.4
$\overline{\overline{\beta}}$ after correction	0.38	0.60	1.2	1.3

Table 1: Average CMSEs, times 100, for $N = 1000$ and $n = 10$ of retained relevant and irrelevant variables (excluding β_0), with and without bias correction.

Greatly reduces MSEs of irrelevant variables in both unconditional and conditional distributions.

Coefficients of retained variables with $|\mathrm{t}| \leq c_\alpha$ are not bias corrected–insignificant estimates set to zero.

Bias correcting conditional distributions



Implications of selection

Despite selecting from $N = 1000$ potential variables when only $n = 10$ are relevant:

(1) nearly unbiased estimates of coefficients and equation standard errors can be obtained;

(2) little loss of efficiency from checking many irrelevant variables;

(3) some loss from not retaining relevant variables at large c_α;

(4) huge gain by not commencing from an under-specified model;

(5) even works well for 'fat-tailed' errors at tight α when IIS used.

Now add path search for non-orthogonal data sets.

Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) **1-cut model selection**

(4) **Automatic estimation**

(5) **Automatic model selection**

(6) Model evaluation

(7) Excess numbers of variables $N > T$

(8) An empirical example: food expenditure

Conclusions

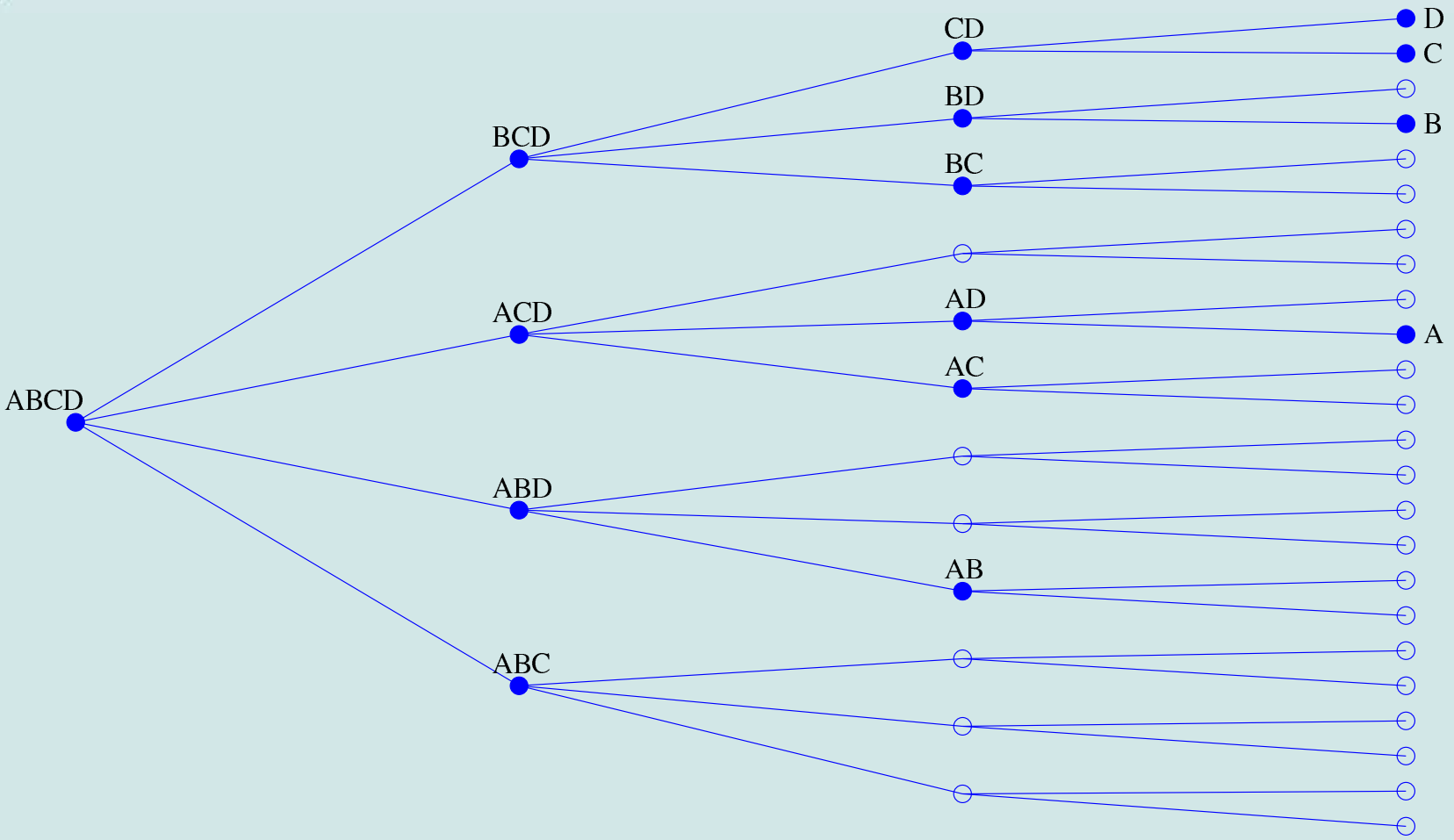
Autometrics improves on previous algorithms

- **Search paths:** Autometrics examines whole search space; discards irrelevant routes systematically.
- **Likelihood-based:** Autometrics implemented in likelihood framework.
- **Efficiency:** Autometrics improves computational efficiency: avoids repeated estimation & diagnostic testing, remembers terminal models.
- **Structured:** Autometrics separates estimation criterion, search algorithm, evaluation, & termination decision.
- **Generality:** Autometrics can handle $N > T$.

If GUM is congruent, so are all terminals: undominated, mutually-encompassing representations.

If several terminal models, all reported: can combine, or one selected (by, e.g., Schwarz, 1978, criterion).

Autometrics *tree search*



Search follows branches till no insignificant variables; tests for congruence and parsimonious encompassing; backtracks if either fails, till first non-rejection found.

Selecting by Autometrics

Even when 1-cut applicable, little loss, and often a gain, from using path-search algorithm *Autometrics*.

***Autometrics* applicable to non-orthogonal problems, and $N > T$.**

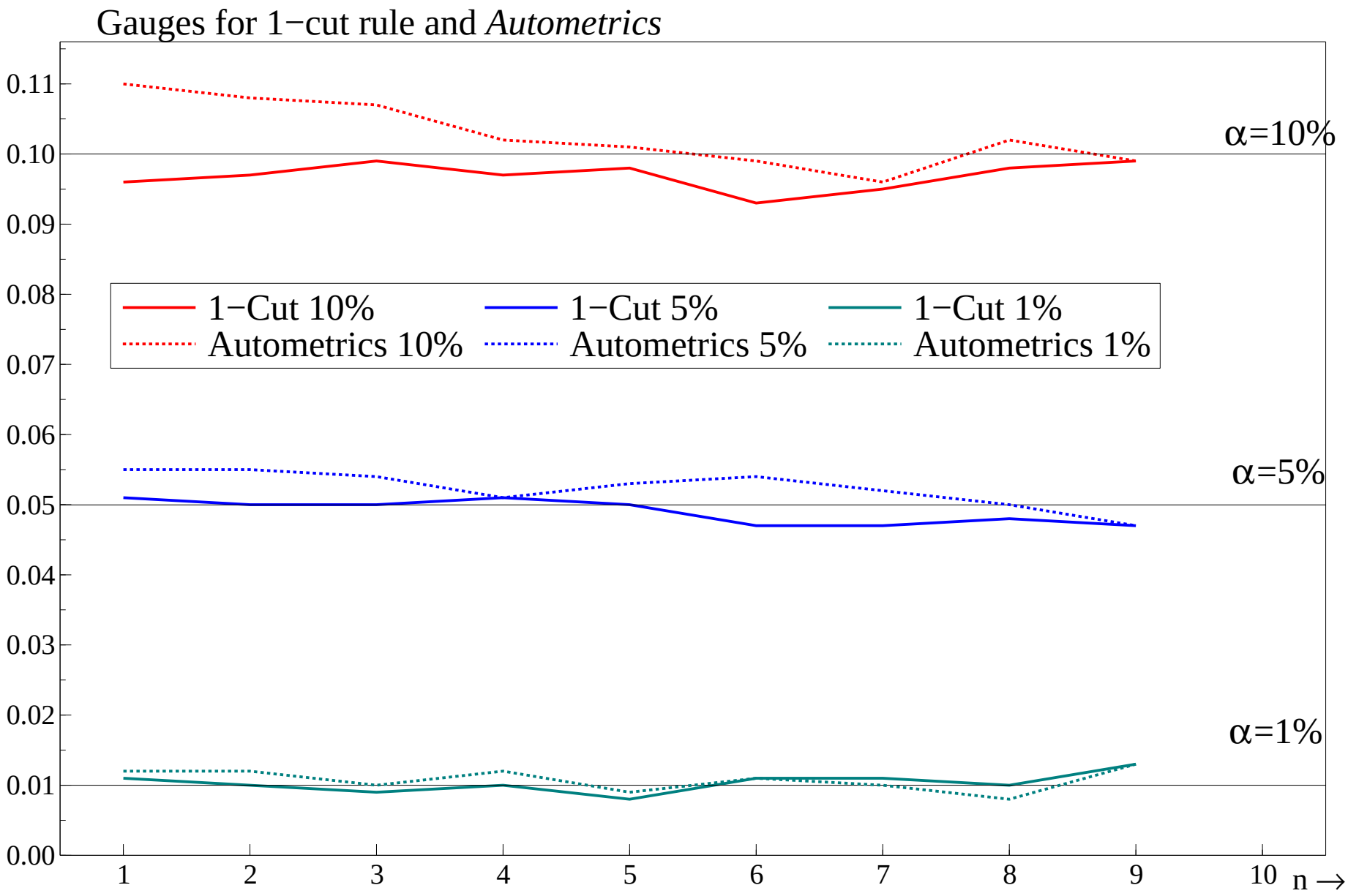
'*Gauge*' (average retention rate of irrelevant variables) close to α.

'*Potency*' (average retention rate of relevant variables) near theory value for a 1-off test.

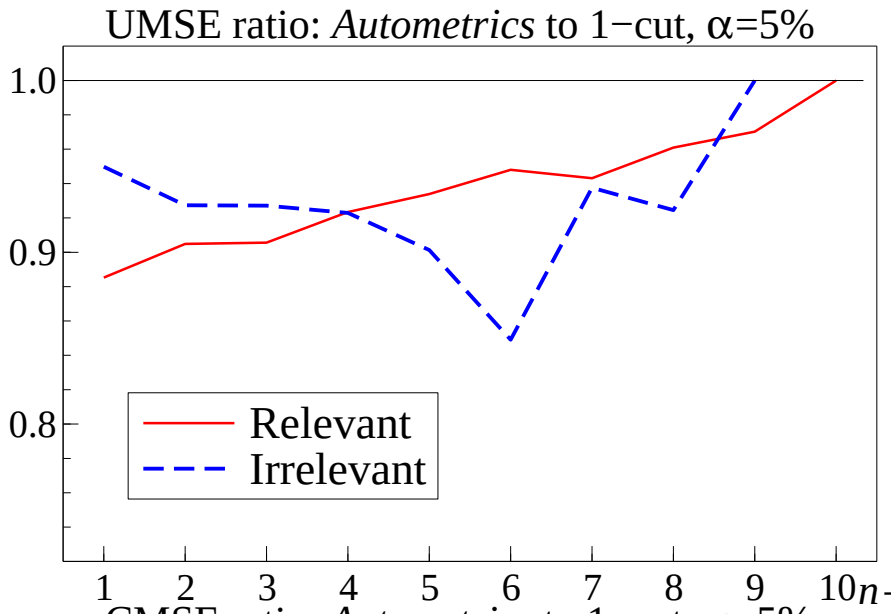
Goodness-of-fit not directly used to select models & no attempt to 'prove' that a given set of variables matters, but choice of c_α affects R^2 and n through retention by $|t_{(n)}| \geq c_\alpha$.

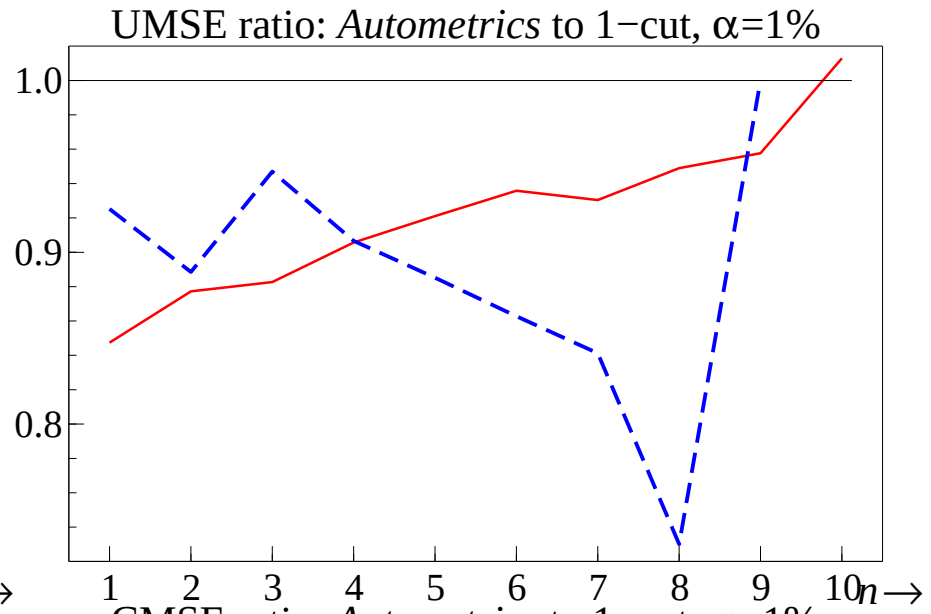
Conclude: 'repeated testing' is not a concern.

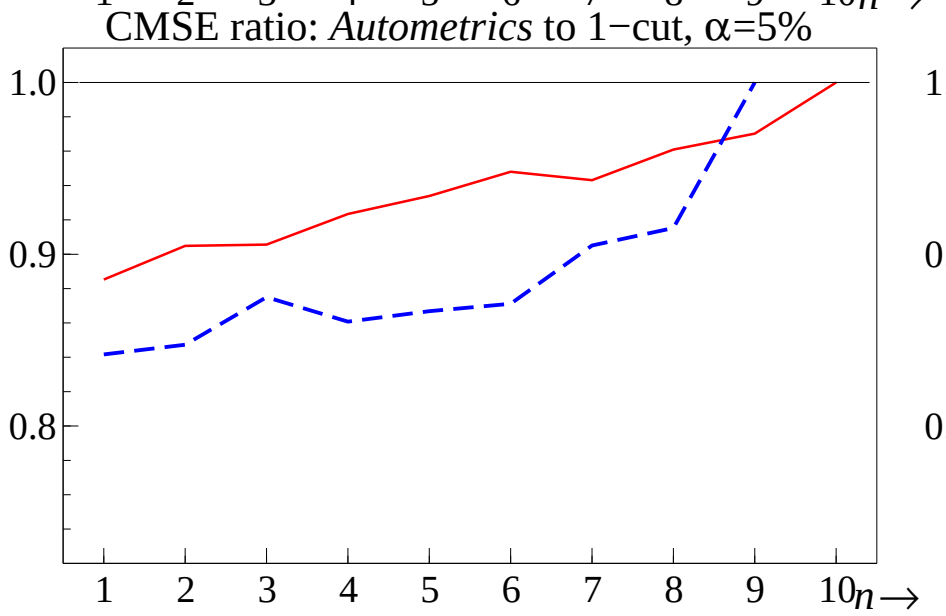
Gauges for 1-cut & Autometrics

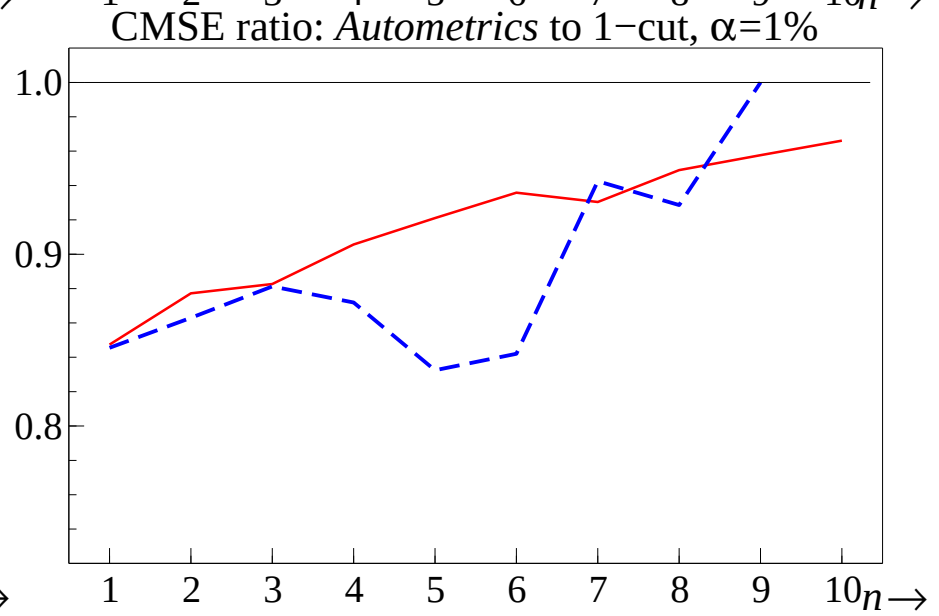


Ratios of MSE*s for Autometrics to 1-cut*









Retaining economic theory insights

Must stress that approach is not atheoretic.

Theory formulations should be embedded in GUM, can be retained without selection.

Call such imposition 'forcing' variables–ensures they are retained, but does not guarantee they will be significant.

Can also ensure theory-derived **signs** of long-run relation maintained, if not significantly rejected by the evidence.

But much observed data variability in economics is due to features absent from most economic theories: which empirical models must handle.

Extension of LDGP candidates, $\mathbf{x}_t$, in GUM allows theory formulation as special case, yet protects against contaminating influences (like outliers) absent from theory.

'Extras' can be selected at tight significance levels.

Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) **1-cut model selection**

(4) **Automatic estimation**

(5) **Automatic model selection**

(6) **Model evaluation**

(7) Excess numbers of variables $N > T$

(8) An empirical example: food expenditure

Conclusions

Role of encompassing

Variables removed only when new model is a valid reduction of GUM.

Reduction fails if result does not parsimoniously encompass GUM at c_α: (see Hendry, 1995, §14.6).

If so, variable retained despite being insignificant on t-test, as in Doornik (2008).

***Autometrics* without encompassing loses both gauge and potency**

***Autometrics* with encompassing is well behaved:**

Gauge is close to nominal rejection frequency α.
Potency is close to theory maximum.

Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) **1-cut model selection**

(4) **Automatic estimation**

(5) **Automatic model selection**

(6) **Model evaluation**

(7) **Excess numbers of variables** $N > T$

(8) An empirical example: food expenditure

Conclusions

Simulating Autometrics

Hoover and Perez (1999) experiments:

HP7 $y_{7,t} = 0.75y_{7,t-1} + 1.33x_{11,t} - 0.9975x_{11,t-1} + 6.44u_t$ $R^2 = 0.58$

HP8 $y_{8,t} = 0.75y_{8,t-1} - 0.046x_{3,t} + 0.0345x_{3,t-1} + 0.073u_t$ $R^2 = 0.93$

where $u_t \sim \mathsf{IN}[0, 1]$; $x_{i,t-j}$ are US macro data

The GUM has 3 DGP variables plus 37 irrelevant.

Also consider 141 irrelevant, larger than $T = 139$.

Hoover–Perez experiments with $\alpha = 0.01$

$T = 139$, **3 relevant and 37 irrelevant variables**

	Hoover–Perez		step-wise		Autometrics	
	HP7	HP8	HP7	HP8	HP7	HP8
	1% nominal size					
Gauge %	3.0*	0.9*	0.9	3.1	1.6	1.6
Potency %	94.0	99.9	100.0	**53.3**	99.2	**100.0**
DGP found %	24.6	78.0	71.6	**22.0**	68.3	**68.8**

* Only counting significant terms (but tiebreaker is best-fitting model)

$T = 139$, **3 relevant and 141 irrelevant variables**

	step-wise		Autometrics	
	HP7	HP8	HP7	HP8
	1% nominal size			
Gauge %	0.9	1.7	1.4	1.1
Potency %	99.9	**50.9**	96.9	**100.0**
DGP found %	32.3	**10.0**	42.5	**47.4**

Almost no impact of 104 additional irrelevant variables

Hoover–Perez experiments with $\alpha = 0.001$

$T = 139$, **3 relevant and 37 irrelevant variables**

	step-wise		Autometrics	
	HP7	HP8	HP7	HP8
	0.1% nominal size			
Gauge %	0.1	1.9	0.8	0.3
Potency %	99.9	**40.5**	98.3	**100.0**
DGP found %	95.1	**10.5**	87.5	**92.8**

$T = 139$, **3 relevant and 141 irrelevant variables**

	step-wise		Autometrics	
	HP7	HP8	HP7	HP8
	0.1% nominal size			
Gauge %	0.1	0.7	0.3	0.1
Potency %	99.7	**40.3**	97.4	**100.0**
DGP found %	87.4	**9.0**	82.9	**90.2**

Large increase in probability of locating DGP relative to $\alpha = 0.01$ not monotonic in α–so should not select by 'goodness of fit'

Testing super exogeneity

Parameter invariance essential in policy models:
else mis-predict under regime shifts.
Super exogeneity combines parameter invariance with valid conditioning so crucial for economic policy.

New automatic test:
impulse-indicator saturation in marginal models,
retain all significant outcomes and
test their relevance in conditional model

No *ex ante* knowledge of timing or magnitudes of breaks:
need not know DGP of marginal variables

Test has correct size under null of super exogeneity
for a range of sizes of marginal-model saturation tests

Power to detect failures of super exogeneity when location shifts in marginal models

Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) **1-cut model selection**

(4) **Automatic estimation**

(5) **Automatic model selection**

(6) **Model evaluation**

(7) **Excess numbers of variables** $N > T$

(8) **An empirical example: food expenditure**

Conclusions

Modelling expenditure on food

Many correct decisions needed for successful modelling:

expenditure depends on **many** relevant variables: incomes, prices, interest rates, taxes, demography, etc.

All effects could vary with changes in 'outside factors': legislation, policy regimes, financial innovation, etc.

Dependence could be linear or non-linear

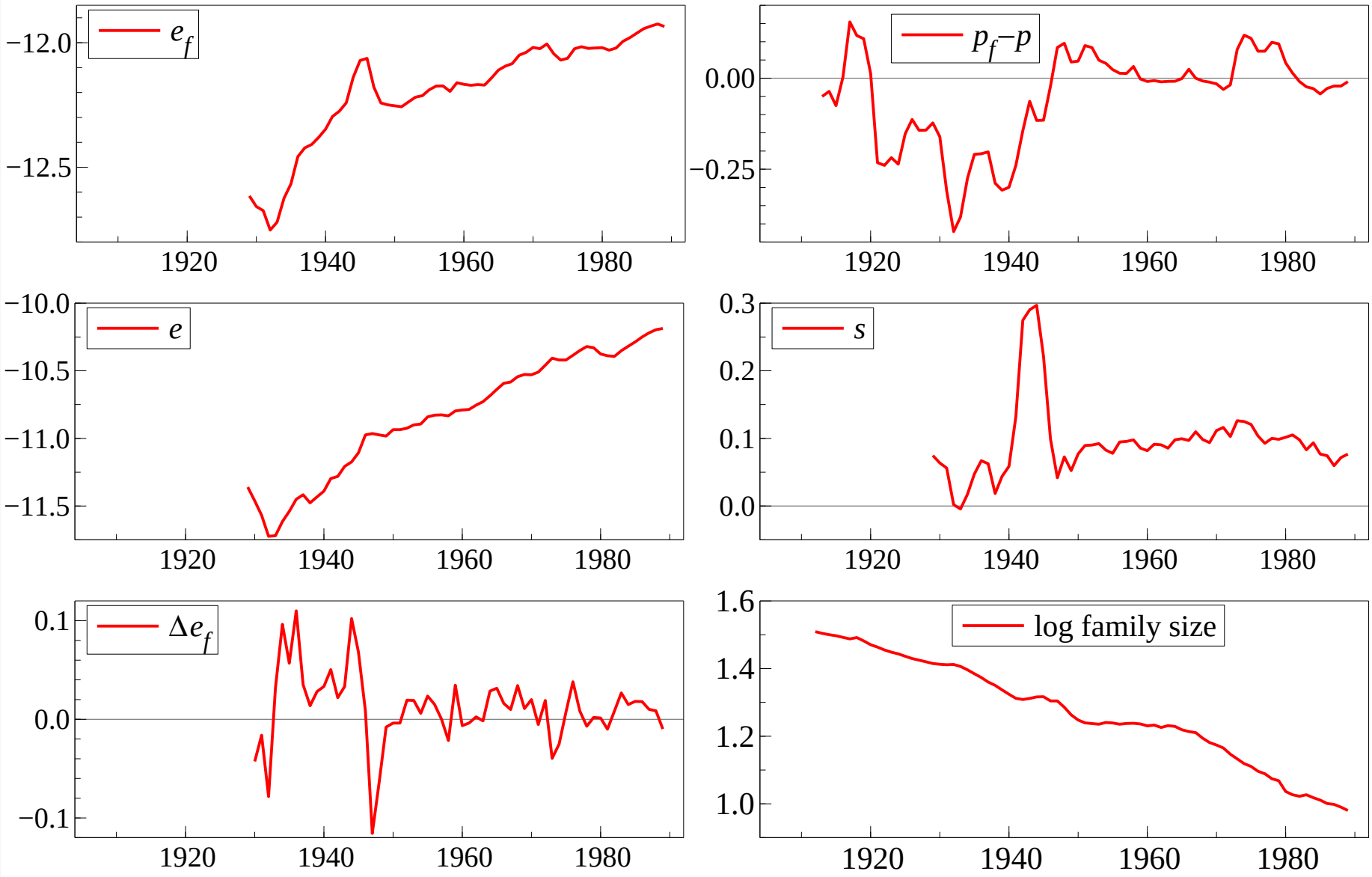
Short-run, long-run and seasonal responses may differ

Relationship may evolve over time

Level of aggregation matters: national or regional, by income levels, categories of transactions, etc.

Non-stationarities entail that any mis-specifications have deleterious effects

US real food expenditure and price data



Modelling problems

'Econometric Experiment' in Magnus & Morgan (1999)
Extension of Tobin (1950): data over 1929–1989
Per capita constant price expenditure on food, e_f, related to: constant price total final expenditure, e; real food prices, p_f; savings rate, s; family size, a; & previous values

Most participants abandoned interwar period:
perturbed by Great Depression and Food Relief.

When model reformulated to explain changes:
excellent properties – equation standard error = 0.75%,
no significant diagnostics, yet **fitted to whole sample**

Autometrics equation is similar to Hendry (1999)

Model derived in a fraction of time it took earlier:
invaluable for labour saving.

Can even 'forecast' post-war from 1952 on.

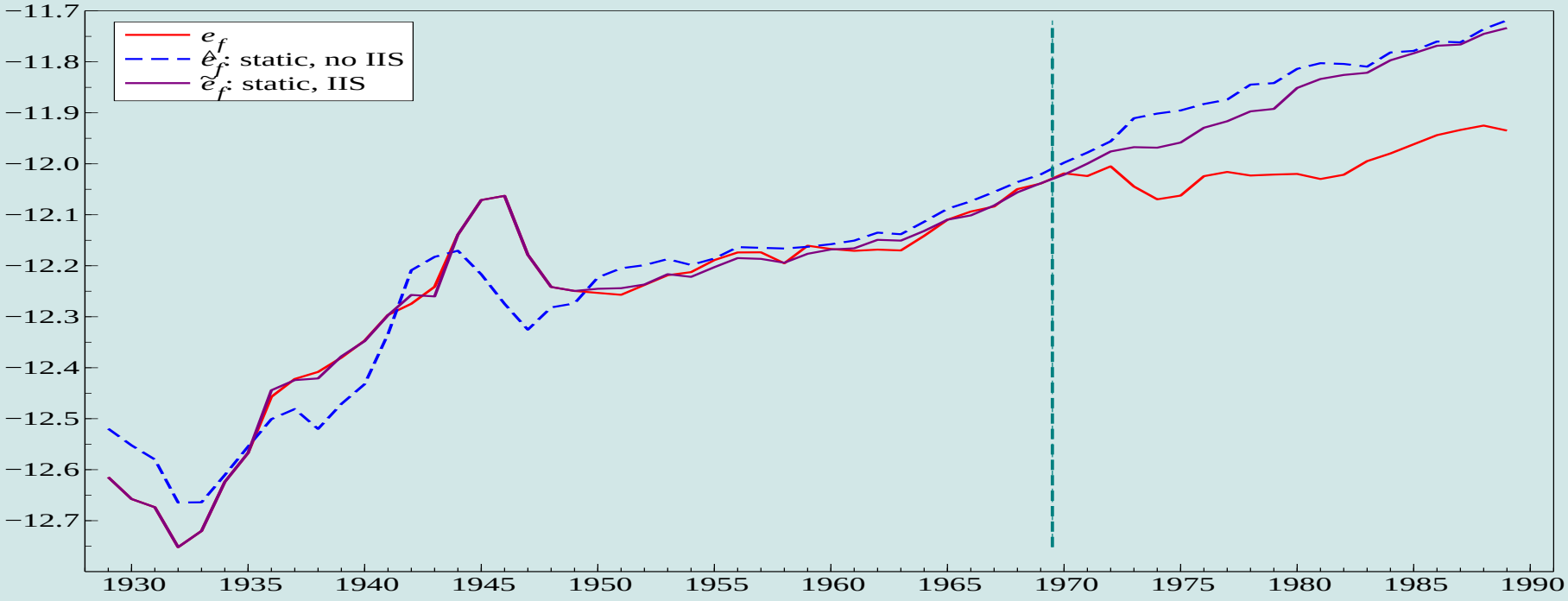
Enforcing theory

$$\underset{}{e_{f,t}} = \underset{(0.71)}{-\ 7.42} + \underset{(0.09)}{0.45}\ e_t - \underset{(0.13)}{0.034}\ p_{f,t} + \underset{(0.16)}{0.86}\ s_t + \underset{(0.27)}{0.073}\ a_t$$

$$\mathsf{R}^2 = 0.91 \quad \widehat{\sigma} = 0.066 \quad \mathsf{F}_{\mathsf{M}}(4,56) = 137.7^{**} \quad \mathsf{F}_{\mathsf{ar}}(2,54) = 55.6^{**}$$

$$\chi^2(2) = 10.4^{**} \quad \mathsf{F}_{\mathsf{arch}}(1,59) = 73.65^{**} \quad \mathsf{F}_{\mathsf{reset}}(2,54) = 14.2^{**}$$

$$\mathsf{F}_{\mathsf{het}}(8,52) = 10.5^{**} \quad \mathsf{F}_{\mathsf{Chow}}(20,36) = 0.58$$



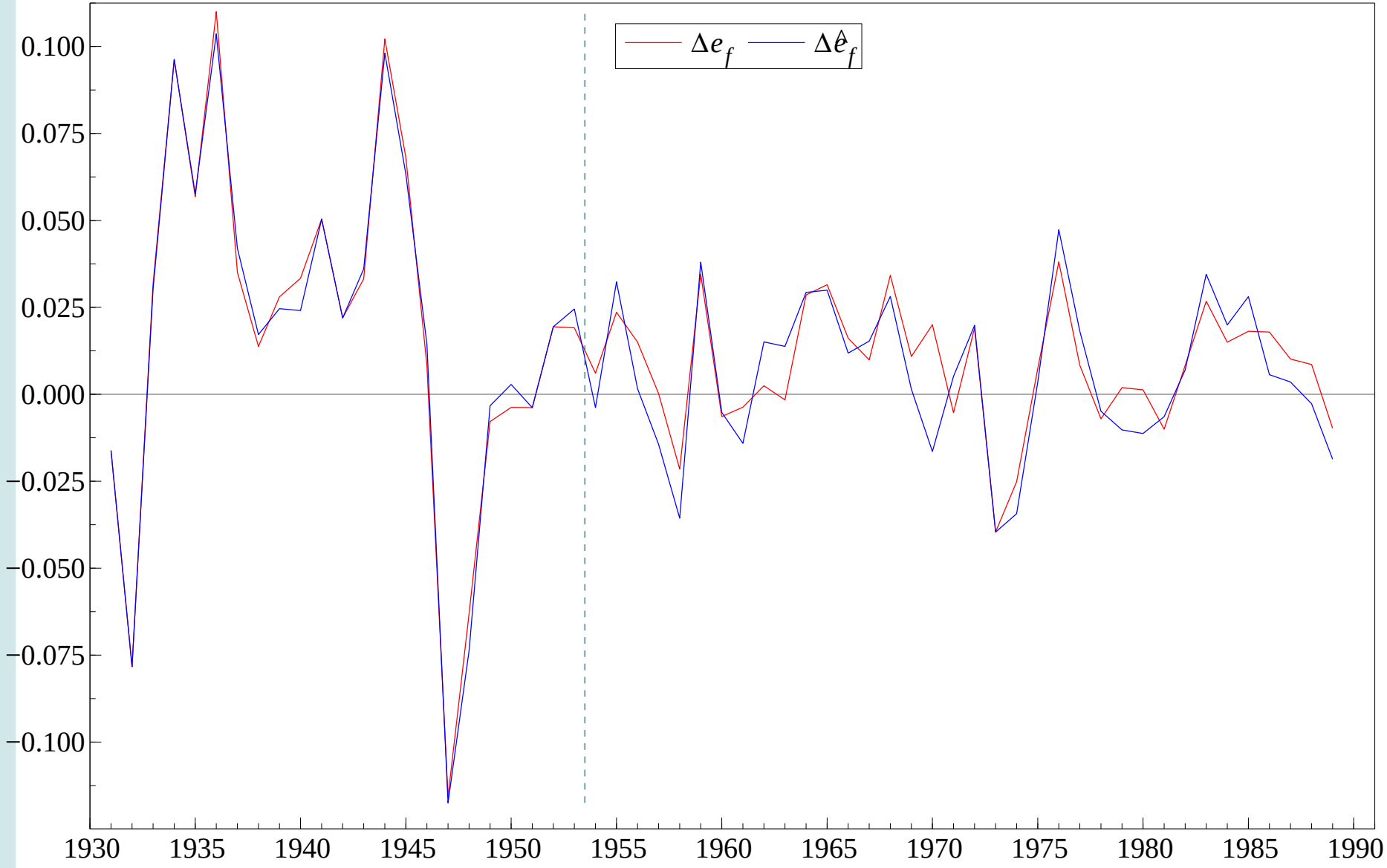
Selecting

IIS removes 1929–1935 & 1944–1949 but now does reject constancy with $\mathsf{F}_{\mathsf{Chow}}(20,23) = 3.89^{**}$ whereas:

$$\begin{aligned}
\Delta e_{f,t} = {} & \underset{(0.02)}{0.33}\, s_{t-1} - \underset{(0.02)}{0.32}\, c_{0,t-1} + \underset{(0.05)}{0.77}\, \Delta e_t + \underset{(0.03)}{0.11}\, \Delta e_{t-1} \\
& - \underset{(0.04)}{0.69}\, \Delta(p_f - p)_t - \underset{(0.01)}{0.09}\, \mathsf{I}_{31} - \underset{(0.01)}{0.10}\, \mathsf{I}_{32} + \underset{(0.01)}{0.03}\, \mathsf{I}_{34} \\
& + \underset{(0.01)}{0.03}\, \mathsf{I}_{41} + \underset{(0.01)}{0.06}\, \mathsf{I}_{42} + \underset{(0.01)}{0.04}\, \mathsf{I}_{51} + \underset{(0.01)}{0.02}\, \mathsf{I}_{52} \\
(\mathsf{R}^*)^2 = {} & 0.99 \;\; \widehat{\sigma} = 0.0067 \;\; \mathsf{F}_{\mathsf{M}}(12,10) = 108^{**} \;\; \mathsf{F}_{\mathsf{ar}}(1,9) = 0.09 \\
\chi^2(2) = {} & 0.72 \;\; \mathsf{F}_{\mathsf{arch}}(1,21) = 0.07 \;\; \mathsf{F}_{\mathsf{reset}}(2,8) = 2.39 \\
& \mathsf{F}_{\mathsf{Chow}}(36,10) = 1.82
\end{aligned}$$

$$c_0 = e_f + 7.99 - 0.4e + 0.36(p_f - p) \qquad (6)$$

‘Forecasting’ the post-war period



Route map

(1) **Discovery**

(2) **Automatic model extension**

(3) **1-cut model selection**

(4) **Automatic estimation**

(5) **Automatic model selection**

(6) **Model evaluation**

(7) **Excess numbers of variables** $N > T$

(8) **An empirical example: food expenditure**

Conclusions

Conclusions

All essential steps feasible once target LDGP defined:

1. automatically create general model from investigator's $\mathbf{x}_t$: extra variables, lags, non-linearity, & impulse indicators;
2. select congruent, parsimonious encompassing model;
3. compute near-unbiased parameter estimates; and
4. stringently evaluate results.

'1-cut' selection involves **no repeated testing**, yet *Autometrics* outperforms even in orthogonal case.

Generalizes to $N > T$ with expanding and contracting searches: see HP8 when $N = 145$, $T = 139$ at $\alpha = 0.001$.

Little difficulty in eliminating almost all irrelevant variables from the GUM (a small cost of search).

Avoids huge costs from under-specified models.

Conclusions II

Autometrics' gauge close to α;
increased slightly by diagnostic testing for congruence;
stabilized by encompassing tests against the GUM.

Potency is near theory power for a 1-off test.

Bias corrections reduce MSEs of retained irrelevant variables in both unconditional & conditional distributions.

Autometrics with IIS performs well: small cost for checking data contamination and multiple breaks.

Test of super exogeneity based on IIS in marginals:
F-test on significant indicators added to conditional model.
Applied to NKPC shows lack of invariance, insignificance of feed-forward term

Application to US food demand shows viable approach.

Overall conclusions

When the LDGP would be retained by *Autometrics* if commencing from it, then a close approximation is generally selected when starting from a GUM which nests that LDGP.

Non-linearities can be detected, modelled, and simplified using encompassing tests.

Theory formulations can be embedded in the GUM, to be retained without selection.

Model selection by *Autometrics* with tight significance levels and bias correction is a successful approach which allows multiple breaks to be tackled.

All the ingredients for empirical model discovery are in place.

The way ahead

Host of developments in automatic empirical model discovery already achieved

Now implementing *automatic*:

modelling of simultaneous systems

selecting cointegration vectors

testing expectations models for invariance

model averaging across terminals for forecasting.

The future is bright:

the future is *Autometrics*

References

Caceres, C. (2007). Asymptotic properties of tests for mis-specification. Economics, Oxford.

Castle, J. L., and Hendry, D. F. (2010). A low-dimension, portmanteau test for non-linearity. *JEcts*, DOI:10.1016/j.jeconom.2010.01.006.

Castle, J. L., and Shephard, N. (eds.)(2009). *Methodology and Practice of Econometrics*. OUP.

Davidson, J. E. H. (1998). Structural relations, cointegration & identification. *JEcts*, **87**, 87–113.

Doornik, J. A. (2008). Encompassing and automatic model selection. *OxBull*, **70**, 915–925.

——(2009). Autometrics. In Castle, and Shephard (2009), pp. 88–121.

Hendry, D. F. (1995). *Dynamic Econometrics*. OUP.

——(1999). An econometric analysis of US food expenditure, 1931–1989. In Magnus, J. R., and Morgan, M. S. (eds.), *Methodology and Tacit Knowledge*, pp. 341–361. Wiley.

Hendry, D. F., Johansen, S., and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *CompStats*, **33**, 317–335. Erratum, 337–339.

Hendry, D. F., and Krolzig, H.-M. (2005). Automatic Gets modelling. *EJ*, **115**, C32–C61.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered. *EctJ*, **2**, 167–191.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *JEDC*, **12**, 231–254.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated VARs*. OUP.

Johansen, S., Mosconi, R., and Nielsen, B. (2000). Cointegration analysis in the presence of structural breaks in the deterministic trend. *EctJ*, **3**, 216–249.

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator. In Castle, and Shephard (2009), pp. 1–36.

Leeb, H., and Pötscher, B. M. (2005). Model selection and inference. *EctTh*, **21**, 21–59.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica*, **58**, 113–144.

Tobin, J. (1950). A statistical demand function for food in the U.S.A.. *JRSS*, *A, 113*(2), 113–141.

Wooldridge, J. M. (1999). Asymptotic properties of some specification tests. In Engle, R. F., and White, H. (eds.), *Cointegration, Causality and Forecasting*, pp. 366–384. OUP.

Retracing route

(1) Discovery

(2) Automatic model extension

(3) 1-cut model selection

(4) Automatic estimation

(5) Automatic model selection

(6) Model evaluation

(7) Excess numbers of variables $N > T$

(8) An empirical example: food expenditure

Conclusion

Limitations

Limits of automatic empirical model discovery apply when:
(A) LDGP would not be reliably selected by the given inference rules applied to itself as the initial specification.

(B) Relevant variables have small t-statistics as parameters are $O(1/\sqrt{T})$, especially when highly correlated with other regressors (see Leeb and Pötscher, 2005).
Selection will not work well if a parameter cannot be consistently estimated.

(C) If LDGP is not nested in GUM, selected approximation must be incorrect:
but could be undominated, & LDGP might still be found by a progressive research strategy when intermittent breaks in both relevant and irrelevant variables.

(D) No approach yet for non-constant second moments, like GARCH errors.

Integrated data

Autometrics conducts inferences for I(0)
Most selection tests remain valid:
see Sims, Stock and Watson (1990)
Only tests for a unit root need non-standard critical values

Implemented PcGive cointegration test in *PcGets* 'Quick Modeler'

Most diagnostic tests also valid for integrated series:
see Wooldridge (1999)

Heteroscedasticity tests an exception:
powers of variables then behave oddly
see Caceres (2007)

Gauge and potency

Let $\widetilde{\beta}_i$ denote estimate of β_i after model selection.

'*Gauge*' is average retention frequency of irrelevant variables.

'*Potency*' is average retention frequency of relevant variables.

In simulation experiments with M replications:

$$\begin{aligned}
\text{retention rate: } \widetilde{\mathsf{p}}_k &= \tfrac{1}{M}\textstyle\sum_{i=1}^{M} 1_{(\widetilde{\beta}_{k,i}\neq 0)}, \qquad k=0,\ldots,N,\\
\text{potency:} &= \tfrac{1}{n}\textstyle\sum_{k=1}^{n} \widetilde{\mathsf{p}}_k,\\
\text{gauge:} &= \tfrac{1}{N-n+1}\left(\widetilde{\mathsf{p}}_0 + \textstyle\sum_{k=n+1}^{N} \widetilde{\mathsf{p}}_k\right).
\end{aligned}$$

Gauges not significantly different from nominal sizes α: selection is not 'over-sized' even with $N = 1000$ variables

Potencies close to average powers even when selecting just $n = 10$ relevant regressors from 1000 variables.

Simultaneous equations models

Linear simultaneous equations models are reductions of systems

Given endogenous $\mathbf{y}_t$ and exogenous variables $\mathbf{z}_t$, congruent linear conditional statistical system formulated:

$$\mathbf{y}_t = \mathbf{\Psi z}_t + \mathbf{v}_t \text{ where } \mathbf{v}_t \sim \mathsf{IN}_m[\mathbf{0}, \mathbf{\Omega}_v] \tag{7}$$

Always identified, so all later selections are also. Model is:

$$\mathbf{B y}_t = \mathbf{C z}_t + \boldsymbol{\epsilon}_t \text{ where } \boldsymbol{\epsilon}_t \sim \mathsf{IN}_m[\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\epsilon}}] \tag{8}$$

Rank condition $\mathbf{B\Psi} = \mathbf{C}$ imposed as constraint on searches

Endogenous variables added to every equation in (7) and reductions checked to eliminate exogenous regressors contingent on maintaining rank condition: (8) needs to be identified–but you do not need to know restrictions.

Cointegration

Use Johansen (1988, 1995) implemented in (9) for $\mathbf{x}_t$:

$$\Delta \mathbf{x}_t = \boldsymbol{\gamma} + \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{x}_{t-1} + \sum_{j=1}^{s-1} \boldsymbol{\Gamma}_j \Delta \mathbf{x}_{t-j} + \boldsymbol{\epsilon}_t \tag{9}$$

Use 'similar' form (restricted trend, unrestricted intercept)

Normalize on largest $\alpha_i \beta_j$: find minimal ('irreducible') cointegration vectors as in Davidson (1998).

Do not need to know identifying restrictions, as with SEMs.

Complications:

I(2) variables;

lag length needs joint selection;

outliers and breaks, so different critical values if IIS (e.g., Johansen, Mosconi and Nielsen, 2000).

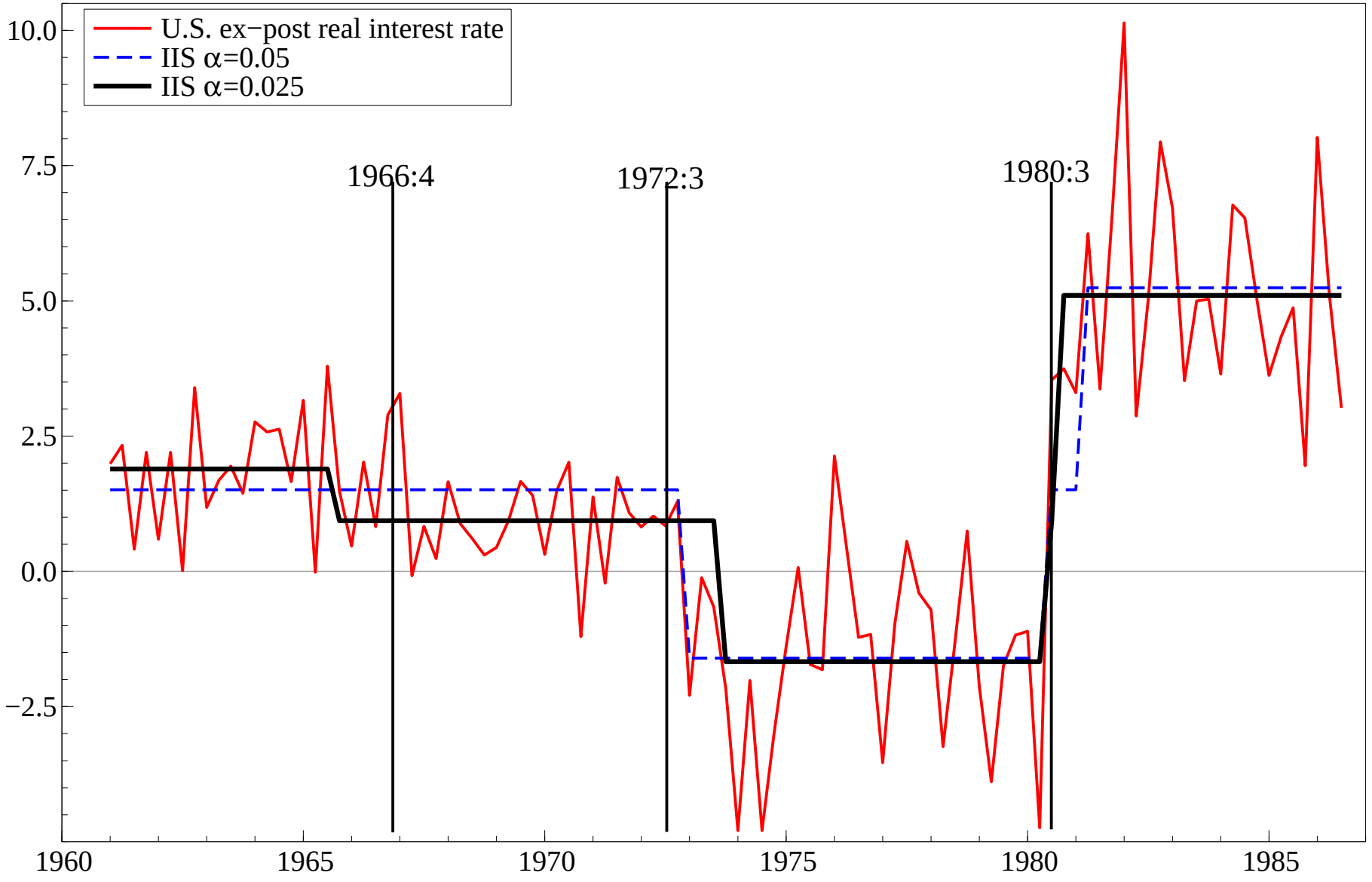
Order of proceeding unknown: we tackle cointegration reductions after finding congruent representation.

Super exogeneity in food expenditure

Build 'automatic' lagged equations with IIS for e; $p_f - p$; s; f.

Finds 25 new impulse-indicators for:
(e): 1946; ($p_f - p$): 1936, 1937, 1940, 1950, 1958, 1967, 1978; (s): 1943, 1968, 1984, 1987; (f): 1947, 1954, 1957, 1961, 1963; ($p_f - p$,s): 1944, 1945, 1973; (s,f): 1949; (e,f): 1980; (e,$p_f - p$,s): 1933, 1938.
Already had (impulses in common with model of e_f in **bold**): **1931** (e,$p_f - p$,s), **1932** (e,s), **1934** ($p_f - p$,s), **1941** (s), **1942** ($p_f - p$,s), **1951** ($p_f - p$), 1952, 1970.
Adding 25 to selected model of e_f yields test:
$\mathsf{F}_{22}^{24} = 1.63$ ($\mathsf{p} = 0.13$)
Does not formally reject, but common impulses might do.
However, no impulses in common in the post-1952 period.

Original BP and IIS on US real interest rates



Extended BP and IIS on US real interest rates