Multi-Path General-to-Specific Modelling with OxMetrics

Genaro Sucarrat

(Department of Economics, UC3M)

http://www.eco.uc3m.es/sucarrat/

1 April 2009

(Corrected for errata 22 November 2010)

Outline:

1. General-to-Specific (GETS) modelling:
   $\longrightarrow$ Motivation, properties, advantages, disadvantages

2. The basics of OxMetrics 5:
   $\longrightarrow$ Loading, editing and transforming data (logs, differencing), creating "special" series (cointegration relations, trends, etc.)

3. An overview of Autometrics:
   $\longrightarrow$ Key concepts and characteristics

4. Single-equation modelling with Autometrics
   $\longrightarrow$ Formulation, Advanced Autometrics settings, fixing variables, example (2007 Econometric Game, Q1)

5. Multiple-equation modelling with Autometrics
   $\longrightarrow$ Formulation, fixing variables, example (2007 Econometric Game, Q2)

Common (in-sample) modelling strategies:

1. Select model that minimises information criterion
2. Simple-to-general
3. "1-shot" General
4. Single-path GETS

• Multi-Path GETS: Combines 1 and 4 iteratively

• Multi-Path GETS algorithms: Hoover and Perez (1999), PcGets (Hendry and Krolzig 2001, 2005), Autometrics (Doornik and Hendry 2007a, Doornik 2009), AutoSEARCH (Sucarrat and Escribano 2009, Sucarrat 2009)

• Autometrics: A feature in OxMetrics that automates Multi-Path GETS

Autometrics automates GETS modelling of an OLS or IV estimable linear regression

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \epsilon$$

where the $\{\epsilon\}$ can be homoscedastic, heteroscedastic and/or autocorrelated

• NOTE: Only the case where $\{\epsilon\} \sim IIN(0, \sigma^2)$ has been extensively studied through Monte Carlo simulation (see in particular Hendry and Krolzig 2005, and Doornik 2009)

• Analytical analysis either not possible or yields limited insight

GETS modelling summarised:

1. Formulate a General Unrestricted Model (GUM)

2. Delete step-wise, along different paths, insignificant regressors at the chosen regressor significance level α ("target size", optional), while checking a range of (optional) diagnostics at each deletion using a different (optional) significance level

3. If simplification results in more than one terminal model, then select the model with lowest value on the chosen information criterion (default: Schwartz), or their union (optional)

Main benefits of GETS modelling:

- Estimation and inference is conducted while controlling for the influence of other variables
- In simulations multi-path GETS compares favourably to other (in-sample) modelling strategies
- GETS modelling results in a parsimonious model that is particularly useful for scenario analysis (conditional forecasting, policy analysis, counterfactual analysis, etc.)

Main disadvantages of GETS modelling:

- Slight tendency to retain irrelevant variables (the more correlated the regressors, the higher the tendency)
- Finite sample behaviour can depend substantially on the properties of the data (regressor inter-correlation, homoscedastic vs. heteroscedastic errors, fat-tailed errors, etc.)

Define k_0 as the number of relevant variables in GUM, k_1 as the number of irrelevant variables in GUM (and so $k_0 + k_1 = K$ total number of variables in the GUM):

- $\hat{k}_0/k_0$ is the relevance proportion or "potency" (analogous to "power" in statistical hypothesis testing)

- $\hat{k}_1/k_1$ is the irrelevance proportion or "gauge" (analogous to "size" in statistical hypothesis testing)

Main statistical properties of Autometrics (default options):

- $E(\hat{k}_0/k_0) \to 1$ as the sample size goes to ∞

- $E(\hat{k}_1/k_1) \to \alpha$ as the sample size goes to ∞

Target size:

• User defined regressor significance level α. For example, if 5% is chosen, then the insignificant regressors at 5% are deleted

Diagnostic test *p*-value:

• The acceptable diagnostic test significance level. For example, if deleting an insignificant variable results in a diagnostic test *p*-value above the acceptable level, then the variable is re-included into the model

Branch:

• Suppose we choose a regressor significance level of 5%, and consider the following GUM:

Regressor	Coef.	P-value
x_1	2.851	0.35
x_2	0.343	0.00
x_3	1.069	0.07

The GUM contains TWO insignificant variables (x_1 and x_3) $\Rightarrow$ TWO branches each made up of *paths*

Path:

• A deletion sequence. For example if x_1 is deleted first and then x_3 before no regressors are significant at the chosen regressor significance level, then $\{x_1, x_3\}$ is a deletion path or sequence

Rounds:

• If simplification results in more than one terminal model, then Autometrics initiates a second round by forming a new GUM made up of the union of the terminal models

• Specification search terminates when either only one terminal model results, or when the GUM at round n equals the GUM at round $n-1$. If this is the case, then a "Tiebreaker" (an information criterion) is used to select among the models

Backtesting:

• Parsimonious encompassing test. By default, this is a joint test of the final model against the initial GUM ("GUM 0"), that is, an F-test of whether the deleted regressors are jointly insignificant at α

OxMetrics basics:

• Load data: File → Open, etc.

• Edit sample/dates: Edit → Change Sample

• Missing values (my recommendation): Set to "missing" by double-clicking the data cell in question

• Graph series: Model → Graphics (or click on the graphics button) → Actual series or All plot types

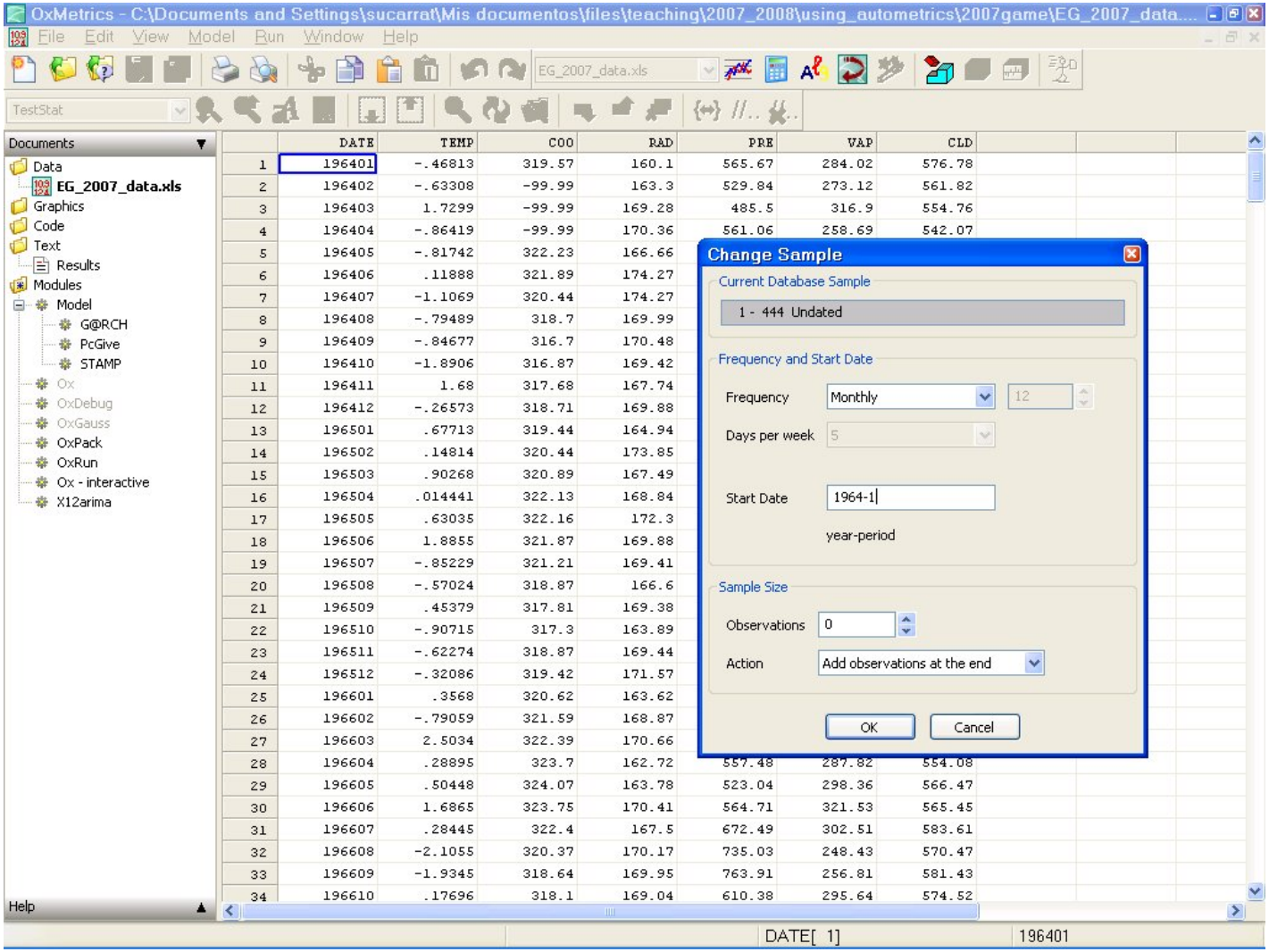
• Transform data (algebra feature):

Edit → Algebra (or click on the "Alg" button) → "Code", e.g. "DCOO = diff(COO,1);" → Run (→ Done)

(NOTE: Case sensitivity in variable names!)

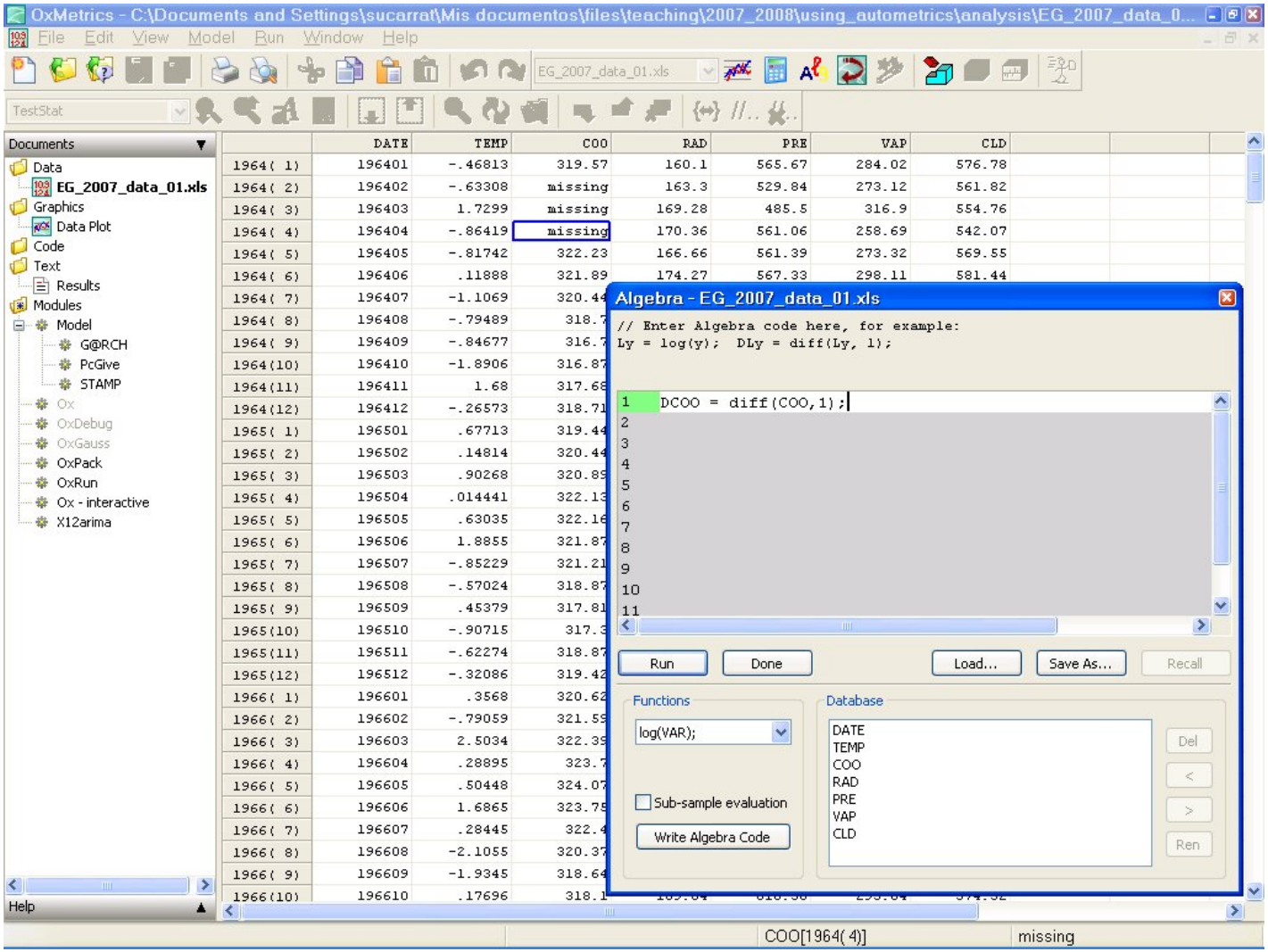
• Create special series (calculator feature): Model → Calculator (or click on the calculator button) → ...

- *Example*. Edit dates (2007 Econometric Game Case):



	DATE	TEMP	COO	RAD	PRE	VAP	CLD
1	196401	-.46813	319.57	160.1	565.67	284.02	576.78
2	196402	-.63308	-99.99	163.3	529.84	273.12	561.82
3	196403	1.7299	-99.99	169.28	485.5	316.9	554.76
4	196404	-.86419	-99.99	170.36	561.06	258.69	542.07
5	196405	-.81742	322.23	166.66			
6	196406	.11888	321.89	174.27			
7	196407	-1.1069	320.44	174.27			
8	196408	-.79489	318.7	169.99			
9	196409	-.84677	316.7	170.48			
10	196410	-1.8906	316.87	169.42			
11	196411	1.68	317.68	167.74			
12	196412	-.26573	318.71	169.88			
13	196501	.67713	319.44	164.94			
14	196502	.14814	320.44	173.85			
15	196503	.90268	320.89	167.49			
16	196504	.014441	322.13	168.84			
17	196505	.63035	322.16	172.3			
18	196506	1.8855	321.87	169.88			
19	196507	-.85229	321.21	169.41			
20	196508	-.57024	318.87	166.6			
21	196509	.45379	317.81	169.38			
22	196510	-.90715	317.3	163.89			
23	196511	-.62274	318.87	169.44			
24	196512	-.32086	319.42	171.57			
25	196601	.3568	320.62	163.62			
26	196602	-.79059	321.59	168.87			
27	196603	2.5034	322.39	170.66			
28	196604	.28895	323.7	162.72	557.48	287.82	554.08
29	196605	.50448	324.07	163.78	523.04	298.36	566.47
30	196606	1.6865	323.75	170.41	564.71	321.53	565.45
31	196607	.28445	322.4	167.5	672.49	302.51	583.61
32	196608	-2.1055	320.37	170.17	735.03	248.43	570.47
33	196609	-1.9345	318.64	169.95	763.91	256.81	581.43
34	196610	.17696	318.1	169.04	610.38	295.64	574.52

- *Example*. Create a differenced series:



Autometrics is a multi-path GETS modelling feature in OxMetrics:

• The objective of Autometrics is to automate Multi-Path GETS specification search of a *data coherent*, *General Unrestricted Model* (GUM) in the form of a linear OLS/IV estimable regression (or regressions)

• Default definition of data-coherency: Stable parameters and Gaussian, serially uncorrelated, homoscedastic errors. NOTE: These assumptions can be relaxed through the "Advanced Autometrics settings", and if the GUM fails one or several diagnostic checks Autometrics proceeds anyway

• GUM: A general model (advice: Not too general!) that includes the variables and lags that are believed to possibly have an impact

• Further reading: Doornik and Hendry (2007a, pp. 70-77), Hendry and Krolzig (2001) (Autometrics is an evolution of PcGets)

Single-equation estimation. *Example*: 2007 Econometric Game, Question 1

- A "rough" GUM:

$$\Delta COO_t = b_0 + b_1 \Delta COO_{t-1} + b_2 \Delta COO_{t-2} + \sum_{j=1}^{11} c_j d_{j,t} + e_t \quad (1)$$

- Formulating a model: (Model →) PcGive → Category: "Models for time series data" → Model class: "Single-equation dynamic modelling using PcGive" (→ Options) → "Formulate"

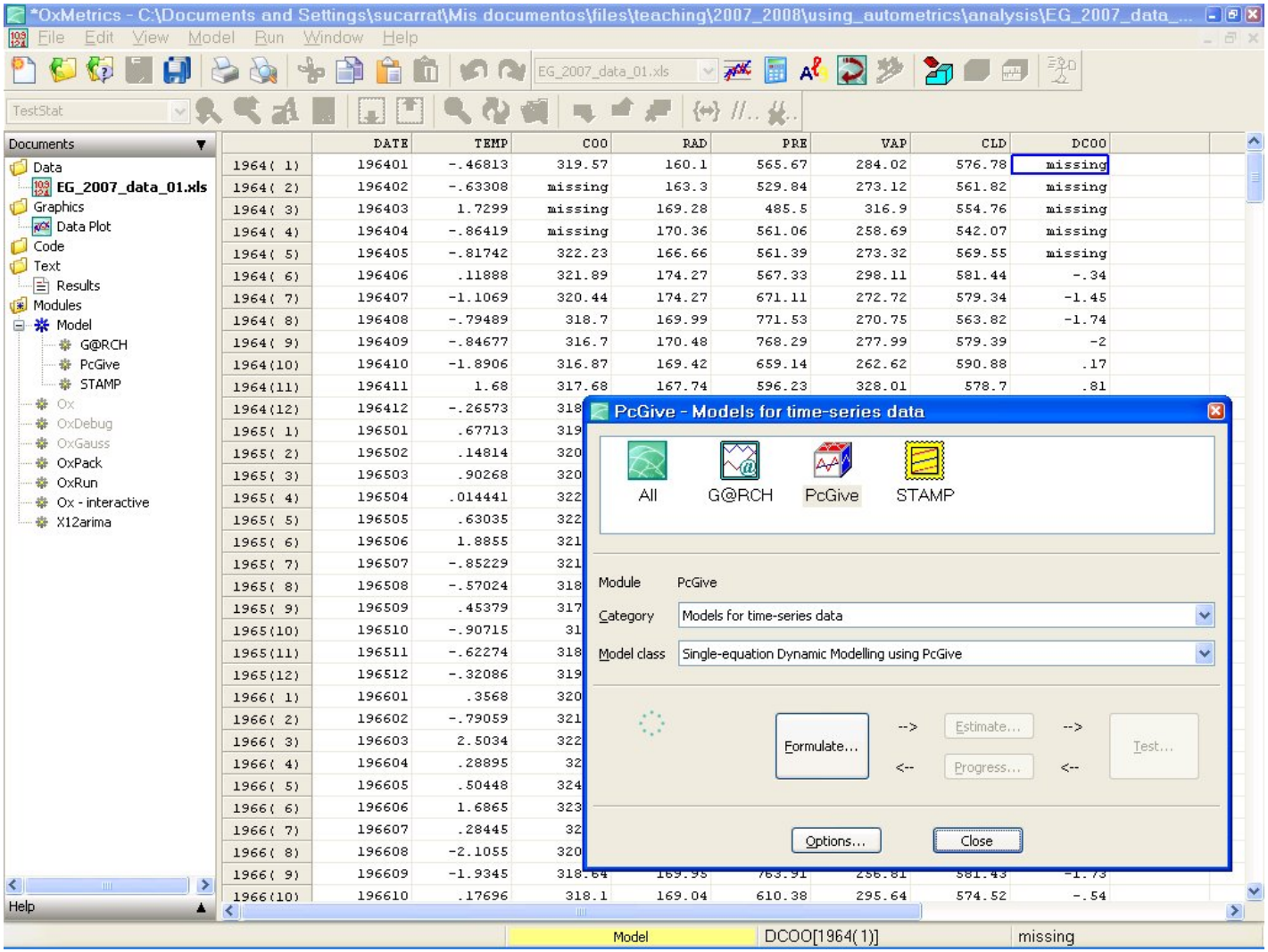
- Some estimation options (→ Options):

  → White (1980) standard errors: Tick "Heteroscedasticity consistent standard errors"

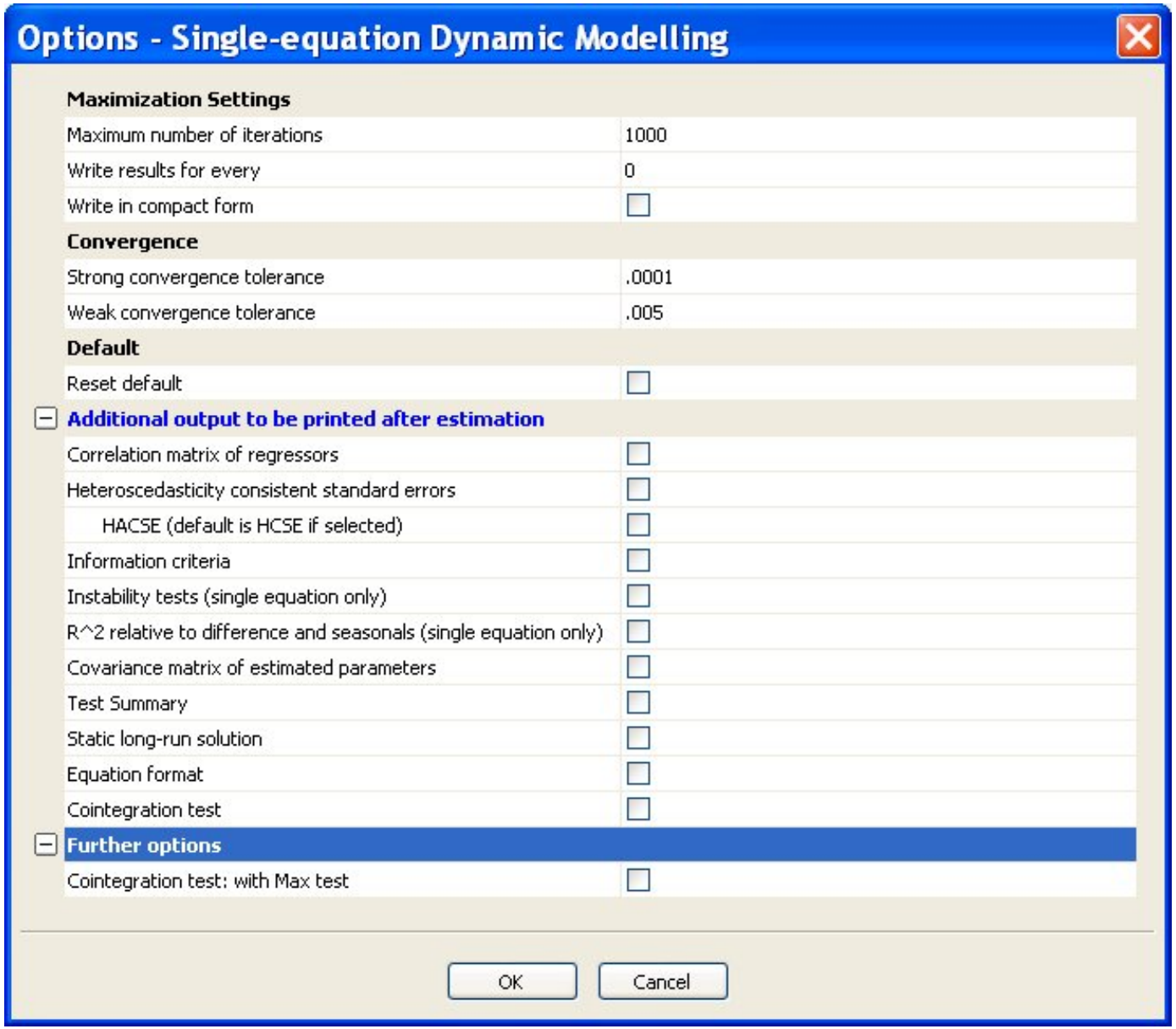
  → Newey and West (1987) standard errors: Tick "Heteroscedasticity consistent standard errors" *and* "HACSE"

  → Selected diagnostic tests: Tick "Test summary"

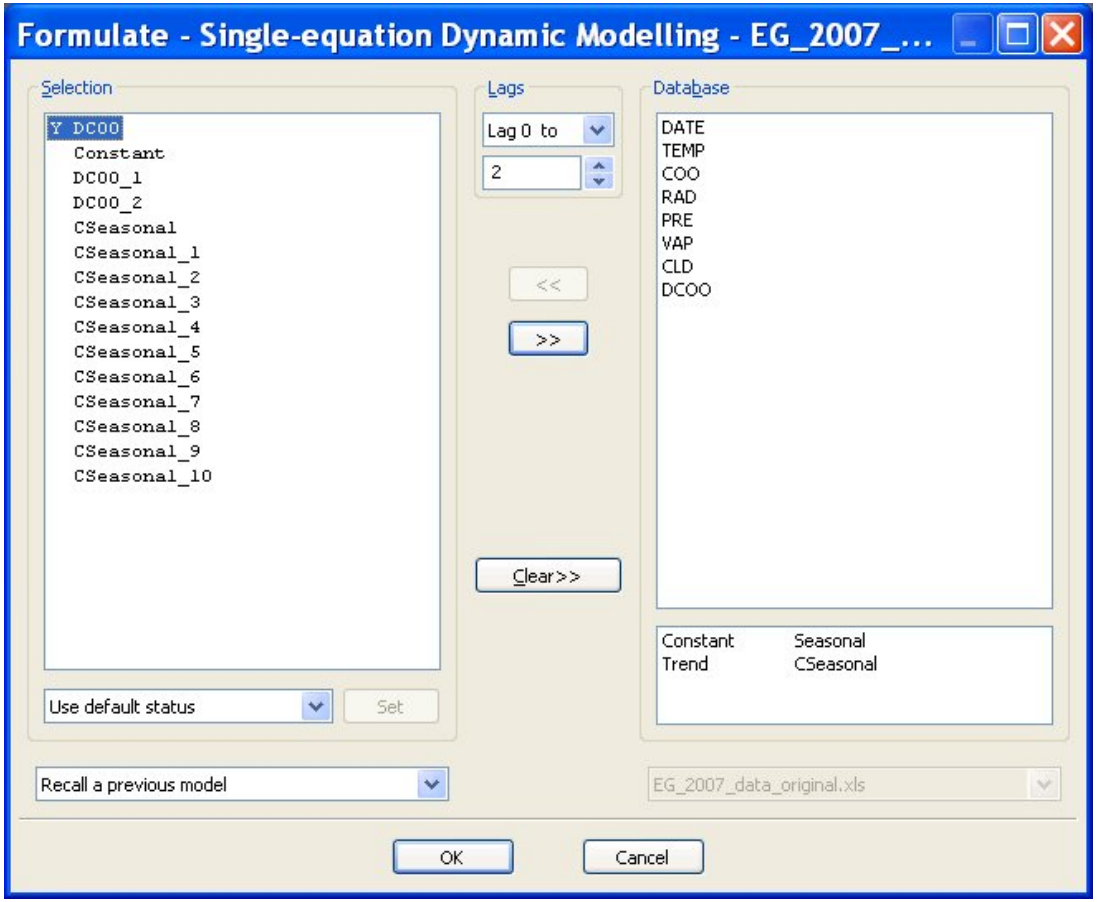
- Formulate a model: (Model →) PcGive → Category: "Models for time series data" → Model class: "Single-equation dynamic modelling using PcGive" → "Formulate"



- Estimation options (→ Options):



- Specify model (... → Formulate):



- "Seasonal", "CSeasonal": Seasonal dummies and centred seasonal dummies, respectively

- Estimate with default options: Ok → Ok → Ok

Single-equation GETS modelling with Autometrics. *Example*: 2007 Econometric Game, Question 1

- Recall the "rough" GUM:

$$\Delta COO_t = b_0 + b_1 \Delta COO_{t-1} + b_2 \Delta COO_{t-2} + \sum_{j=1}^{11} c_j d_{j,t} + e_t$$

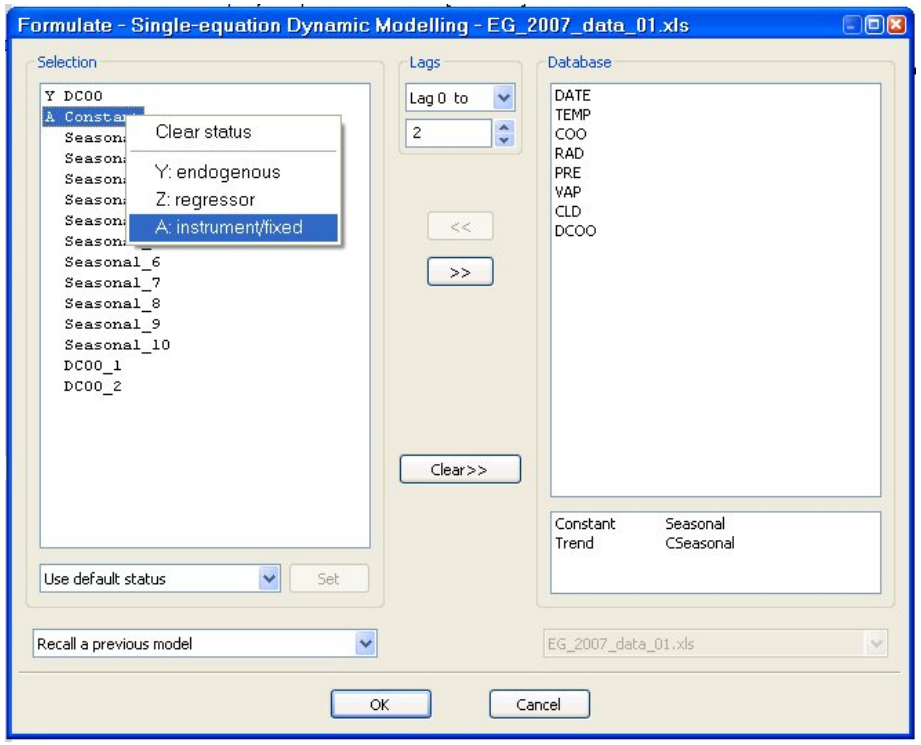
- The specific model proposed by Autometrics using the default options:

$$\Delta COO_t = b_0 + b_1 \Delta COO_{t-1} + \sum_{j=1}^{3} c_j d_{j,t} + \sum_{j=5}^{11} c_j d_{j,t} + e_t$$

- Level representation:

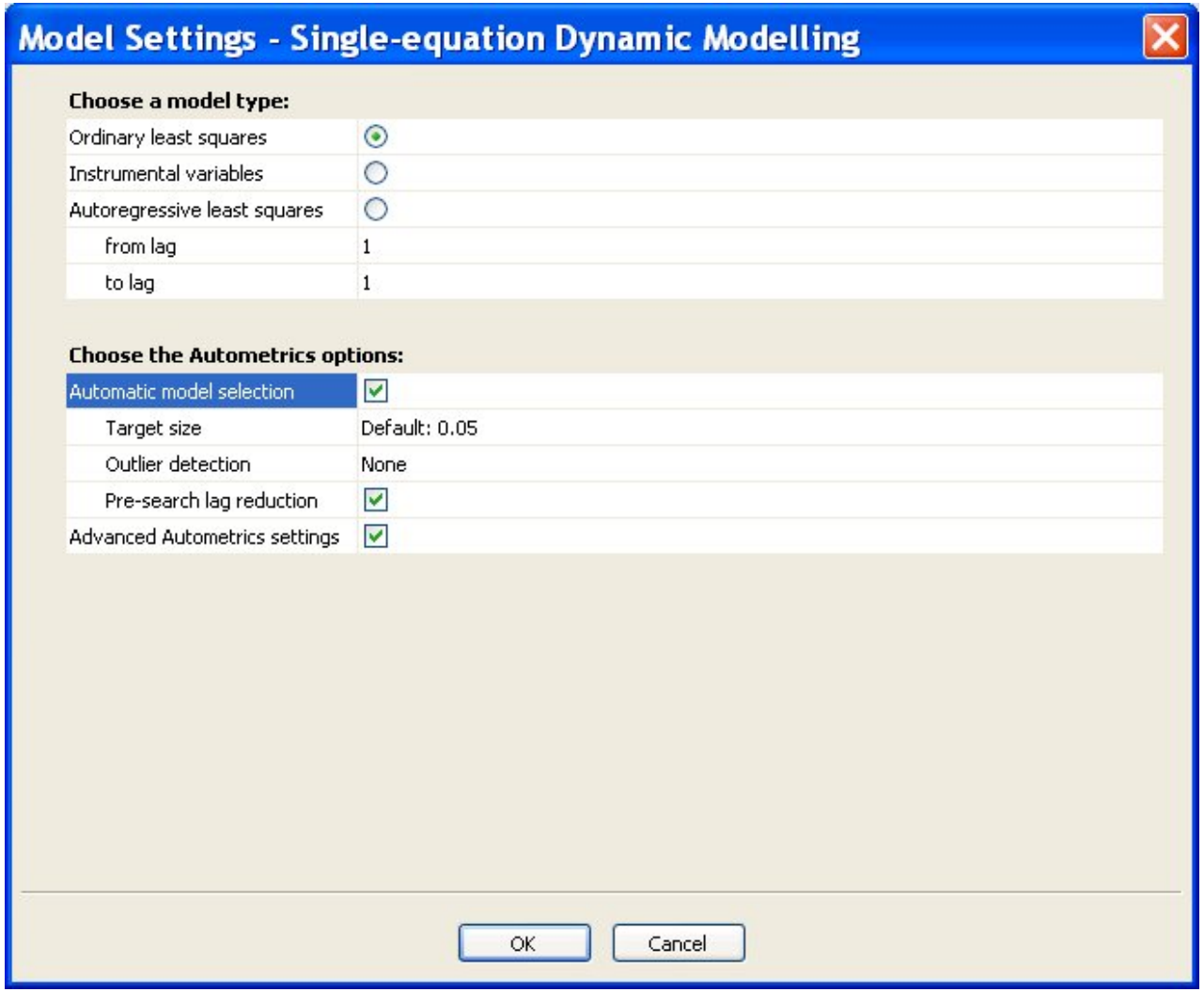
$$COO_t = b_0 + (1+b_1) COO_{t-1} + b_1 COO_{t-2} + \sum_{j=1}^{3} c_j d_{j,t} + \sum_{j=5}^{11} c_j d_{j,t} + e_t$$

- Specify model:



USEFUL FEATURE: Fixing regressors (that is, preventing Autometrics from deleting them). Select the regressors to fix → Right-click mouse → A: instrument/fixed. NOTE: Do the same thing to define instruments if IV is used instead of OLS

- GETS modelling with Autometrics: Tick "Automatic model selection"



- Main Autometrics options:

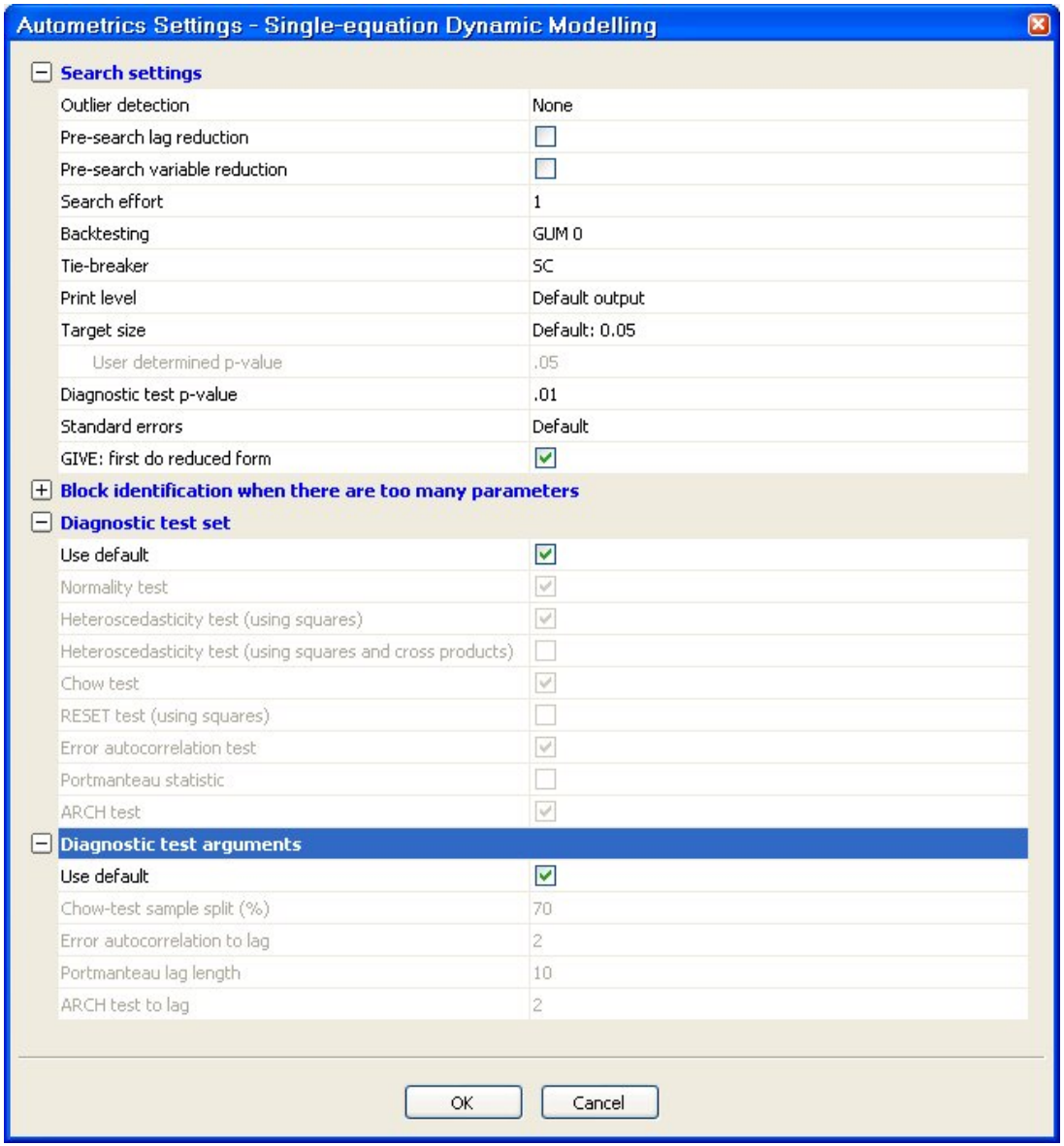
  → Target size: Regressor and backtesting significance level

  → Outlier detection: Neutralises large residuals in the GUM by means of impulse dummies

  → Pre-search lag reduction: Speeds up simplification; GENERAL ADVICE: Turn off!

  → Advanced Autometrics settings: Tick if default settings are unsatisfactory

- Advanced Autometrics settings:



- Selected advanced Autometrics settings:

  → Backtesting: "None" may be preferable if the final model does not encompass the initial GUM. "GUM0" is the initial GUM, which generally does not correspond to the "Current GUM"

  → Tiebreaker: The information criterion used to select between terminal models. SC (Schwartz) and min(k) (the model with the least regressors) are the most conservative

  → Diagnostic test *p*-value: The acceptable diagnostic test significance level. If deleting an insignificant variable results in a diagnostic test *p*-value above the acceptable level, then the variable is re-included into the model

  → Standard errors: Ordinary ("Default"), White (1980) ("HCSE") and Newey and West (1987) ("HACSE")

  → Heteroscedasticity tests: White (1980)

- "Recursive estimation": Slows down the computations (slightly), but it enables some very useful recursive stability analysis features

- Specific model proposed by Autometrics:

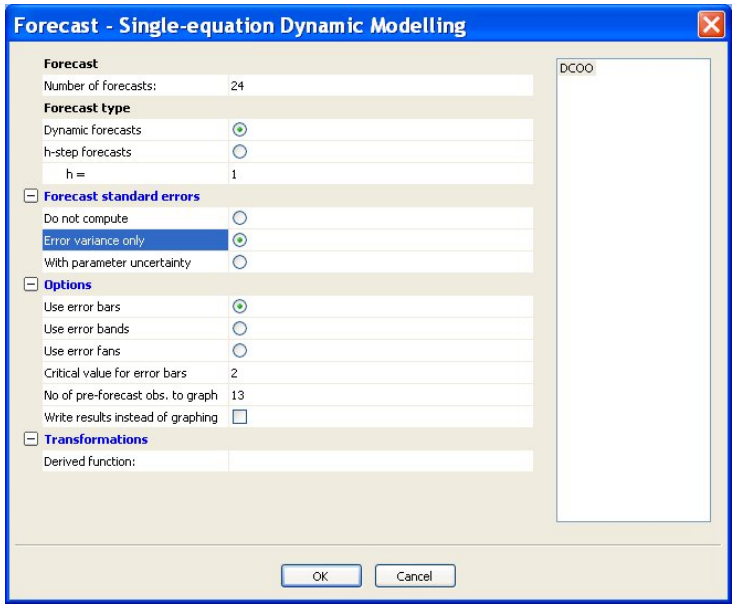
```
EQ( 2) Modelling DCOO by OLS
       The dataset is: C:\Documents and Settings\sucarrat\Mis documentos\files\teaching
       The estimation sample is: 1964(8) - 2000(12)

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
DCOO_1              -0.213999    0.04512    -4.74  0.0000   0.0503
Seasonal            -0.199927    0.05880    -3.40  0.0007   0.0265
Seasonal_1          -0.507123    0.05851    -8.67  0.0000   0.1502
Seasonal_2          -0.451573    0.06034    -7.48  0.0000   0.1165
Seasonal_4          -0.707806    0.05885    -12.0  0.0000   0.2540
Seasonal_5           -1.94870    0.06413    -30.4  0.0000   0.6848
Seasonal_6           -3.02784    0.09527    -31.8  0.0000   0.7038
Seasonal_7           -3.77271     0.1275    -29.6  0.0000   0.6733
Seasonal_8           -3.69728     0.1505    -24.6  0.0000   0.5868
Seasonal_9           -1.89487     0.1424    -13.3  0.0000   0.2941
Seasonal_10         -0.176490    0.07692    -2.29  0.0222   0.0122
Constant      U       1.50663    0.06042     24.9  0.0000   0.5940

sigma                0.286776  RSS                 34.9520821
R^2                  0.946555  F(11,425) =   684.3 [0.000]**
log-likelihood       -68.1549  DW                        2.05
no. of observations       437  no. of parameters           12
mean(DCOO)           0.112334  var(DCOO)              1.49654

AR 1-7 test:      F(7,418)  =   1.0154 [0.4196]
ARCH 1-7 test:    F(7,411)  =  0.78778 [0.5979]
Normality test:   Chi^2(2)  =   2.5460 [0.2800]
Hetero test:      F(12,412) =  0.85756 [0.5908]
Hetero-X test:    F(22,402) =   1.2915 [0.1715]
RESET test:       F(1,424)  =0.00067988 [0.9792]
```

Some further diagnostic tests:

- Residuals graphs: Model $\rightarrow$ Test $\rightarrow$ Graphical analysis $\rightarrow \ldots$
- User specified residuals tests: Model $\rightarrow$ Test $\rightarrow$ Test $\rightarrow \ldots$
- Recursive graphics (VERY useful!): Model $\rightarrow$ Test $\rightarrow$ Recursive graphics $\rightarrow \ldots$

Single equation dynamic forecasting:

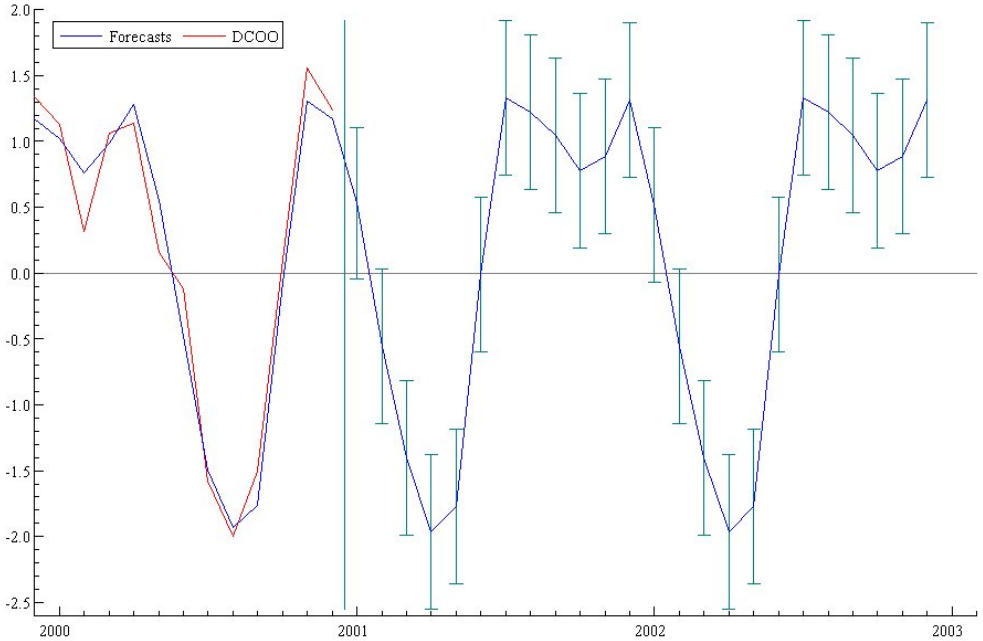
- The parsimonious model suggested to us by Autometrics contains lags and deterministic terms only, so we may readily generate dynamic forecasts beyond 2000(12)
- Forecasting DCOO dynamically 24 months beyond 2000(12): Model → Test → Forecast and then



yields (graph on next slide)

Single equation dynamic forecasting (cont.):

- Out-of-sample forecasts of DCOO from 2001(1)-2002(12):



- In order to generate forecasts of the *level* of COO, recall that any variable y_T satisfies $y_T = y_0 + \sum_{t=1}^{T} \Delta y_t$. In other words, tick "Write results instead of graphing" and use Algebra or Calculator (Model $\rightarrow$ Calculator) to obtain the forecasts of the levels COO

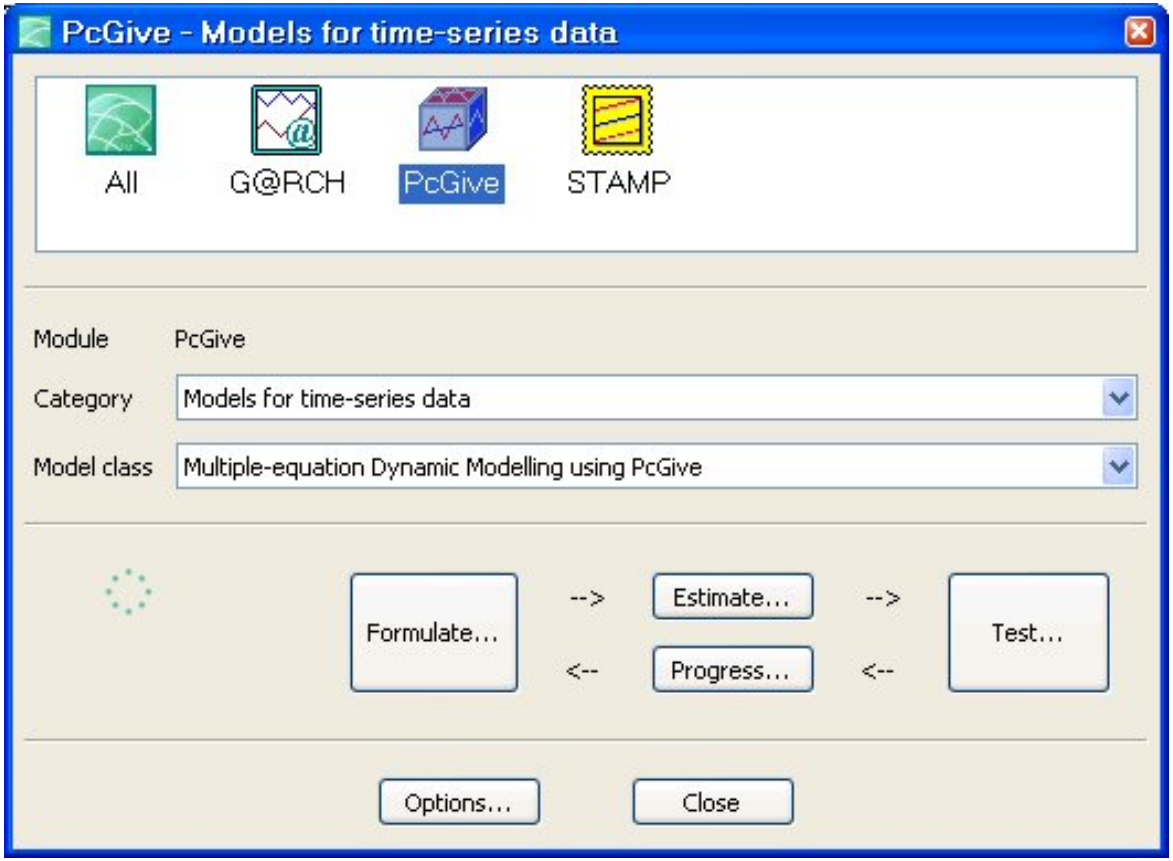
Multiple-equation modelling with Autometrics (two approaches):

1. Seemingly Unrelated Regression (SUR) using OLS/IV, that is, single-equation GETS modelling of each equation separately (requires stationarity of regressors)

2. Simultaneous variable deletion (or non-deletion) across equations using vector diagnostic tests but estimation still by OLS (does not require stationarity of regressors), see Doornik and Hendry (2007b, pp. 29-31). (NOTE: IV estimation not available with this strategy)

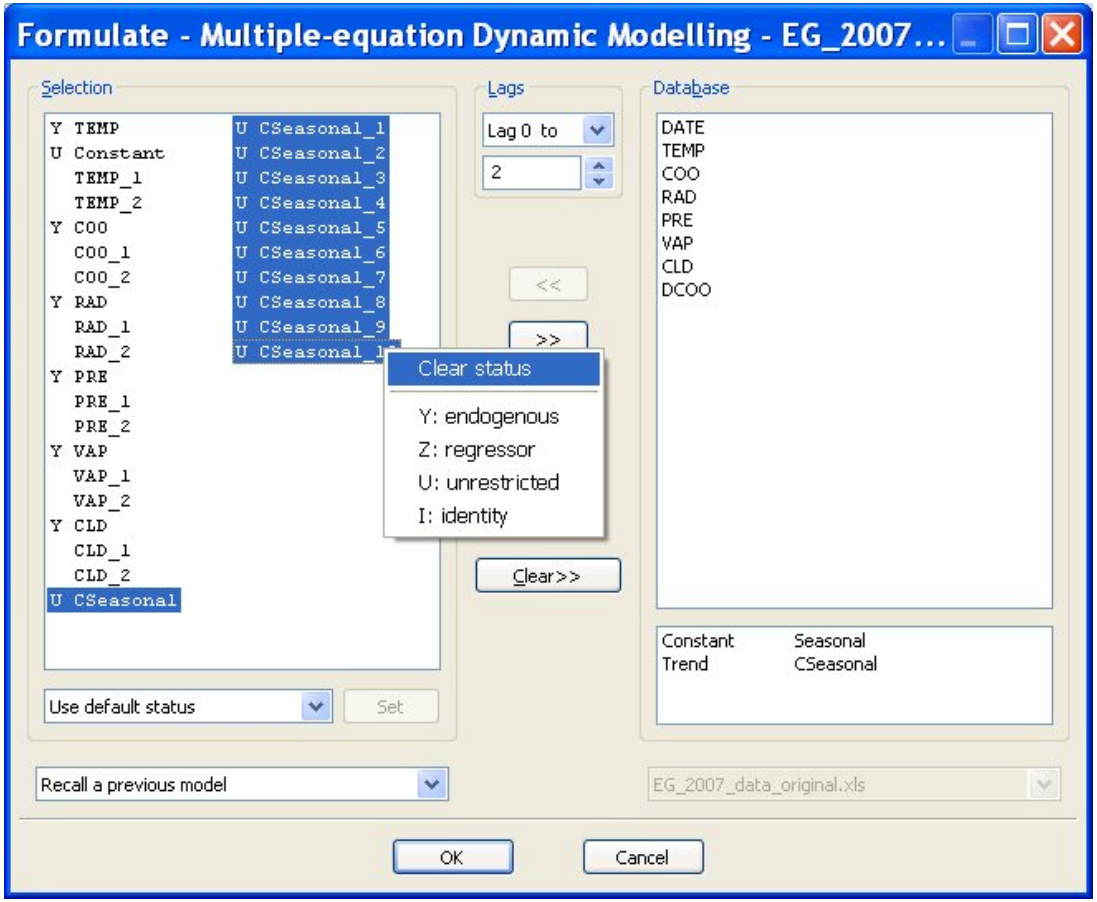
→ Model type: "Unrestricted system" (system of URFs), see Doornik and Hendry (2007b, chapter 3)

Formulate a system: (Model →) PcGive → Category: "Models for time series data" → Model class: "Multiple-equation dynamic modelling using PcGive" → "Formulate"



Multiple-equation modelling with Autometrics using second approach. *Example*: 2007 Econometric Game, Question 2

- My GUM: A six-dimensional VAR(2) of $y_t = (TEMP_t, COO_t, RAD_t, PRE_t, VAP_t, CLD_t)$, with a constant and 11 centered seasonals in each of the six equations:



- Fixing variables (that is, restricting Autometrics to keep them) now differs:

  ◇ Fixing VAR-lags is not possible, only the exogenous regressors can be fixed

  ◇ Select the exogenous regressors to fix → Right-click mouse → U: Unrestricted

  ◇ To unfix exogenous regressors, select the regressors to unfix → Right-click mouse → Clear status

- Lag-deletion is undertaken across equations. For example, $TEMP_1$ is either deleted from all six equations or from none, etc.

Results with default settings:

• NOTE: Autometrics simplifies even though the GUM does not pass all diagnostic checks

• Four variables are removed from all of the equations: The second lag of TEMP, VAP and CLD, and CSeasonal_10

Other type of analysis:

• Cointegration analysis (applied on the Unrestricted system, *not* on the simplified model): Model → Test → Dynamic Analysis and Cointegration Tests → ...

See Doornik and Hendry (2007b, chapter 4)

References:

Doornik, J. (2009). Autometrics. In J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*. Oxford: Oxford University Press.

Doornik, J. A. and D. F. Hendry (2007a). *Empirical Econometric Modelling - PcGive 12: Volume I*. London: Timberlake Consultants Ltd.

Doornik, J. A. and D. F. Hendry (2007b). *Empirical Econometric Modelling - PcGive 12: Volume II*. London: Timberlake Consultants Ltd.

Hendry, D. F. and H.-M. Krolzig (2001). *Automatic Econometric Model Selection using PcGets*. London: Timberlake Consultants Press.

Hendry, D. F. and H.-M. Krolzig (2005). The Properties of Automatic Gets Modelling. *Economic Journal 115*, C32–C61.

Hoover, K. D. and S. J. Perez (1999). Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search. *Econometrics Journal 2*, 167–191. Dataset and code: `http://www.csus.edu/indiv/p/perezs/Data/data.htm`.

Newey, W. and K. West (1987). A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica 55*, 703–708.

Sucarrat, G. (2009). Forecast Evaluation of Explanatory Models of Financial Variability. *Economics – The Open-Access, Open-Assessment E-Journal 3*. Available via: `http://www.economics-ejournal.org/economics/journalarticles/2009-8`.

Sucarrat, G. and Á. Escribano (2009). Automated Model Selection in Finance: General-to-Specific Modelling of the Mean, Variance and Density. `http://www.sucarrat.net/research/autofim.pdf`.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica 48*, 817–838.