Notes on Automatic Econometric Model Selection

David F. Hendry
Department of Economics, Oxford University
May 2, 2005

Framework of paper

(1) Introduction.
(2) Review progress in PcGets.
(3) Assisting non-experts in model formulation.

Solving five ‘intractable’ problems:

(4) Unbiased selection with accurate standard errors.
(5) Resolving perfect multicollinearity.
(6) Simultaneous equations models.
(7) Selecting when too many regressors.
(8) Selecting non-linear models.
(9) Conclusion.

1 Introduction


Model selection

How to find the source of the Nile? Explore every path – N, S, E, and W – succeeded. Gets does this for regression model selection. Search all reduction paths in general model. Two costs of selection: costs of inference and costs of search. First inevitable if tests of non-zero size and non-unit power, even if commence from data generation process (DGP). Costs of search additional to initial model being the DGP. How expensive is it to search? Automatic model selection does not have high costs, and is essential if many potential candidate variables.

Summary of PcGets

Retains relevant variables close to theory maximum.
Eliminates irrelevant at chosen significance level.
PcGets model selection is consistent.
Equation standard error close to that of DGP.
Selected estimates have appropriate standard errors.
Can be bias corrected despite selection — which also down-weights adventitious significance.
Can tackle: more variables than observations; perfectly multicollinear variables; simultaneous equations without prior information; and non-linear models.
General-to-specific modelling

Start from general statistical model (GUM): embed previous findings and available theory. Check GUM captures essential characteristics of data: ensures congruence and valid inferences. Eliminate insignificant variables to reduce complexity: diagnostic checks on validity of reductions – ensures congruence of final model. Multi-path searches avoid path-dependence. Selection encompasses rival contenders. Check reliability on sub-samples.

Specification of GUM

(i) Larger GUM: retain more adventitious effects.
(ii) Smaller GUM: omit key variables.
(iii) Prior analysis is essential: relevant variables, functional form, indicators, etc.
(iv) Previous evidence to ensure encompassing.
(v) The less ‘orthogonality’, the more ‘confusion’.
(vi) Central role for theory in ‘prior simplification’.

Repeated testing

Does repeated testing distort selection?

(a) Severe illness: more tests increase probability of correct diagnosis.
(b) Mis-specification tests: if \( r \) independent tests \( \tau_j \) conducted under null for small significance level \( \eta \) (critical value \( c_\eta \)):

\[
P(|\tau_j| < c_\eta \mid j = 1, \ldots, r) = (1 - \eta)^r \simeq 1 - r\eta.
\]

More tests increase probability of false rejection. Suggests significance level \( \eta \) of 1% or tighter.
(c) Repeated diagnostic tests: probabilities unaltered. Conclude: no generic answer.

Selection of parsimonious model


Selection strategies

M-variable DGP nested in \( N > M \)-variable GUM. Assume \( T \) observations, with \( N = K + M \) for \( K \) irrelevant; and \( n = k + m \) sized selected model.

Pre-defined search strategies:

Liberal strategy: reduce non-selection probability.
Conservative strategy: reduce non-deletion probability.
Expert strategy: can re-define all options.
Quick modeller: makes all decisions (Liberal strategy).

Selection not based on fit: but minimal congruent encompassing model, which will fit best.

Choice of strategy

Almost no analyses of ‘strategies’. Analogy of sieving for gold (t-values > 0) in dirt pile.

I. Large nuggets in large pile: strict significance levels—conservative strategy.

II. Medium nuggets in moderate pile: relatively loose significance levels. If \( M = 6, \psi = 3, K = 20, T \geq 100, \alpha = 0.05 \). \( k = 1 \): one irrelevant variable retained. Power is:
\[ P (|t| > 1.985 \mid \psi = 3) \simeq P (t > -1.015 \mid H_0) \simeq 0.844. \]

so on average omit \( 6 \times (1 - 0.844) \simeq \text{one relevant variable} \). Basis of Liberal strategy.

III. Small nuggets in large pile. Close to impossible—null and alternative overlap. Balance risks of losing relevant against keeping irrelevant given the purpose of modelling exercise.

IV. Gold dust in large pile. ‘Data snooping’: Sullivan, Timmermann and White (2001). If huge number of potential hypotheses, significance levels hard to control. Could still try with Conservative strategy.

V. Pile exceeds sieve capacity. This is \( N > T \). If \( T > (m + k) \) – just keep sieving! Return to this case shortly.

2 Review progress in PcGets

Table 1: Monte Carlo designs

| Design | \( N \) | \( M \) | \( K \) | \( |t|\)-values |
|--------|--------|--------|--------|----------------|
| HP0    | 41     | 0      | 41     | 0              |
| HP2\*  | 41     | 1      | 40     | 5.77           |
| HP2    | 41     | 1      | 40     | 11.34          |
| HP7    | 41     | 3      | 38     | (10.9, 16.7, 8.2) |
| JEDC   | 22     | 5      | 17     | (2,3,4,6,8)    |
| S0     | 34     | 0      | 34     | 0              |
| S2     | 34     | 8      | 26     | (2,2,2,2,2,2,2) |
| S3     | 34     | 8      | 26     | (3,3,3,3,3,3,3) |
| S4     | 34     | 8      | 26     | (4,4,4,4,4,4,4) |
| S0\*   | 42     | 0      | 42     | 0              |
| S2\*   | 42     | 8      | 34     | (2,2,2,2,2,2,2) |
| S3\*   | 42     | 8      | 34     | (3,3,3,3,3,3,3) |
| S4\*   | 42     | 8      | 34     | (4,4,4,4,4,4,4) |

Properties of the strategies

Figures 1 shows four main aspects.

(a) ‘Unbiased’ fit: \( \hat{\sigma} \) close to \( \sigma \). Liberal slight downward bias (< 5%), Conservative upward.

(b) Sizes of strategies close to their targets.

(c) Power close to upper bound.

(d) Probabilities of locating DGP show ‘problem’ for search algorithm may be cost of inference: DGP is sometimes never retained as model.

Choice of strategy

Conservative no higher power than Liberal: cost of avoiding spurious variables can be high. In Hoover–Perez, Conservative best. For \( t\)-values of 2 or 3, Liberal does well: sometimes outperforms DGP start if 1% significance. Relative performance depends on state of nature. Confirms strategies have desired operating characteristics. Key cites: Hendry and Krolzig (2001, 2003) and Krolzig and Hendry (2001).
3 Quick modeller for non-experts

User specifies regressand and basic regressors: *PcGets* creates GUM and selects model. Sets maximum lag length from data frequency; checks congruence of resulting GUM; estimates levels unrestrictedly; uses *PcGive* unit-root test; transforms to differences and cointegration; re-estimates; then selects parsimonious undominated model.

Example: modelling UK consumers’ expenditure. Key component of GNP (65%); and highly variable, as figure 2 shows: **15 percentage point swings between quarters**. Consider study by Davidson et al. (1978) [DHSY]. Constant price consumers’ expenditure is modelled using: constant price disposable income; annual inflation, and past values of all three variables. Finally, model is reformulated to explain annual changes in consumption. *PcGets* reports same equation as DHSY. Excellent properties – equation error = 0.6%. Accurate forecasts. Model derived in a fraction of the time it took the original authors. Invaluable for labour saving! Main *caveat* is non-orthogonality, so consider collinearity.

4 Unbiased selection estimates

Conditional on retaining, coefficients upward biased: more so for Conservative and for small $t$. If $\beta \geq 0$:

$$E \left[ \hat{\beta} \mid \hat{\beta} \geq \sigma_\beta c_\alpha \right] = \beta \left( 1 + \psi^{-1} R(\psi, c_\alpha) \right)$$  \hspace{1cm} (1)

with:

$$R(\psi, c_\alpha) = \frac{\phi(c_\alpha - \psi) - \phi(-c_\alpha - \psi)}{1 - \Phi(c_\alpha - \psi) + \Phi(-c_\alpha - \psi)}.$$ \hspace{1cm} (2)

Unconditionally, coefficient estimates downward biased. Conditional distributions for non-DGP variables bimodal and symmetric, except for lagged endogenous variable: impact of famous Hurwicz (1950) bias is clear.

Figure 1: Accuracy, size, power and success
Figure 2: Consumption and Income data

Key: ‘pre-test’ effects not changed by search: coefficient biases similar from the DGP and the GUM.

Table 2: Conditional estimated SEs and SDs

<table>
<thead>
<tr>
<th>variable</th>
<th>DGP</th>
<th>D-Lib</th>
<th>D-Cons</th>
<th>GUM</th>
<th>G-Lib</th>
<th>G-Cons</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_a$</td>
<td>0.103</td>
<td>0.102</td>
<td>0.102</td>
<td>0.113</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>$Z_b$</td>
<td>0.102</td>
<td>0.102</td>
<td>0.102</td>
<td>0.112</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>$Z_c$</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
<td>0.113</td>
<td>0.101</td>
<td>0.102</td>
<td>0.100</td>
</tr>
<tr>
<td>$Z_d$</td>
<td>0.102</td>
<td>0.103</td>
<td>0.104</td>
<td>0.113</td>
<td>0.101</td>
<td>0.103</td>
<td>0.100</td>
</tr>
<tr>
<td>$Z_e$</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
<td>0.113</td>
<td>0.101</td>
<td>0.103</td>
<td>0.100</td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_a$</td>
<td>0.103</td>
<td>0.066</td>
<td>0.061</td>
<td>0.115</td>
<td>0.070</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td>$Z_b$</td>
<td>0.102</td>
<td>0.082</td>
<td>0.075</td>
<td>0.113</td>
<td>0.084</td>
<td>0.075</td>
<td></td>
</tr>
<tr>
<td>$Z_c$</td>
<td>0.103</td>
<td>0.095</td>
<td>0.089</td>
<td>0.115</td>
<td>0.098</td>
<td>0.090</td>
<td></td>
</tr>
<tr>
<td>$Z_d$</td>
<td>0.103</td>
<td>0.102</td>
<td>0.104</td>
<td>0.116</td>
<td>0.108</td>
<td>0.106</td>
<td></td>
</tr>
<tr>
<td>$Z_e$</td>
<td>0.106</td>
<td>0.100</td>
<td>0.102</td>
<td>0.119</td>
<td>0.111</td>
<td>0.110</td>
<td></td>
</tr>
</tbody>
</table>

Conditional ESEs close to unbiased for SDs of DGP. Astonishing result: despite selection, estimated uncertainty for relevant variables correct for uncertainty in DGP without selection.

Intuition: SDs in estimated DGP correctly estimated by ESEs. Latter based on $\hat{\sigma} \approx \sigma$ times square-roots of $(X'X)^{-1}$; approximately same in selected model if variable retained.

Conditional sampling standard deviations smaller.

Note that $PcGets$ does not ‘overfit’: $\hat{\sigma}$ little biased; and while retain $k = \alpha K$ spurious variables, lose some relevant. Graphs show obvious quality of discrimination.

Little impact of selection on test statistics: only a slight change in quantiles above nominal significance.
Figure 3: Conditional distributions: Liberal

Figure 4: Unconditional distributions: Liberal
but increasing impact as quantile decreases. Bound to occur: models with significant heteroscedasticity not selected.

Not a ‘distortion’ of sampling properties: decision is taken for GUM. Conditional on that, no change should occur. Figure 5 shows ratios of actual to nominal sizes.

<table>
<thead>
<tr>
<th>Model</th>
<th>DGP</th>
<th>ARCH test</th>
<th>GUM</th>
<th>Heteroscedasticity test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JEDC</td>
<td>S2</td>
<td>S2*</td>
<td>S2</td>
<td>S2*</td>
</tr>
<tr>
<td>S3</td>
<td>S3*</td>
<td>S3*</td>
<td>S3</td>
<td>S3*</td>
</tr>
<tr>
<td>S4</td>
<td>S4*</td>
<td>S4*</td>
<td>S4</td>
<td>S4*</td>
</tr>
<tr>
<td>HP2</td>
<td>HP2*</td>
<td>HP2*</td>
<td>HP9</td>
<td>HP9</td>
</tr>
</tbody>
</table>

Figure 5: Heteroscedasticity test outcomes

**Can correct final estimates for selection.** Convenient approximations that:

\[
t_{\hat{\beta}} = \frac{\hat{\beta}}{\hat{\sigma}_\beta} \sim \frac{\bar{\beta}}{\sigma_{\bar{\beta}}} \sim N \left[ \frac{\beta}{\sigma_{\beta}}, 1 \right] = N \left[ \psi, 1 \right].
\]

when non-centrality of t-test is \( \psi = \frac{\beta}{\sigma_\beta} \). Using Gaussian approximation:

\[
\phi(w) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} w^2 \right)
\]

\[
\Phi(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} \exp \left( -\frac{1}{2} x^2 \right) dx.
\]

then:

\[
E \left[ t_{\hat{\beta}} \mid t_{\hat{\beta}} > c_\alpha; \psi \right] = \psi + \frac{\phi(c_\alpha - \psi)}{1 - \Phi(c_\alpha - \psi)}
\]

\[
E \left[ t_{\hat{\beta}} \mid t_{\hat{\beta}} < -c_\alpha; \psi \right] = \psi - \frac{\phi(-c_\alpha - \psi)}{\Phi(-c_\alpha - \psi)}
\]

Doubly-truncated distribution, so expected truncated t-value is \( E[|t_{\hat{\beta}}| \mid |t_{\hat{\beta}}| > c_\alpha; \psi] = \psi^* \) with:

\[
\psi^* = \psi + \frac{\phi(c_\alpha - \psi) - \phi(-c_\alpha - \psi)}{1 - \Phi(c_\alpha - \psi) + \Phi(-c_\alpha - \psi)} = \psi + R(\psi; c_\alpha).
\]
Thus, observe \( \psi^* \) when true non-centrality is \( \psi \). Know mapping, so correct via ‘inversion’. Applies as well to \( \beta \). For \( \psi = 0 \), truncated distribution is symmetric around zero. Figure 7a illustrates for \( \psi^* = 0 \). However, even if \( \psi = 1 \) and \( c_\alpha = 2 \), lower tail contributes almost nothing to mean: see figure 7b for \( \psi = 2 \), Panels c, d confirm. Matches closeness of (1) to simulation outcomes for non-central t-statistics.

Figure 6: Non-linearity of bias function

\( \psi \) is unknown so estimate by \( \tilde{\psi} \) based on \( t_\beta \) after iteration:

\[
\tilde{\psi} \simeq t_\beta - R(\psi, c_\alpha).
\]

Then \( E[\tilde{\psi}] \simeq \psi \). Use \( \tilde{\psi} \) in:

\[
\tilde{\beta} = \hat{\beta} \left( \frac{\tilde{\psi}}{t_\beta} \right).
\]

Figure 8 compares biases and RMSEs. Figure 9 plots double-corrected conditional distributions.

Implications

If regress \( y_t \) on \( \{z_{j,t}\} \) \( (j = 1, \ldots, N) \) over \( t = \frac{1}{n}, \ldots, T \); select \( \{z_{j,t}\} \) \( (j = 1, \ldots, n) \) by \( PcGets \), then in:

\[
\hat{y}_t = \sum_{j=1}^{n} \tilde{\beta}_j \left( \frac{\tilde{\psi}}{t_\beta} \right) z_{j,t} + \hat{\epsilon}_t \frac{\hat{\sigma}}{\sigma}
\]

a. Estimates near unbiased, \( E[\tilde{\beta}_j] \simeq \beta_j \), for parameter in DGP when bias-corrected estimates reported.
b. SEs accurate for SDs of estimated DGP equation: \( V[\tilde{\beta}_j] \simeq V[\beta_j] \) for that \( \tilde{\beta}_j \) estimated in DGP equation.
c. Estimated equation standard error nearly unbiased \( (\hat{\sigma} \simeq \sigma) \); and
d. Relevant variables retained with almost same probabilities as commencing from DGP.
Figure 7: Probability densities

Figure 8: Biases and RMSEs
5 Resolving perfect multicollinearity

Assume DGP is identifiable and estimable. GUM cannot be estimated: but no bivariate collinearity. Illustrate by simple example where DGP is:

\[ y_t = \beta_0 1_{\{0,d\}} z_t + \beta_1 1_{\{1,d\}} z_{t-1} + \beta_2 1_{\{2,d\}} (z_t + z_{t-1}) + \beta_3 1_{\{3,d\}} \Delta z_t + v_t \]  

(6)

1_{\{j,d\}} indicators for regressor entering DGP. At most two indicators non-zero: DGP is identified. 11 cases to consider, but only 6 distinct: (a) none; (b) \( z_t \); (c) \( z_{t-1} \); (d) \( \Delta z_t \); (e) \( (z_t + z_{t-1}) \); (f) \( z_t \) and \( \Delta z_t \).

Select for which combination of \( z_t \), \( z_{t-1} \), \( (z_t + z_{t-1}) \) and \( \Delta z_t \) enters. GUM is unidentified, but sub-models are estimable:

\[ y_t = \gamma_0 z_t + \gamma_1 z_{t-1} + \gamma_2 (z_t + z_{t-1}) + \gamma_3 \Delta z_t + u_t. \]  

(7)

Inversion routines arbitrarily assign singularity. But all 4 can be entered if use multi-path searches. **Explore all ‘drop any two’ paths.** Once non-collinear, revert to usual algorithm. (a) will hold 95% of time at 5% F-test. In an artificial data illustration: \( \text{PcGets} \) correct in (b)–(e), but in (f): selects orthogonal representation \( (z_t + z_{t-1}) \) and \( \Delta z_t \). **New tool yields new insights.**

\( \text{PcGets} \) provides ‘collinearity analysis’: correlation matrix and its eigenvalues. Eigenvalues only invariant under orthogonal transforms; correlations unreliable indicators of collinearity: source of correlation matters. Correlation of 0.99 unproblematic if DGP is random walk. But disaster if bivariate normal regressors: no information on one variable after orthogonal transform. Serious indirect cost of collinearity is to preclude use of ordered \( t \)-values in GUM for selection. Transforming to a ‘near orthogonal’ representation requires analyzing the regressors, so heading for a system variant of \( \text{Gets} \).

Illustrate properties of \( \text{PcGets} \) by Monte Carlo on JEDC. Figure 10 summarizes size, power and probability of finding the DGP from:

(i) GUM in levels
(ii) orthogonalised GUM
(iii) orthogonalised GUM & constant effective sample.
Collinearity: significant loss in power and increasing size. (iii) stabilizes size and power, \( \rho \)-invariant. Partly due to changing time-series properties – check that by variant of JEDC design.

\[
y_t = \sum_{j=1}^{5} \beta_j y_{t-j} + u_t, \quad u_t \sim \text{IN}[0, \omega^2],
\]

\[
z_t = v_t, \quad v_t \sim \text{IN}[0, \Sigma]
\]

for \( t = 1, \ldots, T \).

Limited effects of collinearity on size or power.

\[
y_t = \pi_{0.0} + \pi_{0.1} y_{t-1} + \sum_{k=1}^{10} \sum_{i=0}^{1} \pi_{k,i} z_{k,t-i} + w_t, \quad w_t \sim \text{IN}[0, \sigma_w^2].
\]

Figure 10: Selection by \textit{PcGets} for varying \( \rho \)

6 Simultaneous equations models

Embed selection of linear simultaneous models in theory of reduction: SEMs are reductions of systems. Given endogenous \( y_t \) and non-modelled variables \( z_t \), congruent linear conditional statistical system formulated:

\[
y_t = \Psi z_t + v_t, \quad v_t \sim \text{IN}[0, \Omega_v].
\]

Always identified, so all later selections are also. Model is:
\[ \mathbf{B} \mathbf{y}_t = \mathbf{C} \mathbf{z}_t + \epsilon_t \quad \text{where} \quad \epsilon_t \sim \mathcal{N}_m [\mathbf{0}, \Sigma_{\epsilon}] \] (11)

Rank condition \( \mathbf{B} \Psi = \mathbf{C} \) imposed as constraint on searches. Endogenous variables added \textit{seriatim} to each equation and reductions checked to eliminate additional regressors, \textbf{contingent on maintaining identification}. SEMs as reductions of systems is known for programming, modelling, model evaluation, and testing theory. Two main possibilities if \textbf{selecting} in any equation on adding an endogenous regressor and deleting an exogenous:

(1) no reductions are found;
(2) at least one further reduction is feasible.

In (1), equation either not identified, or an ‘equivalent’ just-identified equation found. In (2), equation over identified and testable: eliminated regressors must be determinants of endogenous effects.

\textbf{Rank condition imposed as a constraint on searches}, so current ‘partial structure’ is fully identified at every step. ‘Structure’ denotes equation with more than one endogenous variable: not necessarily structural (invariant to extensions of information for new variables, over time, and across regimes). Weak instruments imply poorly determined initial system. Framework for other sources of endogeneity, such as measurement errors.

Add \( y_{k,t} \) \textbf{to equation for} \( y_{l,t} \). Drop regressors in turn. \textbf{If rank condition accepted, could have}:

a) no reduction possible; equation for \( y_{l,t} \) just identified: accept parsimonious system.
b) \( y_{k,t} \) significant, and eliminates other regressors; resulting equation more parsimonious than system, is congruent, and identified, so select. Excluded regressors must be instruments for \( y_{k,t} \).

\textbf{Interpretation} depends on economic theory: \textit{identification} on properties of DGP.
Two-equation example

Conventional supply-demand example with 4 instruments:

\[ \begin{align*}
\epsilon_t \sim \mathcal{N}_2(0, \Sigma), \\
y_{1,t} + \beta_{1,2}y_{2,t} &= \gamma_{1,1}z_{1,t} + \gamma_{1,2}z_{2,t} + \epsilon_{1,t} \\
y_{2,t} + \beta_{2,1}y_{1,t} &= \gamma_{2,1}z_{1,t} + \gamma_{2,3}z_{3,t} + \gamma_{2,4}z_{4,t} + \epsilon_{2,t}
\end{align*} \]

(12)

\[ \begin{pmatrix}
y_{1,t} \\
y_{2,t}
\end{pmatrix} = d \begin{pmatrix}
\gamma_{1,1} - \beta_{1,2}\gamma_{2,1} & \gamma_{1,2} - \beta_{1,2}\gamma_{2,3} & -\beta_{1,2}\gamma_{2,4} \\
\gamma_{2,1} - \beta_{2,1}\gamma_{1,1} & -\beta_{2,1}\gamma_{1,2} & \gamma_{2,3}
\end{pmatrix} \begin{pmatrix}
z_{1,t} \\
z_{2,t} \\
z_{3,t} \\
z_{4,t}
\end{pmatrix}
\]

+ \begin{pmatrix}
\epsilon_{1,t} - \beta_{1,2}\epsilon_{2,t} \\
\epsilon_{2,t} - \beta_{2,1}\epsilon_{1,t}
\end{pmatrix}

(13)

Unrestricted projection on unmodelled variables is:

\[ \begin{align*}
y_{1,t} &= \phi_{1,1}z_{1,t} + \phi_{1,2}z_{2,t} + \phi_{1,3}z_{3,t} + \phi_{1,4}z_{4,t} + w_{1,t} \\
y_{2,t} &= \phi_{2,1}z_{1,t} + \phi_{2,2}z_{2,t} + \phi_{2,3}z_{3,t} + \phi_{2,4}z_{4,t} + w_{2,t}
\end{align*} \]

(14)

uniquely identified for all parameter values.

First consider special case \( \beta_{2,1} = 0 \): supply equation has no contemporaneous variables. From (13), \( \phi_{2,2} = 0 \), so \( z_{2,t} \) in \( y_{2,t} \) equation is 0% of time. Add \( y_{1,t} \), instrumenting by \( z_{2,t} \), being excluded. \( \beta_{2,1} = 0 \), so \( y_{1,t} \) detected 0% of time. Second equation confirmed as (14) with \( \phi_{2,2} = 0 \). Add \( y_{2,t} \) to \( y_{1,t} \) equation, instrument by excluding \( z_{i,t} \). If reduction results, postulated ‘structure’ over-identified: add eliminated regressors to instruments for Sargan test. Search continues till all paths explored. Store model, revert to original system, drop next \( z_{i,t} \). Even when \( \beta_{2,1} \neq 0 \), similar algorithm works. Generalizes to 3-equation example, but tedious analysis. Some steps involve single instrument estimation, so finite-sample moments for the associated estimators. Care required in Monte Carlo.

**Treat simultaneity as a reduction, practical algorithm appears feasible:** can determine ‘structural’ form in 2-equation systems *even when no a priori restrictions available.*

7 More variables than observations

‘Saturate’ regression by indicator variables. \( y_t \sim \mathcal{N}[\mu, \sigma^2_t] \) for \( t = 1, \ldots, T \). \( \mu \) is parameter of interest. Uncertain of outliers, so add \( T \) indicators \( I_t = 1_{(t=t_i)} \). **Perfect fit: nothing is learned.** Instead, add half of indicators, select as usual; then other half, select again. Now combine terminal models and select as usual. Just ‘dummying out’ \( T/2 \) observations for estimating \( \mu \). \( \alpha T \) indicators selected on average. **Feasible algorithm exists:** Hendry, Johansen and Santos (2004).

Can investigate several combinations: \( T/2 \) randomly; \( T/3 \) split and random etc. No impact on null rejection frequency. \( \hat{\mu} \) is unbiased; can bias correct \( \hat{\sigma}^2 \). If \( \mu \) takes 2 values \( \mu^* \), pre and post \( T^* \); blocks of \( I_t \) significant with average value \( \mu - \mu^* \): reveals step shift.

If \( y_{t-1} \) falsely included, and no indicators, get similar, opposite-sign impulses at break end points: conventional ‘outlier removal’ accepts incorrect model. But if saturate in subsets, reveals location shift, and that \( y_{t-1} \) irrelevant (on artificial data).

Consider general case of 4 groups of candidate variables \( x_{i,t} \), for \( i = 1, \ldots, 4 \), \( n_i << T \) but \( n = n_1 + \cdots + n_4 > T \). DGP is:

\[ \begin{align*}
y_t &= \sum_{i=1}^{4} \beta^i x_{i,t} + \epsilon_t \quad \text{where} \quad \epsilon_t \sim \mathcal{N}[0, \sigma^2_t],
\end{align*} \]

(15)
Figure 12: Null distributions of $\tilde{\mu}$ and $\hat{\mu}$

Figure 13: Null distributions of $\tilde{\sigma}^2$ and $\hat{\sigma}^2$
but \( \beta_i \) contain many zeros, so \( k << T \). Fit six ‘general models’ \((i, j = 1, 4; \ i \neq j)\):

\[
y_t = \gamma'_i x_{i,t} + \gamma'_j x_{j,t} + u_t^{(i,j)} \quad \text{where} \quad u_t^{(i,j)} \sim \text{IN} \left[ 0, \sigma^2_{u(i,j)} \right],
\]

(16) does not coincide with (15). Select terminal models from all pairings of blocks.

Assume sub-models congruent against own information, but if non-congruent, use HAC standard errors. Union of all resulting terminal models is next GUM. Now use standard \( \text{PcGets} \) approach. \( 2C_4 = 6 \) combinations \([(1,2) (3,4), (1,3) (2,4), (1,4), (2,3)]\): but procedure easily automated.

When \( x_{i,t}, x_{j,t} \) mutually orthogonal: delivers ‘correct’ answer if initial loose significance, followed by stringent critical values at final stage. For \( n_1 = \cdots = n_4 = 50 << T = 150 \) and 1% level: 2 irrelevant variables retained on average despite \( n = 200 \).

Two variants merit exploring. (A) Cumulatively retain highly significant variables in terminal models for addition to next step. Will improve intermediate goodness of fit, so can use more stringent selection criteria, reducing numbers of variables in intermediate models. (B) Put first \( k \) principal components in all sub-models. Alternative to ‘factor forecasting’ where \( n >> T \). Test if factors sufficient, or also need some salient regressors.

When all \( x_{i,t} \) and \( x_{j,t} \) positively correlated, efficiency of selection lower even if a single stage. Inter-correlations entail proxies improve intermediate fit. When some \( x_{i,t} \) and \( x_{j,t} \) negatively correlated, each required for other to be included. Cross-matching needed so appropriate pairs always jointly included.

8 Selecting non-linear models

Non-linearity inherent in economics Linear representation is a simplifying assumption: could be incorrect. Need general models to ensure all relevant factors included Objective: extend automatic selection to non-linearity in \( \text{PcGets} \). Investigate large class of functions. Establish operating characteristics. Joint research with Castle (2004) on RETINA (Perez-Amaral, Gallo and White, 2003).

Common to begin from previous linear model: simple to general in two respects. Between studies bound to be: perhaps why empirical advances are so difficult. Should not just extend best earlier model. Commence with general non-linear approximation: needs to be identified, and congruent. Add proposed functions (logistic, squashing or whatever) one at a time to test if they simplify the approximation. Avoids lack of identification under the null, and directly tests that postulated functions are valid.

DGP is:

\[
y_t = f(z_{1,t} \ldots z_{k,t}) + \epsilon_t \quad \text{where} \quad \epsilon_t \sim \text{IN} \left[ 0, \sigma^2_{\epsilon} \right].
\]

GUM is:

\[
y_t = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \beta_{i,j} h_{i,j}(z_{i,t}) + v_t.
\]

(18)

Basic relevant variables \((z_{1,t} \ldots z_{k,t})\) known; parsimonious functional forms are not. Example:

\[
\sum_{j=1}^{n_i} \beta_{i,j} h_{i,j}(z_{i,t}) = \sum_{j=1}^{n_i} \beta_{i,j} z_{i,t}^j.
\]

(19)

GUM first tested for congruence: sustains valid inference. (18) over-parameterized, so seek parsimonious approximation:

\[
y_t = \sum_{i=1}^{k} \beta_{i,0} z_{i,t} + \sum_{i=1}^{s} \alpha_i g_i(z_{i,t}) + v_t.
\]

(20)

Test for reducing (18) to linearity:

\[
y_t = \sum_{i=1}^{k} \beta_{i,0} z_{i,t} + v_t.
\]

(21)

Accept, (21) is final model; reject, non-linearity established. Add \( g_i(z_{i,t}) \) to (18): check for eliminating \( h_{i,j}(z_{i,t}) \), etc. If no feasible reduction, (18) not sufficiently general, or \( g_i(z_{i,t}) \) not appropriate approximations.
Problems and Solutions

Key problems with selecting under non-linearity: correlations between the linear and non-linear functions; extreme observations leading to non-normality; many potentially-relevant non-linear functions resulting in excess retention of irrelevant variables. Respective solutions are: reparameterize model to an uncorrelated representation; remove extreme observations using ‘indicator saturation’ techniques; develop a ‘super conservative’ selection strategy.

**RETINA: The basic algorithm**

Data building and sorting; generate non-linear functions. Isolating a candidate model; based on best out-of-sample performance. The search strategy; select a more parsimonious model. Model selection; search all sub-sample combinations.

Level 1 transforms: \(X_{i,k}^\alpha X_{i,j}^\beta\) for \(\alpha, \beta = -1, 0, 1\).

Division of sample into 3 disjoint sub-samples and rank based on \(W\). Include in excess retention of irrelevant variables. Respective solutions are: reparameterize model to an uncorrelated representation; remove extreme observations using ‘indicator saturation’ techniques; develop a ‘super conservative’ selection strategy.

| Case 1: orthogonal model from Hoover and Perez (1999). |
| Case 2: non-orthogonal model – add \(\mu\) to white noise. |

**Monte Carlo Evidence**

Aim: measure size and power of RETINA and PcGets.

**Linear GUM** 22 regressors, 17 nuisance:

\[ y_t = \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \pi_{k,i} x_{k,t-i} + \pi_{0,0} + u_t, \quad u_t \sim \text{IN} [0, \sigma^2] \]

Non-linear GUM: 62 regressors, 57 or 54 nuisance:

\[ y_t = \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \sum_{i=0}^{1} \pi_{k,i} x_{k,t-i} + \pi_{0,0} + u_t, \quad u_t \sim \text{IN} [0, \sigma^2] \]

...
Figure 14: Linear and non-linear GUM and DGP

<table>
<thead>
<tr>
<th>Sample</th>
<th>PcGets: Linear GUM</th>
<th>PcGets: Non-linear GUM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lib</td>
<td>Cons</td>
</tr>
<tr>
<td>T=100</td>
<td>0.608</td>
<td>0.413</td>
</tr>
<tr>
<td>T=1000</td>
<td>0.790</td>
<td>0.632</td>
</tr>
<tr>
<td>T=100</td>
<td>0.911</td>
<td>0.821</td>
</tr>
<tr>
<td>T=1000</td>
<td>1.000</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Note: critical values are sample-size dependent; t values are independent of $T$.

Non-orthogonal case. DGP:

\[
y_t = \beta_j x_{j,t} + \epsilon_t, \quad \epsilon_t \sim \text{IN}[0, 1]
\]

\[
x_t = 10 + \nu_t, \quad \nu_t \sim \text{IN}_2[0, I_2].
\]

for $t = 1, \ldots, T$ and $j = 1$ or 2. GUM:

\[
y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \beta_4 x_{2,t}^2 + \beta_5 \left( \frac{1}{x_1} \right)_t
\]

\[
+ \beta_6 \left( \frac{1}{x_2} \right)_t + \beta_7 \left( \frac{1}{x_1^2} \right)_t + \beta_8 \left( \frac{1}{x_2^2} \right)_t + \beta_9 (x_1 x_2)_t + \beta_{10} \left( \frac{1}{x_1 x_2} \right)_t
\]

\[
+ \beta_{11} \left( \frac{x_1}{x_2} \right)_t + \beta_{12} \left( \frac{x_2}{x_1} \right)_t + \epsilon_t, \quad \epsilon_t \sim \text{IN}[0, \sigma^2].
\]

Substantial collinearity is generated between the linear and non-linear functions: $\text{corr}(x_1, x_1^2) = 0.9775$; $\text{corr}(1, x_1^2) = 0.9899$. To ensure power, large $\beta_j$s are required for significant non-central t-values.
<table>
<thead>
<tr>
<th>Sample</th>
<th>0.05</th>
<th>0.01</th>
<th>0.05</th>
<th>0.01</th>
<th>0.05</th>
<th>0.01</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>t=2</td>
<td>0.465</td>
<td>0.239</td>
<td>0.507</td>
<td>0.285</td>
<td>0.414</td>
<td>0.219</td>
<td>0.453</td>
<td>0.240</td>
</tr>
<tr>
<td>t=3</td>
<td>0.805</td>
<td>0.589</td>
<td>0.848</td>
<td>0.659</td>
<td>0.621</td>
<td>0.432</td>
<td>0.828</td>
<td>0.653</td>
</tr>
<tr>
<td>t=4</td>
<td>0.960</td>
<td>0.864</td>
<td>0.979</td>
<td>0.912</td>
<td>0.645</td>
<td>0.443</td>
<td>0.985</td>
<td>0.915</td>
</tr>
<tr>
<td>t=6</td>
<td>1.000</td>
<td>0.997</td>
<td>1.000</td>
<td>1.000</td>
<td>0.986</td>
<td>0.966</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>t=8</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>t=4 (x_1^2)</td>
<td>0.934</td>
<td>0.828</td>
<td>0.982</td>
<td>0.921</td>
<td>0.746</td>
<td>0.595</td>
<td>0.910</td>
<td>0.800</td>
</tr>
<tr>
<td>t=3 (x_2x_3)</td>
<td>0.782</td>
<td>0.566</td>
<td>0.855</td>
<td>0.670</td>
<td>0.441</td>
<td>0.192</td>
<td>0.735</td>
<td>0.523</td>
</tr>
<tr>
<td>t=4 (x_4x_5)</td>
<td>0.940</td>
<td>0.838</td>
<td>0.971</td>
<td>0.912</td>
<td>0.645</td>
<td>0.443</td>
<td>0.990</td>
<td>0.968</td>
</tr>
</tbody>
</table>

Case 1: \( j = 1, \beta_1 = 0.4 \) for \( T = 100, \beta_1 = 0.12649 \) for \( T = 1000 \): corresponds to \( t_{\beta_1} = 4 \) in orthogonal model.

Case 2: \( j = 1, \beta_1 = 1100 \) for \( T = 100, \beta_1 = 180 \) for \( T = 1000 \): corresponds to \( t_{\beta_1} \approx 4 \).

Case 3: \( j = 2, \beta_1 = 800, \beta_2 = 1600 \) for \( T = 100, \beta_1 = 140, \beta_2 = 280 \) for \( T = 1000 \): corresponds to \( t_{\beta_1} \approx 3, t_{\beta_2} \approx 6 \).

Surprising change if two relevant variables in model: RETINA mis-performs badly, as seen in figure 15b. To orthogonalize, need to double de-mean variables, e.g.: \( \tilde{x}_1 = (x_{1,t} - \bar{x}_1)^2 - (x_{1,t} - \bar{x}_1)^2 \) where \( \tilde{x} \) is the de-meaned variable and \( \bar{x} \) is the mean of \( x \).

Conclude that collinearity can lead to substantial problems. Must orthogonalize first. De-meaning still leaves collinearity with the intercept. Double de-meaning results in more orthogonal GUM. Check for outliers before testing for non-linear functions. Probability of outliers ‘lining up’ in a regression between two unconnected variables increased with non-linear functions. Can use ‘impulse saturation’ to check
Use of sub-samples delivers gains in size but costs in power. Full-sample information at tighter significance is preferable. Separately control test size for linear and non-linear functions. Use of a ‘super-conservative’ strategy in PcGets for non-linear functions. Non-linearities and interactions can be tested for at low cost. Evidence from both programs indicates automatic model selection algorithms achieve considerable success.

9 Conclusions

Models often have many potential candidate variables: search and selection pandemic. **Automatic model selection is invaluable tool.** Tool already demonstrated remarkable performance: selection almost non-distortionary. DGP found from GUM almost as often as from itself. **Power near upper bound of scalar t-test,** so retention of relevant variables close to theory maximum; Elimination of irrelevant at rate of significance level, so stabilized size relative to nominal significance level. **Selected estimates have appropriate standard errors; and can be bias corrected** which also downweights adventitious significance.

No serious additional costs of search over inference: **‘pre-test’ biases arise from simplifying DGP, not search.** $\hat{\sigma}$ within $\pm 5\%$ of $\sigma$, for both strategies.

Higher probability of finding DGP from GUM on Liberal, **than from DGP using Conservative.** Astonishing given Lovell (1983): emphasizes importance of choice of strategy. Sub-sample reliability reduces size at small cost in power: trade-off is relatively costless. Non-orthogonal designs remain problematic: one area where expert knowledge is clearly valuable. Still, ‘quick modeller’ option for non-expert users.

Regression too simple for real-world econometrics. But **Gets** is applicable to: simultaneous equations systems; vector autoregressions (VARs) (Krolzig, 2003); endogenous regressors if sufficient valid instruments; cointegration vectors (Omtzig, 2002, Kurcewicz and Mycielski, 2003); powerful for cross-sections (Hoover and Perez, 2004); general non-linear models (Perez-Amaral et al., 2003; Castle, 2004); and even when more regressors than observations: (Hendry et al., 2004). **Further developments will improve performance, and widen scope of automatic modelling procedures.**

References


